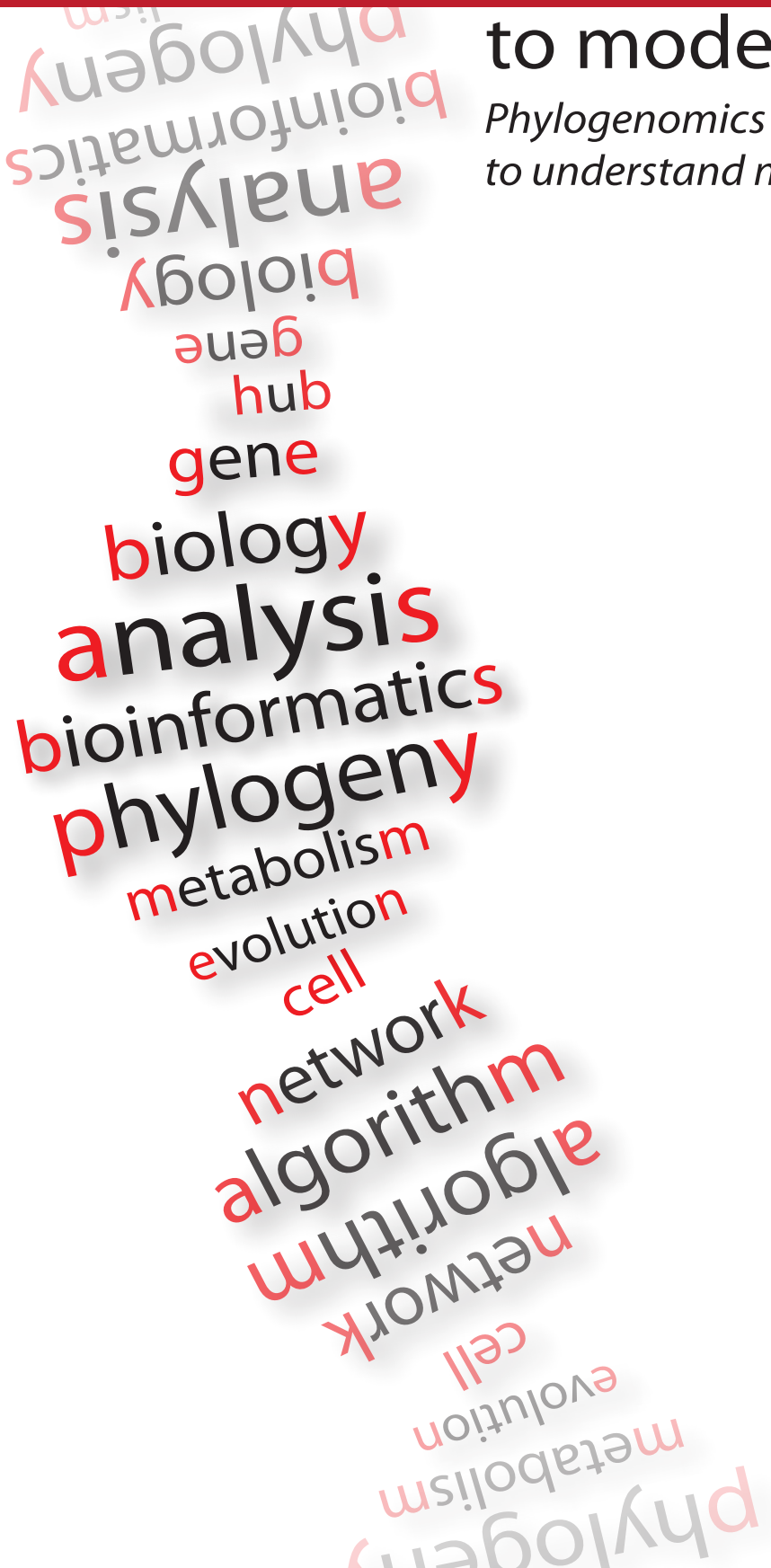
Bioinformatics of genome evolution: from ancestral to modern metabolism

Phylogenomics and comparative genomics to understand microbial evolution



Marco Fondi

Bioinformatics of genome evolution: from ancestral to modern metabolism

Phylogenomics and comparative genomics to understand microbial evolution

Marco Fondi
Dep. of Evolutionary Biology
University of Florence

A dissertation submitted in candidature for the degree of
Doctor of Philosophy in Genetics
Florence, December 2009

Acknowledgements

Fun, (hard) work, hope, (few) delusions, travels, (a lot of) meetings, computers, coffees, friends, papers, joy. This is what my PhD period at the LEMM is made of. I think I could hardly have asked more at the beginning.
Obviously, without the help of the people you will find below I still would be almost at the starting point. I am deeply in debt with:

Renato Fani for supervising me and my work with extreme interest and devotion, and for deeply stimulating my scientific production. He is deeply sustaining me since 2004 and it seems I will never be able to thank him enough.

Matteo Brilli for constant help and discussion. Most of the ideas I had in these three years were born during our conversations in Fani's lab and I probably stole from him the term "Comparative Evolutionary Genomics" during one of these.

I would also like to thank the following persons because they strongly (and positively) influenced me while I was performing my analyses:
Alessio Mengoni, **Giovanni Emiliani**, **Simonetta Gribaldo**, **Pietro Lio'**, **all the past and present staff of the Fani's lab** (there isn't enough space to mention everyone, sorry guys!).
A special thank to **Pino** (pino), **Antonio** and **Giorgio**. Real friends.

This was science. Now the serious stuff:
My family allowed me to live the life I wanted to live, with every means but, most of all, with love. **Sara** is the brand new part of my family and she is doing exactly the same. I'm sure I will never be able to thank you enough.

Most of the data presented in this thesis has been published on the following, *peer reviewed* papers:

- **Fondi M**, Brilli M, Emiliani G, Paffetti D, Fani R: The primordial metabolism: an ancestral interconnection between leucine, arginine, and lysine biosynthesis. BMC Evol Biol 2007, 7 Suppl 2:S3.

- **Fondi M**, Brilli M, Fani R: On the origin and evolution of biosynthetic pathways: integrating microarray data with structure and organization of the Common Pathway genes. BMC Bioinformatics 2007, 8 Suppl 1:S12.

- **Fondi M**, Emiliani G, Lio P, Gribaldo S, Fani R: The Evolution of Histidine Biosynthesis in Archaea: Insights into the his Genes Structure and Organization in LUCA. J Mol Evol 2009, 69:512-526.

- **Fondi M**, Emiliani G, Fani R: Origin and evolution of operons and metabolic pathways. Res Microbiol 2009, 160(7):502-512.

- Fani R, **Fondi M**: Origin and evolution of metabolic pathways, Physics of Life Reviews. 2009 6(1):23-52.

- Fani R, Brilli M, **Fondi M**, Lio P: The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. BMC Evol Biol 2007, 7 Suppl 2:S4.

- Brilli M, Mengoni A, **Fondi M**, Bazzicalupo M, Lio P, Fani R: Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. BMC Bioinformatics 2008, 9:551.

- Papaleo MC, Russo E, **Fondi M**, Emiliani G, Frandi A, Brilli M, Pastorelli R, Fani R: Structural, evolutionary and genetic analysis of the histidine biosynthetic "core" in the genus Burkholderia. Gene 2009, 448(1):16-28.

- Emiliani G, **Fondi M**, Fani R, Gribaldo S: A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. Biol Direct 2009, 4:7.

Submitted for publication

- **Fondi M**, Bacci G, Brilli M, Papaleo MC, Mengoni A, Vaneechoutte M, Dijkshoorn L, Fani R, Exploring the evolutionary dynamics of plasmids:the Acinetobacter pan plasmidome, Submitted for publication to BMC Evolutionary Biology.

- M Brilli, **M Fondi**, R Fani, A Mengoni, L Ferri, M Bazzicalupo and EG Biondi. The diversity and evolution of cell cycle regulation in alphaproteobacteria: a comparative genomic analysis. Submitted for Pubblication to BMC Systems Biology

- E Perrin, **M Fondi** , MC Papaleo, I Maida, S Buroni, MR Pasca, G Riccardi, R Fani, Exploring the HME and HAE1 efflux systems in the genus Burkholderia, Submitted for Pubblication to BMC Evolutionary Biology

Book chpaters

- **M Fondi**, G Emiliani, R Fani. The primordial metabolism: on the origin and evolution of metabolic pathways and operons, In "Life and Time", edited by; S. Casellato, P. Burighel, A. Minelli, CLEUP (Coop. Libraria Editrice, Universit di Padova), pp. 87-113.

- G Emiliani, **M Fondi**, P Li, R Fani, Evolution of metabolic pathways and evolution of genomes, In "Geomicrobiology: Molecular and Environmental Perspective". In press.

Abstract

The whole body of data embedded in this thesis can be easily subdivided into two different major parts: the first (namely "Origin and Evolution of Metabolic Pathways", Part I) deals with evolutionary events that likely played a key role in the assembly and in the shaping of modern biosynthetic routes. In particular, four different metabolic circuits have been explored, namely i) histidine and ii) lysine biosynthetic pathways, iii) the nitrogen fixation process and iv) the phenylpropanoid metabolism. Results obtained in this part have underlined, from one side, the importance of some molecular mechanisms (gene duplication, gene fusion, horizontal gene transfer) in the assembly and in the shaping of microbial (and plant) metabolism and, from the other, have allowed to infer the timing of appearance of some crucial biosynthetic steps in the past. The second part of the work (Part II), instead, deals with comparative genomics and, in particular with the analysis of i) plasmid and ii) antibiotic resistance-related sequences. In particular, a newly developed bioinformatic package allowing the automatic phylogenetic profiling and the visualization of homology relationships in a large number of plasmid sequences is presented. Data obtained with this software and with other *ad hoc* implementations have allowed tracing the evolutionary history of plasmids belonging to Enterobacteriaceae group and to *Acinetobacter* genus. Furthermore, in this second part of the thesis, we tried to evaluate the horizontal flow of antibiotic resistance coding genes (the resistome) across the microbial community and ii) to identify those ecological niches (if any) whose inhabitants mostly contribute to their mobilization. Still in the context of bacterial antibiotic resistance, we have performed a comprehensive computational analysis concerning both the distribution and the phylogeneny of the HAE1 and HME efflux systems (involved in antibiotics and heavy metal recognition) in the genus *Burkholderia*, providing a i) deeper knowledge of the presence, the structure and the distribution of RND proteins in these species and ii) an evolutionary model accounting for their appearance and maintenance in this genus. Despite their "case-by-case" relevance, results presented in this dissertation partially show how bioinformatics has now affected several fields of biology. In fact, the analysis of sequence data can be used in different fields, including evolution (e.g. the assembly and evolution of metabolism), infections control (e.g. the horizontal flow of antibiotic resistance), ecology (bacterial bioremediation), providing crucial hints for the understanding of the main biological systems (including their evolutionary dynamics) and also allowing a more accurate design of *wet lab* experiments

Contents

List of Figures

Chapter 1

Introduction

1.1 From ancestral to modern genomes

Although considerable efforts have been made to understand the emergence of the first living beings, we still do not know when and how life originated [Pereto *et al.*, 2000]. However, it is commonly assumed that early organisms inhabited an environment rich in organic compounds spontaneously formed in the prebiotic world. This heterotrophic origin of life is generally assumed and is frequently referred to as the Oparin-Haldane theory [Lazcano & Miller, 1996; Oparin, 1936, 1967]. If this idea is correct, life evolved from the so called "primordial soup", containing different organic molecules (many of which are still used by the extant life forms), probably formed spontaneously during the Earth's first billion years. This "soup" of nutrient compounds was available for the early

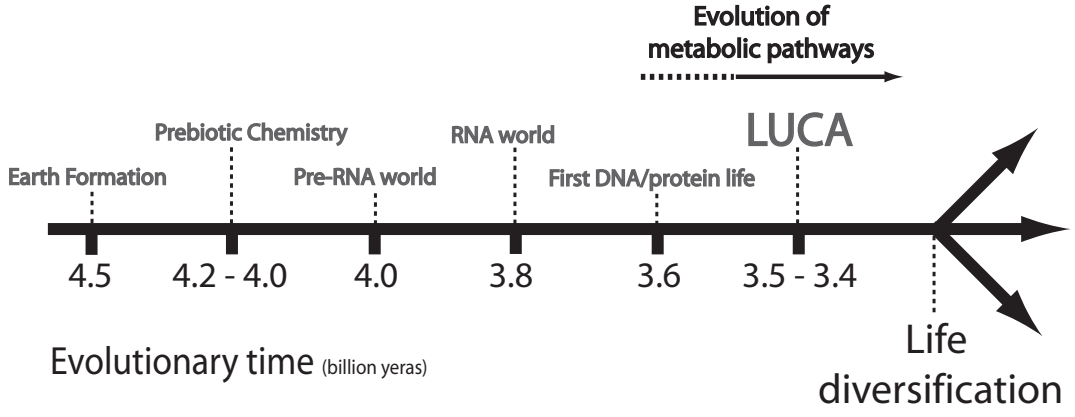


Figure 1.1: Evolutionary time line from the origin of Earth to the diversification of life.

heterotrophic organisms, so they had to do a minimum of biosynthesis. An experimental support to this proposal was obtained in 1953 when Miller [Miller, 1953] and Urey showed that amino acids and other organic molecules are formed under atmospheric conditions thought representative of those on the early Earth. The first living systems probably did stem directly from the primordial soup and evolved relatively fast up to a common

ancestor, usually referred to as LUCA (Last Universal Common Ancestor), an entity representing the divergence starting-point of all the extant life forms on Earth (Figure 1.1). Even though some progress have taken place in the last years, we are nowhere near completely filling the gap existing between prebiotic events and the appearance of the LUCA. It is quite possible that during this extremely complex transition the intermediate stages might have involved simpler organisms with much smaller (genes and) genomes. As stated by Delaye et al. [2005], defining the nature of the LUCA is one of the central goals of the study of the early evolution on Earth; several attempts have been made in this direction and the nature of LUCA is still under debate. It has been recently proposed that LUCA was not a cell, but an inorganically housed assemblage of expressed and replicable genetic elements [Koonin & Martin, 2005]. A very different view was suggested one decade ago by Woese [2002; 1998] who proposed that LUCA could not have been a particular organism or a single organismal lineage, but actually a community of simpler organisms, the progenotes (Figure 1.2). This community evolved into a smaller

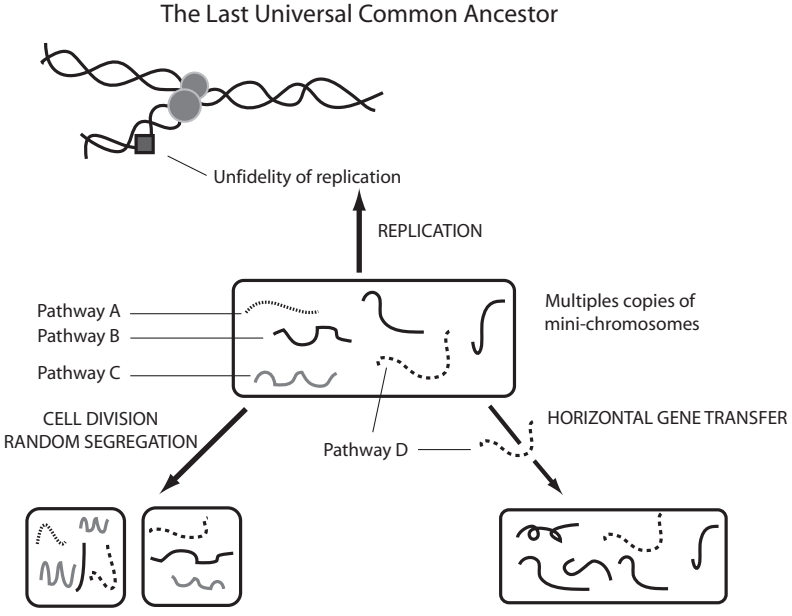


Figure 1.2: Representation of a possible progenotes community

number of more complex cell types, which ultimately developed into the ancestor(s) of all the extant life domains. At the beginning, the major driving force through which these early life forms progressively evolved and increased their complexity, was probably horizontal gene transfer (HGT). This, together with the inaccuracy of the first information processing, determined the high genetic temperature in which, over time, these primordial (micro)organism evolved into a smaller number of increasingly complex cell types with the ancestors of the three primary groupings of organisms arising as a result. It is important to clearly state that LUCA should not be confused with the first cell, but was probably the product of a long period of evolution. Being the last means that LUCA was preceded by a long succession of older ancestors. In this framework, a plethora of cellular lineages

that have left no descendants today may have existed before LUCA [Forterre & Gribaldo, 2007]. It must be taken into account that many of these were probably still present at the time of LUCA, and some have probably even coexisted for some time with its descendants, possibly contributing via horizontal gene transfer to some traits present in modern lineages (Figure 1.3) [Forterre & Gribaldo, 2007]. According to this view, contemporary

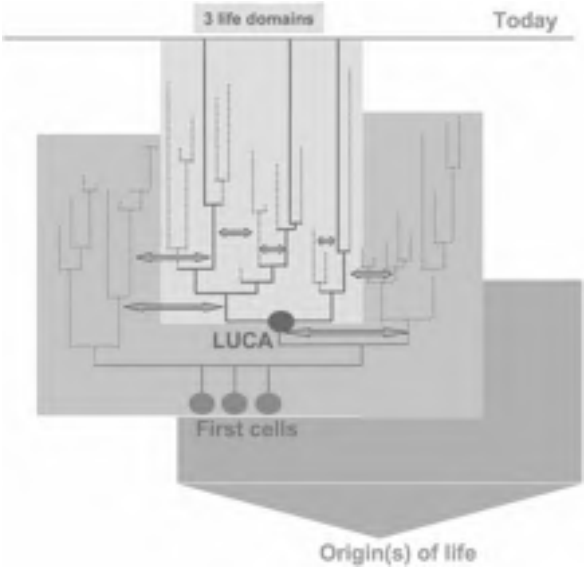


Figure 1.3: LUCA was the last bottleneck in a long series of ancestors to the three present-day cellular domains: Archaea, Bacteria, and Eukarya (from [Forterre & Gribaldo, 2007]).

genomes are the result of 3.54 billions of years of evolution. But how did these ancestral genomes look like? The increasing number of available sequences from organisms belonging to the three domains of life (Bacteria, Archaea and Eukarya) and the implementation of several bioinformatic tools has allowed inferring both the size and the gene content of the genomes of the first living cells that appeared on the Earth. A recent estimate of the minimal gene content of LUCA based on whole-genome phylogenies indicated that ancestral genomes were probably composed by about 1000/1500 genes [Ouzounis *et al.*, 2006]. The results of this analysis are in contrast with the previous notion of a minimal genome (embedding only 500/600 genes) based on comparative genomics analysis of essential genes [Koonin, 2003]. However, despite this small gene content, ancestral genomes were probably fairly complex, similar to those of the extant free-living prokaryotes and included a variety of functional capabilities including metabolic transformation, information processing, membrane/transport proteins and complex regulation [Ouzounis *et al.*, 2006]. Although genome size appears highly variable among organisms with the same level of morphological complexity [Sharov, 2006], it seems well-established that the vast majority of the modern-day organisms (with the exception of secondary genome reductions) (i) possesses much more than the hypothetical gene content of LUCA and (ii) displays a great complexity (gene regulatory and protein interaction networks, mobile genetics element, etc.). Hence, starting from a common pool of highly conserved genetic information, still

shared by all the extant life forms, genomes have been shaped to a considerable extent during evolution, leading to the great diversification of life (and genomes) that we observe nowadays. This raises the intriguing question of how both genome size and complexity could have been increased during evolution. In other words, which are the molecular mechanisms that drove the evolution of the earliest genes and genomes? As we will see in the next sections, the evolution of genes and genomes requires (at least) two main steps, i.e. (i) the acquisition of new genetic material and (ii) its shaping to (eventually) develop a new function. The first is usually achieved with either the HGT process or by the duplication of DNA stretches, whereas the second, that is generally gained through evolutionary divergence, can be satisfied through several different molecular mechanisms such as changes in the catalytic or regulatory domains or fusions involving two (or more) cistrons. However, a demarcation line between the origin(s) and the subsequent evolution of metabolic routes should be traced. There are many indications supporting the idea that in the early stages of cell evolution RNA molecules played a key and central role in cellular processes [Delaye *et al.*, 2005, and references therein]. It is quite possible that the origin of metabolic pathways can be placed within an RNA or an RNA/protein world rather than in a DNA/protein world. Therefore, we will mostly take into account those post-origin evolutionary events that might have played a major role in shaping metabolic pathways.

1.2 Molecular Mechanisms of Genomes Expansion

1.2.1 The Starter Types

It has been recognized that most genetic information is not essential for cell growth and division. Indeed, the analysis of completely sequenced genomes led to the suggestion that 256 genes are close to the minimal gene set that is necessary and sufficient to sustain the existence of a modern-type cell [Mushegian & Koonin, 1996]. However, it is not known if such a set of sequences were already present in the first DNA/protein organisms. As it will be discussed later, most arose by gene duplication. The uncertainty is the number of enzymes that did not arise in this manner, i.e., the starter types. The term starter type genes was firstly conyed by Lazcano and Miller [1994] to refer to the original ancestral genes that underwent (many) duplications and gave rise to the extant paralogous gene families (i.e. those genes that share an ancestral sequence within the same organism, see below for details). It is very unclear how the starter types genes originated. Two years later, the same authors [Lazcano & Miller, 1996] estimated that the number of starter types might have ranged from 20 up to 100. Their idea was based on the similarity of many biochemical reactions, and on the observation that many proteins of related function share the same ancestry within a given organism.

1.2.2 Gene Duplication

Different molecular mechanisms may have been responsible for the expansion of early genomes and metabolic abilities. Data obtained in the last decade clearly indicate that a very large proportion of the gene set of different organisms is the outcome of more or less ancient gene duplication events predating or following the appearance of the LUCA [Ohte, 2000]. These findings strongly suggest that the duplication and divergence of DNA sequences (Figure 1.4) of different size represents one of the most important forces driving the evolution of genes and genomes during the early evolution of life. In fact, the relative

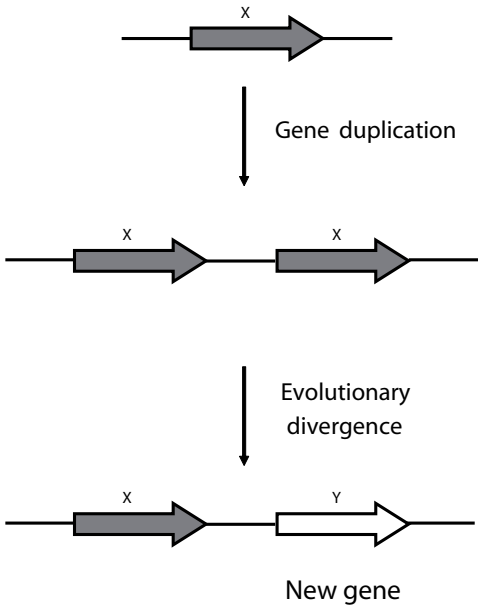


Figure 1.4: Gene duplication.

content of paralogous genes (i.e. the products of a gene duplication event) in extant bacterial genomes has been shown to increase together with genome size (Figure 1.5). Indeed, this process may allow the formation of new genes from pre-existing ones. However, there are a number of additional mechanisms that could have increased the rate of metabolic evolution, including the modular assembly of new proteins by gene fusion events, and horizontal gene transfer, the latter permitting the transfer of entire metabolic routes or part thereof. Even though the lack of fossil records strongly hinders the understanding of biochemical evolution, there is evidence that the basic biosynthetic routes were assembled in a short geological timescale [Pereto *et al.*, 2000]. Indeed, it is quite possible that once an ancestral genetic system (a starter type gene) encoding a functional catalyst (or structural protein) appeared, it will undergo very rapidly paralogous duplications (Lazcano A, *personal communication*). Assuming that Archaean cells had a random rate of duplication fixation, and a rate of spontaneous gene duplications comparable with the present values of $10^5/10^3$, it has been suggested that the time required for the development of a 100 kb genome of a DNA/protein primitive heterotroph into a 7000-genes filamentous cyanobacteria would require only 7106 years [Lazcano & Miller, 1996; Sharov, 2006]. Thus, the

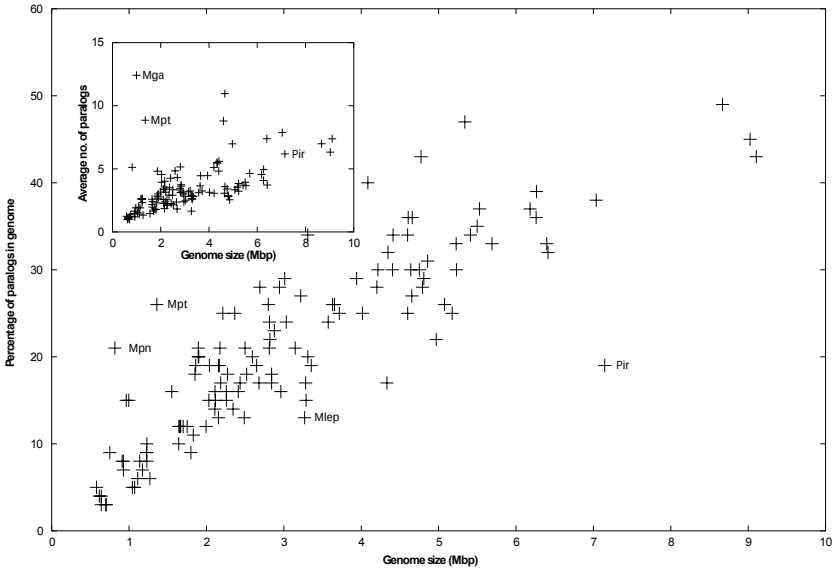


Figure 1.5: Relationship between percentage of genes belonging to paralogous families plotted versus genome size in 127 bacterial genomes

rate of duplication and fixation of new genes can be surprisingly fast on the geological timescale. This idea is supported by directed evolution experiments that have shown that new substrate specificities appear in a few weeks from existing enzymes by recombination events within a gene [Hall & Zuzel, 1980]. The importance of gene duplication for the development of metabolic innovations was firstly discussed by Lewis [1951] and later by Ohno [1972a] and has been recently confirmed by the comparative analysis of complete sequences of archaeal, bacterial and eukaryal genomes. It has already been shown that all of these organisms harbor a remarkable proportion of paralogous genes and that many of them group into numerous families of different sizes [de Rosa & Labedan, 1998; Labedan & Riley, 1995]. In principle, a DNA duplication may involve (i) part of gene, (ii) a whole gene, (iii) DNA stretches including two or more genes involved in the same or in different metabolic pathways, (iv) entire operons, (v) part of a chromosome, (vi) an entire chromosome, and finally (vii) the whole genome [Fani, 2004]. Two structures or sequences that evolved from a single ancestral structure or sequence, after a duplication event, are referred to as homologs. The terms orthology and paralogy were introduced to classify different types of homology (Figure 1.2.2). Orthologous structures or sequences in two organisms are homologs that evolved from the same feature in their last common ancestor but they do not necessarily retain their ancestral function. This is the case of orthologous transcription factors in bacteria that have been shown to have different functions and to regulate different genes [Price *et al.*, 2007]. Therefore, the evolution of orthologs reflects organismal evolution. Homologs whose evolution reflects gene duplication events are called paralogs. Consequently, orthologs usually perform the same function in different organisms, whereas paralogous genes often catalyze different, although similar, reactions.

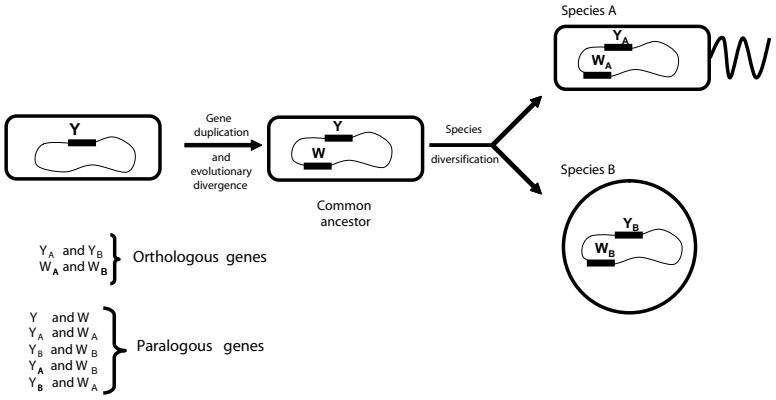


Figure 1.6: Orthologous and paralogous genes.

Two paralogous genes may also undergo successive and differential duplication events involving one or both of them giving rise to a group of paralogous genes, which is referred to as paralogous gene family (Figure 1.7).

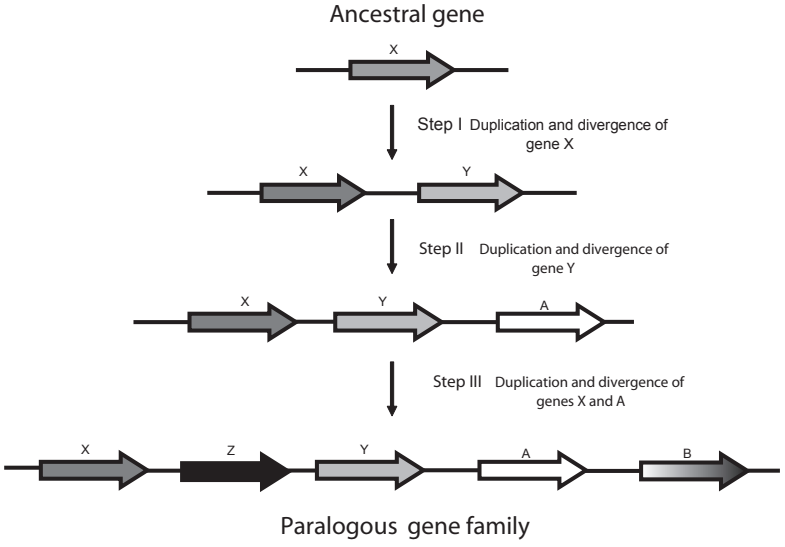


Figure 1.7: Schematic representation of the molecular steps leading to a paralogous gene family.

1.2.3 The Fate of Duplicated Genes

The structural and/or functional fate of duplicated genes is an intriguing issue that has led to the proposal of several classes of evolutionary models accounting for the possible scenarios emerging after the appearance of a paralogous gene pair.

1.2.3.1 Structural fate

Duplication events can generate genes arranged in-tandem. In addition duplication by recombination involving different DNA molecules or transposition can generate a copy of a DNA sequence at a different location within the genome [Fani, 2004; Li & Graur, 1991]. If an in-tandem duplication occurs, at least two different scenarios for the structural evolution of the two copies can be depicted: (i) the two genes undergo an evolutionary divergence becoming paralogs; (ii) the two genes fuse doubling their original size forming an elongated gene (see below). Moreover, if the two copies are not arranged in-tandem: (i) they may become paralogous genes; (ii) one copy may fuse to an adjacent gene, with a different function, giving rise to a mosaic or chimeric gene that potentially may evolve to perform other(s) metabolic role(s). Tandem duplications of DNA stretches are often the result of an unequal crossing-over between two DNA molecules, but other processes, such as replication slippage, may be invoked to explain the existence of tandemly arranged paralogous genes. The presence of paralogous genes at different sites within a microbial genome might be the results of ancient activity of transposable elements, and/or duplication of genome fragments as well as wholegenome duplications [Fani, 2004].

1.2.3.2 Functional fate

The functional fate of the two (initially) identical gene copies originated from a duplication event depends on the further modifications (evolutionary divergence) that one (or both) of the two redundant copies accumulates during evolution (Figure 1.8). It can be

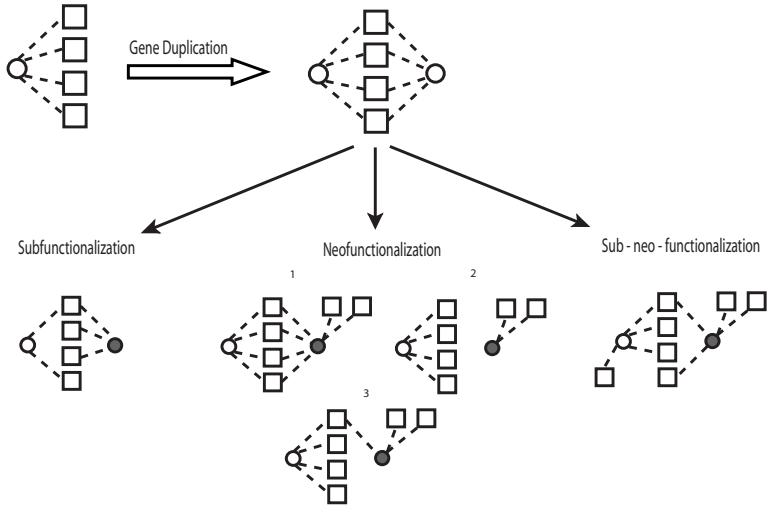


Figure 1.8: Evolutionary models of functional divergence between duplicate genes. Genes and the function(s) they code are represented with circles and squares, respectively. Dotted lines link genes with their functions.

surmised, in fact, that, after a gene duplicates, one of the two copies becomes dispensable and can undergo several types of mutational events, mainly substitutions, that, in turn, can lead to the appearance of a new gene, harboring a different function in respect to the ancestral coding sequence (Figure 1.4 and Figure 1.7). On the contrary, duplicated genes can also maintain the same function in the course of evolution, thereby enabling the production of a large quantity of RNAs or proteins (gene dosage effect); this is the case, for example, of prokaryotic 16S rRNA genes. At least three different models have been proposed to explain the early stages of evolutionary divergence of duplicated gene copies.

1. The classical model of gene duplication (neofunctionalization)
The classical model of gene duplication, or neofunctionalization, predicts that in most cases, after the duplication event, one duplicate may become functionless, whereas the other copy will retain the original function [Lio' *et al.*, 2007; Ohno, 1972b, 1980]. At least in the early stages after the gene duplication event, the two copies are supposed to maintain the same function. Then, it is likely that one of the redundant copy will be lost, due to the occurrence of one (or more) mutation(s) negatively affecting its original function that, in turn, will be preserved by the other redundant copy [Lio' *et al.*, 2007]. However, although less probably, an advantageous mutation may change the function of one duplicate and both copies may be maintained (Figure 1.8). Recent evidences, emerged by large-scale comparative genome analyses revealed that this hypothesis on the fate of duplicated genomes does not fit completely with data. In the case of eukaryotic multigene families, for example, it has been demonstrated that, if the size of the population is large enough, the fate of most duplicated genes is to acquire a new function rather than to become pseudogenes [Walsh, 1995]. Moreover, Nadeau and Sankoff [Nadeau & Sankoff, 1997], studying human and mice genes, estimated that about 50% of gene duplications undergo functional divergence. Other analyses have illustrated the high frequency of paralogous genes preservation following ancient DNA duplications events, being close to values of 30 to 50% over periods of tens to hundreds of millions of years [Lynch & Conery, 2000]. The release of several new fully sequenced genomes has allowed a further validation of these hypotheses. Aury et al. [2006] have observed that gene loss, following a whole genome duplication (WGD), occurs over a long timescale and not as an initial massive event. Accordingly, many genes are maintained after WGD not because of functional innovation but because of gene dosage constraints [Aury *et al.*, 2006]. After the analysis of data coming from comparative genomics and enzymes kinetics, Connant and Wolfe [Conant & Wolfe, 2007] proposed that duplicate copies of glycolysis genes were initially maintained for dosage reasons, but subsequent tuning of enzyme expression levels may have freed one paralog to innovate [Conant & Wolfe, 2007].

2. The sub-functionalization model
The sub-functionalization model for the fate and the maintenance of duplicates relies on the observation that a single gene can be made up of several accessorial components, i.e. promoter regions with a positive or negative effect on transcription of downstream genes,

different functional and/or structural domains of the protein they code for (eventually capable of interacting with different substrates and regulatory ligands, or other proteins) and so on. In this context, these elements can be considered as a sub-functional module for a gene or protein, each one contributing to the global function of that gene or protein. Starting from this idea, Lynch and Force [Lynch & Force, 2000] first proposed that multiple sub-functions of the original gene may play an important role in the preservation of gene duplicates (Figure 1.8). They focused on the role of degenerative mutations in different regulatory elements of an ancestral gene expressed at rates which depend on a certain number of different transcriptional modules (sub-functions) located in its promoter region. After the duplication event, deleterious mutations can reduce the number of active sub-functions of one or both the duplicates, but the sum of the sub-functions of the duplicates will be equal to the number of original functions before duplication (i.e.: the original functions have been partitioned among the two duplicates). Similarly, considering both duplicates, they are together able to complement all the original sub-functions; moreover, they can have partially redundant functions too [Lio' *et al.*, 2007]. The sub-functionalization, or duplication-degeneration-complementation model (DDC) of Lynch and Force [Lynch & Force, 2000], differs from the classical model because the preservation of both gene copies mainly depends on the partitioning of sub-functions between duplicates, rather than the occurrence of advantageous mutations. A limitation of the sub-functionalization model is the requirement for multiple independent regulatory elements and/or functional domains; the classical model is still valid if gene functions cannot be partitioned: for example, when selection pressure acts to conserve all the sub-functions together. This is often the case when multiple sub-functions are dependent on each other [Lio' *et al.*, 2007].

3. The sub-neofunctionalization model

A further implementation of all the models explaining the fate of duplicates has been proposed by He and Zhang [He & Zhang, 2006], starting from the results of a work concerning both yeast protein interaction data and human expression data, which have been tested both under the neo-functionalization and the subfunctionalization models. According to the authors, none of them alone satisfied experimental data for duplicates and the so-called sub-neofunctionalization model was introduced, being a mix of previous ones. The acquisition of expression divergence between duplicates is interpreted by He and Zhang [He & Zhang, 2006] as a (rapid) subfunctionalization event. Then, after this sub-functionalization occurred, both duplicates are essential, in that they can maintain the original expression patterns, and hence they are preserved. Once a gene is established in a genome, it can retain its function or evolve or specialize a new one (i.e. it undergoes neo-functionalization Figure 1.8) [Lio' *et al.*, 2007]. Accordingly, the sub-functionalization appear to be a rapid process, while the neo-functionalization requires more time and continues even long after duplication [He & Zhang, 2006].

1.2.4 Operon Duplication

DNA duplications may also concern entire clusters of gene possibly involved in the same metabolic process, i.e. entire operons or part thereof. For example, one can imagine that if an entire operon a, responsible for the biosynthesis of compound A, duplicates giving rise to a couple of paralogous operons, one of the copy (b) may diverge from the other and evolve in such a way that the encoded enzymes catalyze reactions leading to a different compound, B. If this event actually occurs, it might provoke a (rapid) expansion of the metabolic abilities of the cell and the increase of its genome size [Fani, 2004]. Moreover we should find the vestiges of this duplication by comparing the aminoacid sequence of the proteins encoded by operons A and B. Even though Gevers et al. [Gevers *et al.*, 2004] found only a minority of paralogous operons in some bacterial genomes, this doesn't mean that operon duplication did not occurred more frequently during early cell evolution. In fact, because the molecular clocks and functional constraints are different for each protein, if the duplication event was (very) ancient, it might have been blurred during evolution. In ammonia oxidizing autrophic bacteria multiple copies of ammonia monooxygenase (*amo*) operons have been disclosed (see [Klotz & Norton, 1998] and references therein). In addition to this, paralogous operons have been described in the archeon *Pyrococcus* [Maeder *et al.*, 1999]. A cascade of gene and operon duplications has been also suggested for the origin of nitrogen fixation (*nif*) genes [Fani *et al.*, 2000]. More recently, several interesting examples of operons duplication(s) have been disclosed in different microorganisms. One of the most and intriguing one is represented by the operon(s) whose products are responsible for the assembly and the functioning of the RND drug efflux pump system (RND family). This issue has been extensively analyzed by Guglierame et al. [Guglierame *et al.*, 2006], who discovered 14 paralogous operons embedding all the genes necessary for the assembly of a functional efflux system in the genome of *Burkholderia cenocepacia*. The presence of a large number of paralogous operons in the genome of all the available *Burkholderia* species strongly suggests that they are the outcome of several operon duplication events (followed by evolutionary divergence) that can be dated (at least) in the ancestor of the genus *Burkholderia* (Perrin et al., BMC Evolutionary Biology *submitted for publication*).

1.2.5 Gene Elongation

It is generally accepted that ancestral protein-encoding genes should have been relatively short sequences encoding simple polypeptides likely corresponding to functional and/or structural domains. The size and complexity of extant genes are the result of different evolutionary processes, including gene fusion (see below), accretion of functional domains and duplication of internal motifs [Li & Graur, 1991; Lio' *et al.*, 2007]. The last mechanisms is often referred to as gene elongation, that is the increase in gene size, which represents one of the most important steps in the evolution of complex genes from simple ones. A gene elongation event can be the outcome of an in-tandem duplication of a DNA sequence. Then, if a deletion of the intervening sequence between the two copies occurs followed by a mutation converting the stop codon of the first copy into a sense codon

(Figure 1.9), this results in the elongation by fusion of the ancestral gene and its copy. Hence, the new gene is constituted by two paralogous moieties (modules). In principle, each module or both of them might undergo further duplication events, leading to a gene constituted by more repetitions of amino acid sequences. Many proteins of present-day organisms show internal repeats of amino acid sequences, and the repeats often correspond to the functional or structural domains [McLachlan, 1991]. This type of duplication has

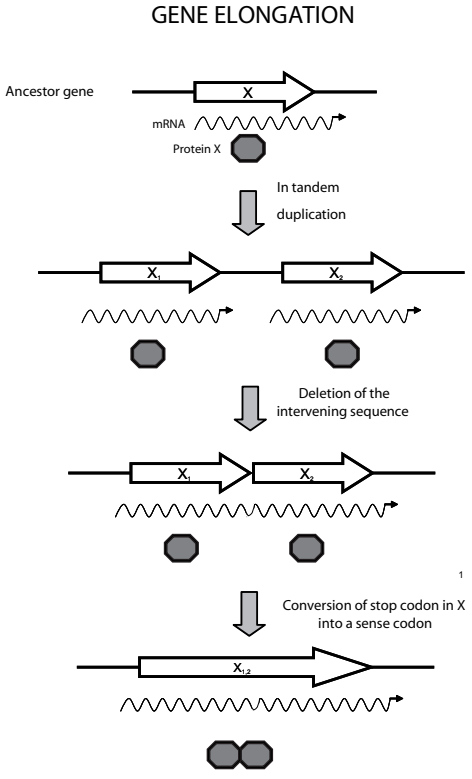


Figure 1.9: Representation of a gene elongation event.

occurred in so many proteins that the process must have considerable evolutionary advantage. The biological significance of gene elongation might rely in: (i) the improvement of the function of a protein by increasing the number of active sites and/or (ii) the acquisition of an additional function by modifying a redundant segment. Several examples of genes sharing internal sequence repetitions have been described in both prokaryotes and eukaryotes. For example the *Escherichia coli thrA*, *thrB* and *thrC* genes of the threonine biosynthetic operon, each shares a short module of about 35 amino acids [Cassan *et al.*, 1986; Parsot, 1986; Parsot *et al.*, 1983]. In another example the *carB* gene of *E. coli*, which specifies a subunit of carbamoylphosphate synthetase, shows an internal duplication of approximately half the size of the entire gene [Nyunoya & Lusty, 1983]. Gupta and Singh [1992] also described an internal repeat in the heat-shock protein 70 (HSP70) of archaea and bacteria. Rubin et al. [1990] found that the two domains of Gram negative bacterial tetracycline efflux proteins are encoded by genes that evolved by duplication of an ancestral module having half the size of the present-day gene(s). Moreover, previous

extensive analyses [Reizer & Saier, 1997] have illustrated the modular assembly and design of bacterial multidomain phosphoryl transferase proteins. The enormous variation in the arrangements of the subunits that has been observed in the ubiquitous ATP binding cassette (ABC) superfamily has led to the conclusion that domain fusions (together with duplication and insertion events) have occurred repeatedly during the evolution of the ABC superfamily [Reizer & Saier, 1997]. However, one of the most the most extensively documented example is represented by the pair of genes, *hisA* and *hisF*, showing an evident split into two modules half the size of the entire gene [Fani *et al.*, 1994]. In other cases, the traces of the common origin of two (or more) portion within a gene (as well as of two or more genes) can be disclosed by comparing the aminoacid sequence of the protein it codes for (see below for references).

1.2.6 Gene Fusion

In addition to gene duplication and gene elongation, one of the major routes of gene evolution is the fusion of independent cistrons leading to bi- or multifunctional proteins [Brilli & Fani, 2004]. Gene fusions provide a mechanism for the physical association of different catalytic domains or of catalytic and regulatory structures [Jensen, 1996]. Fusions frequently involve genes coding for proteins that function in a concerted manner, such as enzymes catalyzing sequential steps within a metabolic pathway [Yanai *et al.*, 2002]. Fusion of such catalytic centres likely promotes the channelling of intermediates that may be unstable and/or in low concentration [Jensen, 1996]; this, in turn, requires that enzymes catalyzing sequential reactions are co-localized within cell [Mathews, 1993] and may (transiently) interact to form complexes that are termed metabolons [Srere, 1987]. The high fitness of gene fusions can also rely on the tight regulation of the expression of the fused domains. Even though gene fusion events have been described in many prokaryotes, they may have a special significance among nucleated cells, where the very limited number, if not the complete absence, of operons does not a low the co-ordinate synthesis of proteins by polycistronic mRNAs. Fusions have been disclosed in genes of many metabolic pathways, such as tryptophan [Xie *et al.*, 2003] and histidine biosynthesis (see below).

1.2.7 The Role Of Horizontal Gene Transfer In The Evolution Of Genomes And Spreading Of Metabolic Functions

The Darwinian view of organism evolution predicts that such process can be interpreted and represented by a "tree of life" metaphor. Any functionally significant (phenotypic) and so selectable evolutionary "invention", arising from gene or genome level molecular processes (point mutations, gene duplication, etc.) is vertically transmitted - if not lethal. Nevertheless, there are exceptions to the tree of life paradigm (that, however, still provides a valid framework): evolutionary landmark events of cellular and genome evolution mediated by symbiosis (i.e. chloroplast and mitochondria) defines an example of non-linear evolution. Such processes defines a different model of evolution - the reticulate one

[Gogarten & Townsend, 2005] - that eventually took place along with the "classical" vertical transmission. Thus, a single bifurcating tree is insufficient to describe the microbial evolutionary process (that is furthermore problematical for the difficulty to define species boundaries in prokaryotes) as "only 0.1% to % of each genome fits the metaphor of a tree of life" [Dagan & Martin, 2006]. Indeed, the phylogenomic and comparative genomic approaches based on the availability of a large number of completely sequenced genomes has highlighted the importance of non-vertical transmission in shaping genomes and evolution processes. Incongruence existing in the phylogenetic reconstructions using different genes is considered as a proof of HGT events [Gribaldo & Brochier-Armanet, 2006; Ochman *et al.*, 2005], some of which probably (very) ancient [Brown, 2003; Huang & Gogarten, 2006]. The extent of HGT events occurred during evolution is still under debate [Dagan & Martin, 2006, 2007] and is especially intriguing in the light of early evolution elucidation as well as the notion of a communal ancestor [Koonin, 2003]. It has been in fact proposed that HGT dominated during the early stages of cellular evolution and was much higher than in modern systems [Woese, 1998, 2000, 2002]. The emergence of a "horizontal genomics" well explains the interest in the role of HGT processes in genome and species evolution. From a molecular perspective HGT is carried out by different mechanisms and is mediated by a mobile gene pool (the so called "mobilome") comprising plasmids, transposons and bacteriophages [Frost *et al.*, 2005]. HGT can involve single genes or longer DNA fragment containing entire operons and thus the genetic determinants for entire metabolic pathways conferring to the recipient cell new capabilities. It has been hypothesized that HGT does not involve equally genes belonging to different functional categories. Genes responsible for informational processes (transcription, translation, etc) are likely less prone to HGT than operational genes [Shi & Falkowski, 2008], even though the HGT of ribosomal operon has been described [Gogarten *et al.*, 2002]. This latter finding and the observation that only a 40% of the genes are shared by three *Escherichia coli* strains (Martin, 1999) raise the question of the stability of bacterial genomes [Itoh *et al.*, 1999; Mushegian & Koonin, 1996]. It is therefore important for phylogenetic and evolutionary analysis to individuate the "stable core" and the "variable shell" in prokaryotic genomes [Shi & Falkowski, 2008]. It is also quite possible that, in addition to HGT (xenology), the early cells might have exchanged (or shared) their genetic information through cell fusion (sinology). The latter mechanism might have been facilitated by the absence of a cell wall in the Archaean cells and might have been responsible for large genetic rearrangements and rapid expansion of genomes and metabolic activities. A summary of the evolutionary forces and mechanisms leading to the acquisition and spreading of novel metabolic traits is schematically reported in Figure 1.10

1.3 Origin and Evolution of Metabolic Pathways

1.3.1 The Primordial Metabolism

All living (micro)organisms possess an intricate network of metabolic routes for biosynthesis of the building blocks of proteins, nucleic acids, lipids and carbohydrates, and the

catabolism of different compounds to drive cellular processes. How these pathways have originated and evolved has been discussed for decades and is still under debate [Copley, 2000]. If we assume that life arose in a prebiotic soup containing most, if not all, of the necessary small molecules, then a large potential availability of nutrients in the primitive Earth can be surmised, providing both the growth and energy supply for a large number of ancestral organisms. Even though it is not still clear what were the properties of the ancestral organisms, i.e. if they possessed cell membranes and if most enzymes evolved prior to compartmentalization of environment into cells, it is plausible that those primordial organisms were heterotrophic and had no need for developing new and improved metabolic abilities since most of the required nutrients were available. If this scenario is correct, at least two questions can be addressed, that is *why* and *how* did primordial cells expand their metabolic abilities and genomes? The answer to the first question is rather intuitive. Indeed, the increasing number of early cells thriving on primordial soup would have led to the depletion of essential nutrients imposing a progressively stronger selective pressure that, in turn, favored those (micro)organisms that have become able to synthesize the nutrients whose concentration was decreasing in the primordial soup. Thus, the origin and the evolution of basic metabolic pathways represented a crucial step in molecular and cellular evolution, because it rendered the primordial cells less dependent on the

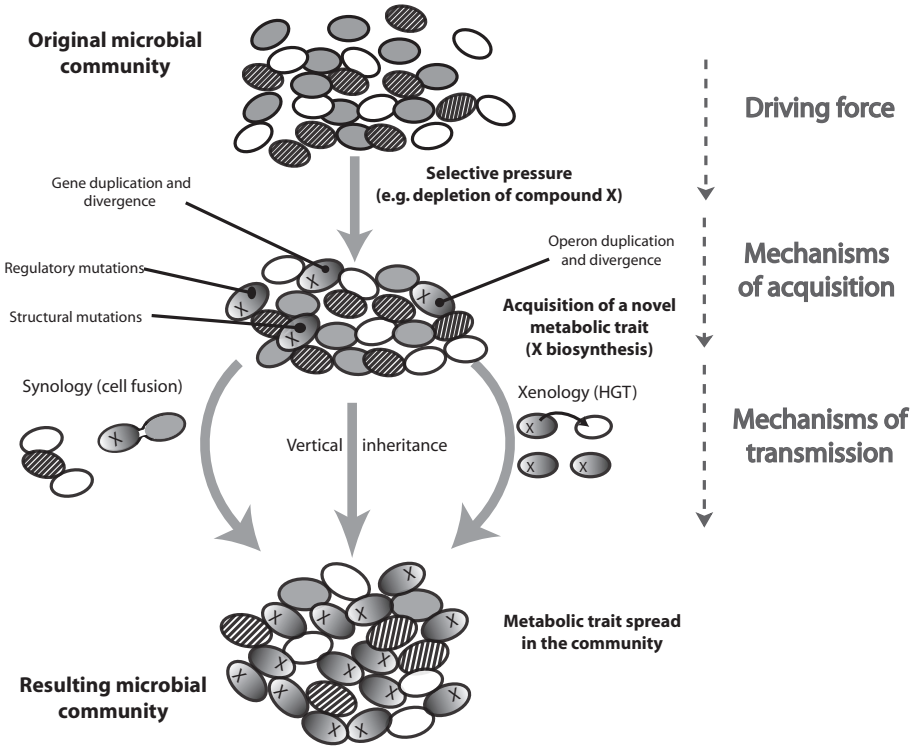


Figure 1.10: Schematic representation of an ancestral cell community with selective pressure allowing for the acquisition and spreading of a new metabolic trait (from [Fondi *et al.*, 2009])

external source of nutrients. Since ancestral cells probably owned small chromosomes and consequently possessed limited coding capabilities, it is plausible to imagine that their metabolism could count on a limited number of enzymes. This raises the question on how could the ancestral cells fulfill all their metabolic tasks possessing such a restricted enzyme repertoire? A possible (and widely accepted) explanation is that these ancestral enzymes possessed broad substrate specificity, allowing them to catalyze several different chemical reactions (see below). Hence, the hypothetical ancestral metabolic network was probably composed by a limited number of nodes (enzymes) that were highly interconnected (i.e., participated in different, although linked, biological processes). On the contrary, network models of extant metabolisms reveal remarkably complex structures (Figure 1.11); thousands of different enzymes form well defined routes that transform many distinct molecules, in an ordered fashion and with a predefined output. The next session will focus on the molecular mechanisms that guided this transition, i.e. the expansion and the refinement of ancestral metabolic routes, leading to the structure of the extant intertwined metabolic pathways.

1.3.2 Mechanisms for metabolic pathways assembly

As discussed in the previous sections, the emergence and refinement of basic biosynthetic pathways allowed primitive organisms to become increasingly less dependent on exogenous sources of amino acids, purines, and other compounds accumulated in the primitive environment as a result of prebiotic syntheses. But how did these metabolic pathways originate and evolve? Then, which is the role that the molecular mechanisms described above (gene elongation, duplication and/or fusion) played in the assembly of metabolic routes? How the major metabolic pathways actually originated is still an open ques-

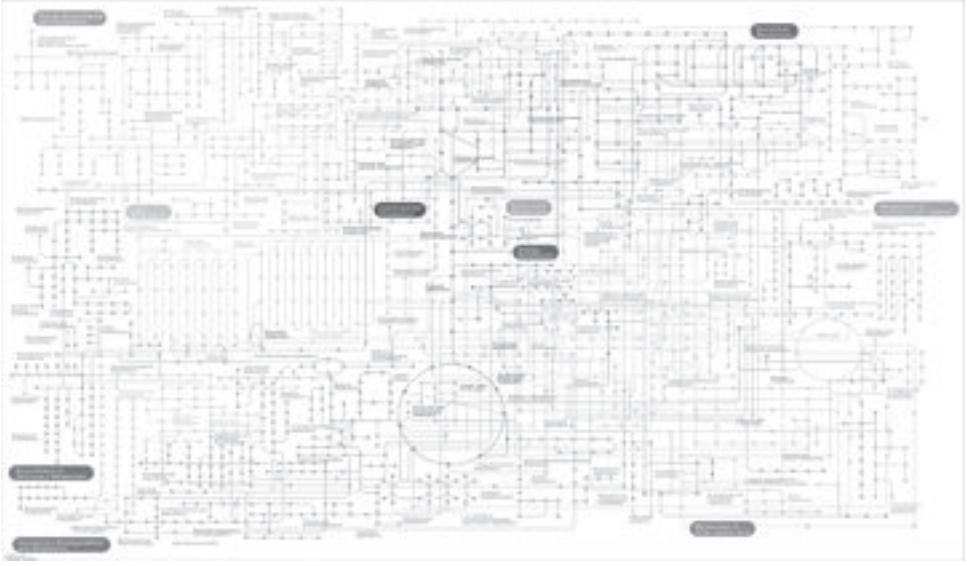


Figure 1.11: Global metabolism map (from www.genome.jp/kegg)

tion, but several different theories have been suggested to account for the establishment of metabolic routes. As we will see, gene duplication plays a major role in all of these ideas.

1.3.2.1 The Retrograde hypothesis (Horowitz, 1945)

The first attempt to explain in detail the origin of metabolic pathways was made by Horowitz [Horowitz, 1945], who based this on two pieces of work. The first was the primordial soup hypothesis and the second was the one-to-one correspondence between genes and enzymes noticed by Beadle and Tatum [Beadle & Tatum, 1941]. Horowitz suggested that biosynthetic enzymes had been acquired via gene duplication that took place in the reverse order found in current pathways. This idea, also known as the Retrograde hypothesis, has intuitive appeal and states that if the contemporary biosynthesis of compound A requires the sequential transformations of precursors D, C and B via the corresponding enzymes, the final product A of a given metabolic route was the first compound used by the primordial heterotrophs (Figure 1.12). In other words, if a compound A was essential

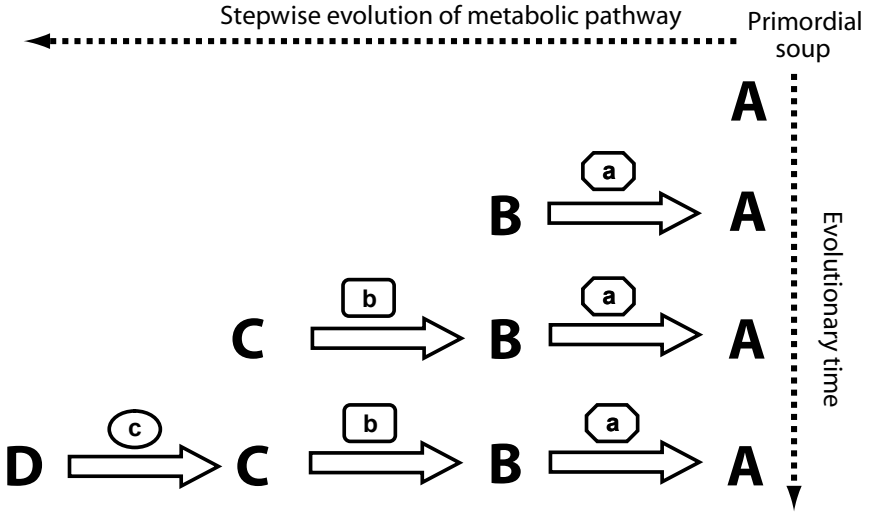


Figure 1.12: Schematic representation of the Horowitz hypothesis on the origin and evolution of metabolic pathways (from [Fondi *et al.*, 2009]).

for the survival of primordial cells, when A became depleted from the primitive soup, this should have imposed a selective pressure allowing the survival and reproduction of those cells that were become able to perform the transformation of a chemically related compound B into A catalyzed by enzyme a that would have lead to a simple, one-step pathway. The selection of variants having a mutant b enzyme related to a via a duplication event and capable of mediating the transformation of molecule C chemically related into B, would lead into an increasingly complex route, a process that would continue

until the entire pathway was established in a backward fashion, starting with the synthesis of the final product, then the penultimate pathway intermediate, and so on down the pathway to the initial precursor (Figure 1.12). Twenty years later, the discovery of operons prompted Horowitz to restate his model, arguing that it was supported also by the clustering of genes, that could be explained by a series of early tandem duplications of an ancestral gene; in other words, genes belonging to the same operon and/or to the same metabolic pathway should have formed a paralogous gene family. The retrograde hypothesis establishes a clear evolutionary connection between prebiotic chemistry and the development of metabolic pathways, and may be invoked to explain some routes. However, The evolution of metabolic pathways in a backward direction requires special environmental conditions in which useful organic compounds and potential precursors have accumulated. Although these conditions might have existed at the dawn of life, they must have become less common as life forms became more complex and depleted the environment of ready-made useful compounds [Copley, 2000]. Furthermore, the origin of many other anabolic routes cannot be understood in terms of their backwards development as they involve many unstable intermediates and it is difficult to explain their synthesis and accumulation in both the prebiotic and extant environments. In addition to this, many of these metabolic intermediates are phosphorylated compounds that could not permeate primordial membranes in the absence of specialized transport systems that were probably absent in primitive cells [Lazcano *et al.*, 1995]. It has been also argued that the Horowitz hypothesis fails to account for the origin of catabolic pathway regulatory mechanisms, and for the development of biosynthetic routes involving dissimilar reactions. In addition to this, if the enzymes catalyzing successive steps in a given metabolic pathway resulted from a series of gene duplication events [Horowitz, 1965], then they must share structural similarities [Hegeman & Rosenberg, 1970]. Even though there is a handful of examples where adjacent enzymes in a pathway are indeed homologous [Belfaiza *et al.*, 1986; Bork & Rohde, 1990; Fani *et al.*, 1994, 2000; Wilmanns *et al.*, 1991], the list of known examples confirmed by sequence comparisons is small. Maybe the most extensively documented examples pertain to the pair of genes *hisA* and *hisF* [Fani *et al.*, 2000] and four of the genes involved in nitrogen fixation (*nifD*, *K*, *E*, and *N*) [Fani *et al.*, 1994](see the relative sections).

1.3.3 The Granick hypothesis

An alternative, although less-well known, proposal is the development of biosynthetic pathways in the forward direction [Granick, 1957, 1965], where the prebiotic compounds do not play any role. Granick proposed that the biosynthesis of some end-products could be explained by forward evolution from relatively simple precursors (see [Pereto *et al.*, 2000] and references therein). This model predicts that simpler biochemical compounds predated the appearance of more complicated ones; hence, the enzymes catalyzing earlier steps of a metabolic route are older than the latter ones. For this to operate it is necessary for each of the intermediates to be useful to the organism, since the development of

multiple genes simultaneously in a sequence is too improbable [Lazcano & Miller, 1996; Pereto *et al.*, 2000]. This might work with heme and chlorophyll as cited by Granick, but problems arise with pathways such as purine and branched chain amino acid syntheses, where the intermediates are of no apparent use. Another example where the Granick proposal has been applied is the development of the isoprene lipid pathway [Ourisson & Nakatani, 1994].

1.3.3.1 The Patchwork hypothesis (Ycas, 1974; Jensen, 1976)

Gene duplication has also been invoked in another theory proposed to explain the origin and evolution of metabolic pathways, the so-called patchwork hypothesis [Jensen, 1976; Ycas, 1974] according to which metabolic pathways may have been assembled through the recruitment of primitive enzymes that could react with a wide range of chemically related substrates. Such relatively slow, non-specific enzymes may have enabled primitive cells containing small genomes to overcome their limited coding capabilities (Figure 1.13). Figure 1.13 shows a schematic three-step model of the patchwork hypothesis;

1. the ancestral enzyme E1 endowed with low substrate specificity is able to bind to three substrates (S1, S2 and S3) and catalyze three different, but similar reactions;

2. a paralogous duplication of the gene encoding enzyme E1 and the subsequent divergence of the new sequence lead to the appearance of enzyme E2 with an increased and narrowed specificity;

3. a further duplication event occurred leading to E3 showing a diversification of function and narrowing of specificity.

In this way the ancestral enzyme E1, belonging to a given metabolic route is recruited to serve other novel pathways. The patchwork hypothesis is also consistent with the possibility that an ancestral pathway may have had a primitive enzyme catalyzing two or more similar reactions on related substrates of the same metabolic route and whose substrate specificity was refined as a result of later duplication events. In this way primordial cells might have expanded their metabolic capabilities. Additionally, this mechanism may have permitted the evolution of regulatory mechanisms coincident with the development of new pathways [Fani, 2004; Lazcano *et al.*, 1995]. Related to this view is that in which enzyme evolution has been driven by retention of catalytic mechanisms [Copley & Bork, 2000]. There is good evidence to suggest that this has occurred within many protein families [Babbitt & Gerlt, 1997; Eklund & Fontecave, 1999; Gerlt & Babbitt, 1998; Lawrence *et al.*, 1997]. The patchwork hypothesis is supported by several lines of evidence. The broad substrate specificity of some enzymes means they can catalyze a class of different chemical reactions and this provides a support for the patchwork theory. As demonstrated by whole genome sequence comparisons, there is a significant percentage of metabolic genes that are the outcome of paralogous duplications described in completely sequenced cellular

genomes. Sequence comparisons of enzymes catalyzing different reactions in the biosynthesis of threonine, tryptophan, isoleucine and methionine indicate that each protein has evolved from a single common ancestral molecule active in several metabolic pathways [Lazcano *et al.*, 1992]. The recruitment of enzymes belonging to different metabolic pathways to serve novel biosynthetic routes is well documented under laboratory conditions. These are the so-called directed evolution experiments, in which microbial populations are subjected to a strong selective pressure leading to heterettophic phenotypes capable of using new substrates (see below). Some fascinating examples of Natures opportunism in assembling new pathways using this patchwork approach have been found [Copley, 2000]. The urea cycle in terrestrial animals clearly evolved by addition of a new enzyme, arginase, to a set of four enzymes previously involved in the biosynthesis of arginine [Takiguchi *et al.*, 1989]. The Krebs cycle is postulated to have evolved by combination of several pre-existing enzymes from pathways for biosynthesis of aspartate and glutamate with four additional enzymes [Copley, 2000; Melendez-Hevia *et al.*, 1996]. Besides, some ancestral biosynthetic routes, such as histidine [Fani *et al.*, 1995, 1998] and tryptophan [Xie *et al.*, 2003] biosynthesis, nitrogen fixation [Fani *et al.*, 2000], as well as lysine, arginine and leucine [Fondi *et al.*, 2007] were highly likely assembled through this mechanism. However, there are also very nice examples of recent adaptation to completely newly compounds by the patchwork mechanism. This is particularly true for metabolic pathways

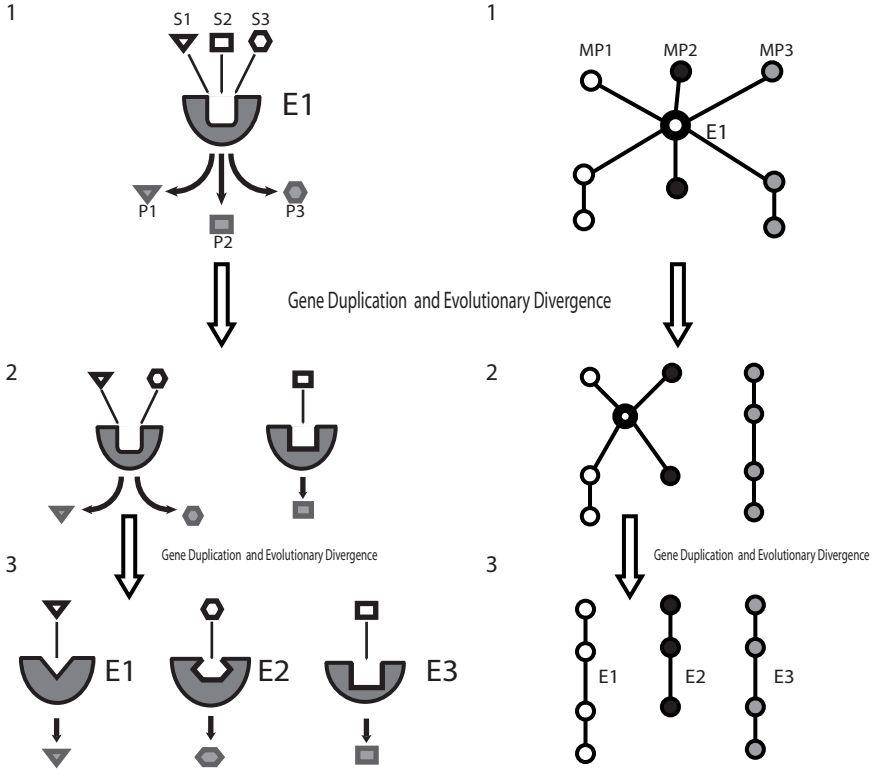


Figure 1.13: Schematic representation of the Jensen hypothesis on the origin and evolution of metabolic pathways (from [Fondi *et al.*, 2009]).

evolved by microorganisms in order to either exploit new carbon sources or detoxify toxic compounds, such as xenobiotic chemicals. One of the most striking examples is the evolution of the pathway for degradation of pentachlorophenol (PCP), a xenobiotic pesticide, in *Sphingomonas chlorophenolica*, which has been suggested to be the outcome of the patchwork combination of enzymes from two different existing pathways [Copley, 2000].

1.3.3.2 Semienzymatic origin of metabolic pathways (Lazcano and Miller, 1996)

In order to explain the origin of the very early metabolic pathways, Lazcano and Miler [Lazcano & Miller, 1996] proposed a different approach that may be applicable to the origin of some but not all metabolic routes. They based their idea on the following assumptions: (a) a set of rather stable prebiotic compounds was available in the primitive ocean. (b) Compounds due to leakage from existing pathways within cells were also available. These compounds need not be particularly stable because they are produced within the cell and used rapidly. (c) Existing enzyme types are assumed to be available from gene duplication and they were non-specific according to Jensen [Jensen, 1996]. (d) Starter-type enzymes are assumed to arise by non-enzymatic reactions followed by acquisition of the enzyme. It is known that most steps in biosynthetic routes are mediated by enzymes, but some occur spontaneously. In other cases the corresponding chemical step can be achieved by changing the reaction conditions and reagents in the absence of the enzyme. An example is the product of the G-type glutamine amidotransferase gene (*hisH*), which takes part in histidine biosynthesis. The reaction adds NH_3 under high ammonia concentrations in the absence of the HisH protein [Martin, 1971]. Experimental evidence has demonstrated prototrophic growth under high ammonia concentrations of a *Klebsiella pneumoniae* strain with a mutated *hisH* gene Rieder *et al.* [1994]. Lazcano and Miller [Lazcano & Miller, 1996] propose that the reaction first took place with NH_3, followed by the development of HisH, followed in turn by the substitution of glutamine or NH3 as this compound disappeared from the prebiotic soup.

1.3.4 Origin and Evolution of Operons

As mentioned above, changes in gene structure across time have greatly affected the assembling and the refinement of (entire) metabolic routes. However, gene organization, that is, the order of genes along the chromosome(s), has also played a pivotal role in metabolism evolution. This idea has been reinforced by the observation(s) that, at least in the microbial world, an important percentage of genes participating in the same biosynthetic route are organized in an operon fashion [Omelchenko *et al.*, 2003]. The term operon was first introduced to define a group of genes regulated by an operator and transcribed into a polycistronic mRNA [Jacob & Monod, 1961]. The same term is now used to describe any group of adjacent genes that are transcribed from a promoter into a polycistronic mRNA. In the last decades, studies were focused mainly on two operons, i.e. trypto-

phan [Xie *et al.*, 2003] and histidine [Alifano *et al.*, 1996], which helped to reveal new and sophisticated mechanisms of transcription control (i.e. attenuation [Alifano *et al.*, 1996]). In addition, the finding that genes belonging to the same metabolic pathway were organized in similar operons in distantly related organisms [de Daruvar *et al.*, 2002] suggested that clustering of genes involved in a biosynthetic route was a common feature of prokaryotic genomes, leading to the idea that operon assembly predated the LUCA and that such genes clusters were vertically or horizontally transferred during evolution. However, the comparative analysis of fully sequenced genomes has challenged the original view of operon structure, origin and evolution [Itoh *et al.*, 1999; Makarova *et al.*, 2001; Price *et al.*, 2005b, 2006; Wolf *et al.*, 2001], disclosing the possibility that some operons might be a recent invention of evolution [Fani *et al.*, 2005] and sometimes the result of convergent evolution (see below).

1.3.4.1 Distribution and Structure of Operons

Operons are widespread in prokaryotes [Fani *et al.*, 2005; Itoh *et al.*, 1999; Langer *et al.*, 1995; Price *et al.*, 2005b, 2006] and represent one of the main strategies of gene organization and regulation in prokaryotes [Omelchenko *et al.*, 2003]. In eukaryotes, gene clusters are very rare (often reflecting multiple alleles of a single cistron), although several *Caenorhabditis elegans* genes appear to be co-transcribed in clusters resembling operons. However, it is not yet clear whether gene clusters arose in this genus independently or were already present in the ancestor of all eukaryotes. It has been estimated that in a typical prokaryotic genome, about half of the protein-coding genes are organized in operons [Price *et al.*, 2006]. However, in spite of the idea that the proximity of functional related genes offers more efficient regulation [Demerec & Demerec, 1956], allowing the maintenance of operon organization during evolution by purifying selection [Rocha, 2006], operon conservation among prokaryotes seems to be far less common than expected [Omelchenko *et al.*, 2003]. Indeed, prokaryotic genomes are rather unstable [Itoh *et al.*, 1999; Mushegian & Koonin, 1996; Watanabe *et al.*, 1997] and only 5,25% of genes belong to strings (probably operons) shared by at least two distantly related species [Wolf *et al.*, 2001]. This suggests that operon conservation might be neutral over evolutionary time [Itoh *et al.*, 1999], even though operon disruption should decrease the transcriptional efficiency and hence reduce cell fitness. Moreover, the very same operon organization in distantly related organisms is strongly maintained only for few key genes coding for physically interacting proteins [Dandekar *et al.*, 1998; Huynen *et al.*, 2000; Itoh *et al.*, 1999; Mushegian & Koonin, 1996; Watanabe *et al.*, 1997], such as the ribosomal proteins, proton ATPases and ABC membrane transport cassettes [Wolf *et al.*, 2001]. In the original definition proposed by Jacob and Monod [Jacob & Monod, 1961], operons contain genes belonging to the same functional pathway [de Daruvar *et al.*, 2002; Rogozin *et al.*, 2002] in order to guarantee them a similar expression level. The analysis of several fully sequenced genomes is challenging this idea. First, some genes without apparent functional relationships, that is, alien [Papaleo *et al.*, 2009] (genes apparently not involved in the same metabolic route and having homologs in other species) or ORFan genes (lacking homologs in closely re-

lated species and probably acquired from bacteriophages) [de Daruvar *et al.*, 2002; Price *et al.*, 2005a, 2006], can be embedded in the same operon. The biological significance of this finding might in some cases rely on the requirement for coordinate regulation of the (apparently unrelated) genes by the same environmental stimulus(i) [Price *et al.*, 2006], but in other cases it remains obscure. Secondly, the discovery of regulons [Maas, 1964] (sets of functionally related genes scattered throughout the genome that can be efficiently co-regulated) revealed that different gene organization strategies may assure similar expression patterns [Price *et al.*, 2006; Sabatti *et al.*, 2002]. Thirdly, many operons do not have a structure consistent with the original definition, since they are under the control of multiple promoters and/or regulators [Price *et al.*, 2006; Vicente *et al.*, 1998]. Lastly, operons exhibit a different degree of compactness. Overall, genes within operons are separated by less than 20 base pairs [Eyre-Walker, 1995; Moreno-Hagelsieb & Collado-Vides, 2002] and often overlap [Eyre-Walker, 1995], because of biases of bacterial genomes towards small deletions [Mira *et al.*, 2001] and/or for translational coupling [Yu *et al.*, 2001]. However, wide spacing exists even in highly expressed operons [Ma *et al.*, 2002; Moreno-Hagelsieb & Collado-Vides, 2002] and this is often related to the presence of internal promoters [Price *et al.*, 2006]. Thus, the operon structure appears to be more heterogeneous than previously thought.

1.3.4.2 Hypothesis on the Origin and Evolution of Operon

Despite the large body of data available regarding structure, distribution and conservation, the biological significance and mechanism(s) of operon formation are still under debate. At least seven different models have been proposed for operon formation.

1. The Natal model predicts that operons originated by in situ gene duplication and divergence, whereby the evolution of metabolic pathways took place in a stepwise fashion in an assembly line of genes [Itoh *et al.*, 1999]. This model corresponds to the Horowitz idea on the origin and evolution of metabolic pathways and was supported by the observation that gene order in some operons (such as the *trp*-operon) reflects e in some organisms e the corresponding biochemical reactions. Even though examples of gene duplication and divergence inside an operon have been reported [Fani *et al.*, 1994, 2000], the low conservation of operon structure and of gene order and the lack of homology between operon genes challenged this model.

2. The Fischer model proposes that the physical proximity of co-adapted alleles in the genome reduces the frequency of the formation of unfavorable combinations of genes by recombination events. This might favor operon assembly.

3. Glansdorff [Glansdorff, 1999] suggested that early adaptation to thermophily played a key role in the emergence of operons. This is supported by transcriptionetranslation coupling, which is seen as a mechanism capable of protecting messenger RNA from degradation caused by high temperatures.

4. The co-regulation model predicts that genes are clustered together because regulation is easier under a single promoter, providing both economy of transcription and equal abundance of products, especially when genes belong to the same metabolic pathway. Operon organization should therefore be the most economical means of regulation, preferred by selection over gene scattering. However, genes organized in regulons can also be co-regulated and operons can also contain functionally unlinked genes. Moreover, it has been argued that co-regulation can provide a selective advantage for operon structure maintenance, but not for gene clustering, since no fitness beneficial effects are expected during operon formation (progressive increase in gene proximity) until co-transcription is possible. Lastly, co-regulation might be more easily obtained by modifying two promoters than by placing two genes in proximity, since the likelihood of rearranging two genes in the correct position is very low [Lawrence, 1999; Lawrence & Roth, 1996].

5. In the molarity model, co-regulation can also guarantee that proteins are synthesized in equimolar amounts, reducing stochastic differences in their concentration levels [Swain, 2004], and can increase the rate of both formation and folding of multisubunit protein complexes [Dandekar *et al.*, 1998; Pal & Hurst, 2004]. However differences in (i) the efficiency of activity of different enzymes, (ii) the efficiency of translation of genes within an operon, and/or (iii) the half-life of different mRNAs can reduce the probability of equimolar production of proteins. Furthermore, even though some highly conserved operons code for protein complexes [Dandekar *et al.*, 1998], the majority of them do not and, vice versa, many protein complexes are not encoded by genes within the same operon [Butland *et al.*, 2005].

6. The selfish operon theory [Lawrence, 1999; Lawrence & Roth, 1996] posits that operons, except for some highly conserved operons, thought to be ancient and to have been formed by other mechanisms, form because such compact organization facilitates HGT (and so survival) of non-essential gene clusters, whose function is only occasionally useful and so prone to random deletion of genes by mutation pressure and genetic drift. HGT would save the cluster from extinction and might confer selective advantage(s) to the recipient organism in some environmental conditions. It might also drive cluster compacting by deletion of intervening DNA stretches in recipient cells, since compactness enhances the HGT probability even without the benefit of co-regulation. Co-regulation can evolve later in the recipient cell by the presence of a promoter in the site of insertion, providing new abilities to the organism and acting on operon conservation. The selfish model is consistent with the compactness of operons and with the finding that operons are horizontally transferred [Hazkani-Covo & Graur, 2005; Lawrence & Roth, 1996; Omelchenko *et al.*, 2003; Price *et al.*, 2006], but does not explain why genes for key functions are found in operons and why operons contain genes coding for unrelated functions. Moreover, essential and non-HGT genes are generally likely to be found in operons [Gerdes *et al.*, 2003; Pal & Hurst, 2004; Price *et al.*, 2005b]. Furthermore, since non-HGT genes form new operons (often containing genes that are apparently functionally

unrelated) [Price *et al.*, 2005b], it has been suggested that HGT acts on the distribution of some operons or on the modification of preexisting ones [Omelchenko *et al.*, 2003], but that it is not the driving force in operon formation [Price *et al.*, 2005a, 2006]. Therefore, operon formation could be driven by gene clustering due to rearrangements and deletions in order to facilitate co-regulation, since complex regulation is more easily reachable by the evolution of one complex promoter than by the evolution of different promoters [Price *et al.*, 2005a, 2006].

7. More recently, a new idea, referred to as the piece-wise model, was proposed to explain the origin and evolution of some operons [Fani *et al.*, 2005] 1.14. According

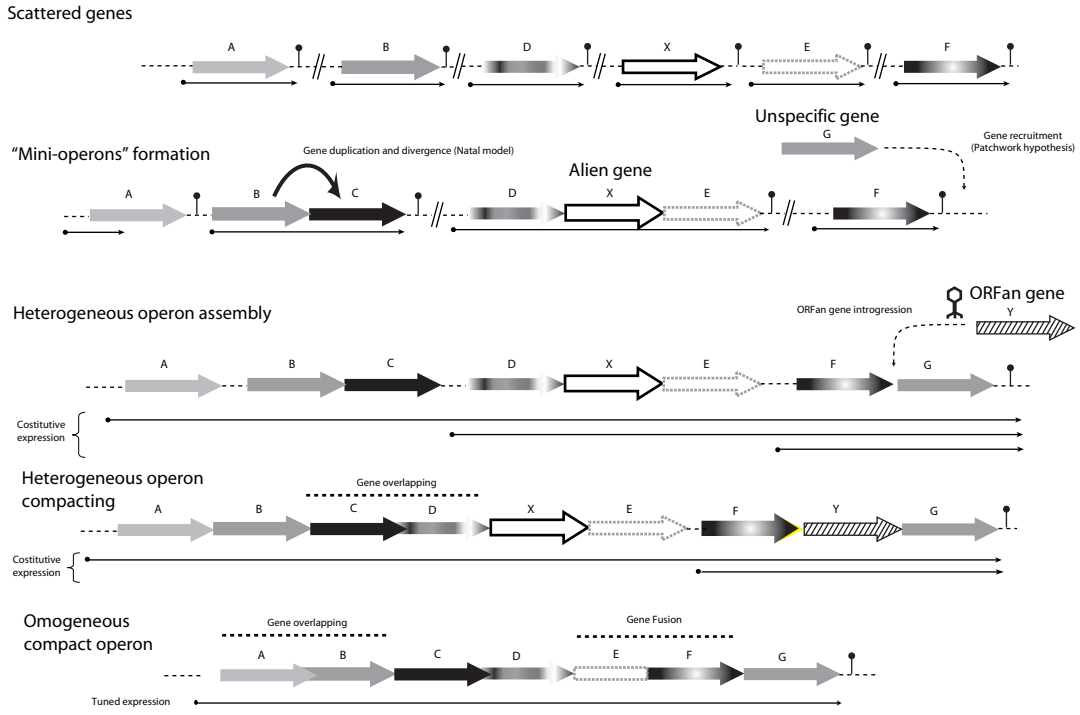


Figure 1.14: Schematic representation of the "piece-wise" model for operon assembly (from [Fondi *et al.*, 2009]).

to this model, long and complex operons can be assembled through progressive clustering of pre-existing suboperons embedding part of the genes of the final, completely assembled operon (Figure 1.14). Even though the model was originally suggested to explain the origin and evolution of the proteobacterial histidine operon [Fani *et al.*, 2005], it might be applied to the origin and evolution of any complex operon. The assembly of scattered genes into sub-operons might proceed through different mechanisms. According to Horowitz [Horowitz, 1945], in-tandem duplication of ancestral genes may lead to bi- or multicistronic operons, while other genes could be recruited via the patchwork mechanism and put close to other genes via recombination or transposition. These sub-operons might be evolutionary fixed by

different forces: the necessity of equimolarity and/or co-regulation or the formation of metabolon-like structures. The model also implies that the construction of compact (homogeneous) operons might proceed through the progressive clustering of sub-operons (parallel to the shortening of the intergenic sequences) that implies intermediate stages represented by heterogeneous operons, which might also include ORFan and/or alien genes, and intergenic sequences possibly containing transcription promoters. Since the heterogeneous operon contains alien and/or ORFan genes, its expression might be under the control of different stimuli and, thus, might be constitutively transcribed [Papaleo *et al.*, 2009]. The final step in construction of homogeneous operons should be the elimination of ORFan and/or alien genes, the shortening of intergenic sequences (with the possible overlap or fusion of some genes) and refinement of the regulator signals controlling operon expression. The piece-wise model is also consistent with the possibility that, at different stages of operon compacting, genes involved in different metabolic pathways can be recruited and specialized by introgressing the heterogeneous or homogeneous operon itself. A paradigmatic example of such a construction is represented by the proteobacterial histidine biosynthetic operon [Fani *et al.*, 2005].

1.3.4.3 A dynamic view of operon life

It is well established that some operons are highly conserved and vertically inherited [Omelchenko *et al.*, 2003; Overbeek *et al.*, 1999; Wolf *et al.*, 2001]. However, such stability in operon organization is relatively rare [Dandekar *et al.*, 1998; Fani *et al.*, 2005; Itoh *et al.*, 1999; Omelchenko *et al.*, 2003] and it is not much higher than in non-operon genome regions [Itoh *et al.*, 1999]. These findings suggest a highly dynamic view of operon formation and evolution Figure 1.15. As proposed by Price et al. [Price *et al.*, 2005b, 2006], the formation of new operons involving native, HGT, ORFan and/or alien genes can occur at quite a high rate [Daubin & Ochman, 2004; Fischer & Eisenberg, 1999]. The same molecular mechanisms driving operon formation (rearrangements, deletions and HGT insertion with consequent splitting of the operon, or gene displacement) may also be responsible for operon death (Figure 1.15). Interestingly, but not surprisingly, new operons as well as operons containing functionally unrelated genes are more prone to be lost [Price *et al.*, 2006]. Existing operons can also be modified, even if at a lower rate than the formation of new operons, and some operons show a rapid evolution for addition of new genes at the end or at the beginning of the pre-existing operon [Price *et al.*, 2006]. Operons can also be modified by in situ xenologous displacement of genes by HGT within the resident operon; this mechanism can lead to the formation of mosaic operons [Omelchenko *et al.*, 2003] that can also be the outcome of de novo assembly of native or HGT or ORFan genes [Price *et al.*, 2006]. Lastly, operon duplication can lead to the appearance of paralogous operon families, increasing the overall number of operons within a genome [Fondi *et al.*, 2009].

1.4 The Reconstruction of the Origin and Evolution of Metabolic Pathways

How can the origin ad evolution of metabolic pathways be studied and reconstructed? By assuming that useful hints may be inferred from the analysis of metabolic pathways existing in contemporary cells, important insights of the evolutionary development of microbial metabolic pathways can be obtained by:

1. laboratory studies in which new substrates are used as carbon, nitrogen, or energy sources. These are the so-called directed-evolution experiments, in which a microbial (typically, bacterial) population is subjected to a (strong) selective pressure that leads to the establishment of new phenotypes capable of exploiting different substrates [Clarke, 1974; Mortlock & Gallo, 1992]. By assuming that the processes involved in acquiring new metabolic abilities are comparable to those found in natural populations, directed-evolution experiments can provide useful insights in early cellular evolution [Fani, 2004].

2. The use of bioinformatic tools, which allow the comparison of gene and genomes from organisms belonging to the three cell domains (Archaea, Bacteria and Eukarya). This approach takes advantage of the availability of the phylogenetic relationships among (micro)organisms, and possibly on the existence of different structure and organization exhibited by orthologous genes. Beside, the more ancient is a pathway, the more information can be retrieved from this comparative analysis

Data presented in this dissertation were obtained adopting this second approach.

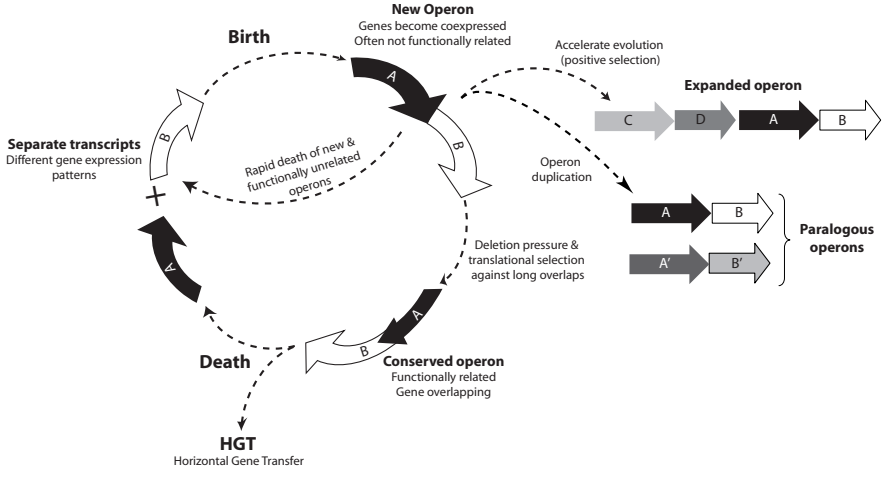


Figure 1.15: The life cycle of operon (from [Fondi *et al.*, 2009], modified from [Price *et al.*, 2006]).

1.5 Bioinformatics of Genomes Evolution

Recent years saw a dramatic increase in genomics data deriving from organisms belonging to all of the three known domains of life (Figure 1.16). By the way, the use of bioinfor-

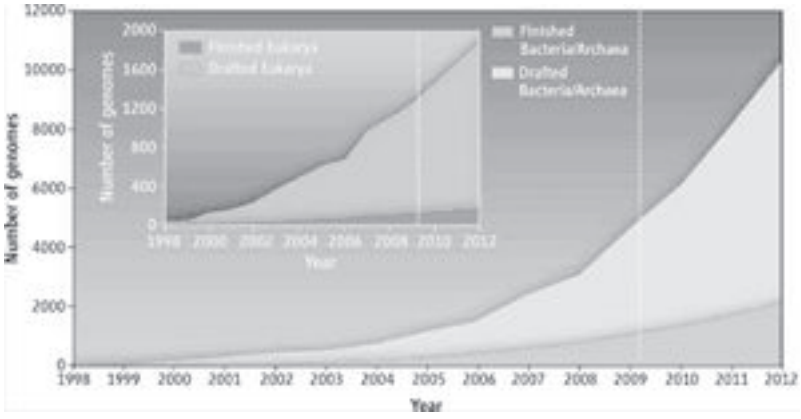


Figure 1.16: Trends in generation of drafted and finished genomes. A conservative estimate of future projects is shaded in light blue. Taken from [Chain *et al.*, 2009]).

matic tools allowed the storage and the interpretation of several sources of information (gene structure and organization, gene regulation, proteinprotein interactions) and, probably more importantly, their integration, a fundamental step for the global understanding of genomes properties and dynamics. This approach is usually referred to as comparative genomics. Combining data gained from comparative genomics with evolutionary studies of different species (i.e. phylogenetic inference), results in a new kind of approach, referred to as phylogenomics. This novel way of investigating the evolutionary history of genes introduced several advantages, in fact, adopting a genome-scale approach theoretically overcomes incongruence derived from molecular phylogenies based on single genes mainly because (i) non-orthologous comparison (i.e. the comparison of those genes erroneously defined as orthologous) is much more misleading when the analysis is performed on a single gene, whereas it is probably buffered in a multigene analysis and (ii) stochastic error naturally vanishes when more and more genes are considered. Next sessions will deal with some different methods and techniques that, taken together, allow a comparative genomics and phylogenomics approaches.

1.5.1 Browsing Microbial Genomes

At present, hundreds of microbial genomes have been sequenced, and hundreds more are currently in the pipeline. Furthermore, functional genomic studies have generated a large and growing body of experimental results for many different organisms belonging to the known domains of life. However, this whole body of data would reveal almost useless if not stored in a proper manner. To this purpose a growing number of public databases have

been developed in recent years, usually providing also user-friendly tools for their interrogation. These tools, despite not allowing automatized large-scale phylogenomic analyses, often represent their first preliminary (and useful) step. This is the case for example of MicrobesOnLine (http://www.microbesonline.org, [Alm *et al.*, 2005; Dehal *et al.*, 2009]) which embeds both structural and functional data on a large (almost 3000) dataset of completely sequenced genomes. These data are retrieved from a wide range of other spe-

Figure 1.17: Output of MicrobesOnLine webserver when probed with "HisF" text search.

cific databases (including KEGG, GeneOntology, RefSeq). Interestingly, MicrobesOnLine also allows to interactively explore the neighborhood of any given gene, hence allowing, for example, a first analysis of the gene organization of a given metabolic pathway (Figure 1.17). The same task can be pursued adopting also operonDB web service (http://odb.kuicr.kyoto-u.ac.jp/, [Pertea *et al.*, 2009]) aiming at collecting all known operons (derived from the literature and from publicly available database) in multiple species and to offer a system to predict operons by user definitions. Several other web sites and software tools have been described that assist in the annotation and exploration of comparative genomic data. The Prolinks [Bowers *et al.*, 2004] and STRING [Jensen *et al.*, 2009] databases offer convenient tools for browsing predicted functional associations among proteins. String, in particular (Figure 1.18), imports protein association knowledge not only from databases of physical interactions, but also from databases of curated biological pathway knowledge. A number resources are included in the current release (MINT [Ceol *et al.*, 2009], HPRD [Keshava Prasad *et al.*, 2009], DIP [Xenarios *et al.*, 2002], BioGRID [Stark *et al.*, 2006], KEGG [Kanehisa & Goto, 2000] and Reactome [Matthews *et al.*, 2009] IntAct [Hermjakob *et al.*, 2004], EcoCyc [Keseler *et al.*, 2009]). Furthermore, this set of previously known and well-described interactions is then complemented by interactions that are predicted computationally, specifically for STRING, using a number of prediction algorithms [Jensen *et al.*, 2009].

1.5.2 Orthologs Identification

Genomics data is a fundamental step for addressing the topic of the evolution of metabolic pathways, and strictly depends on a correct identification of orthologous proteins shared

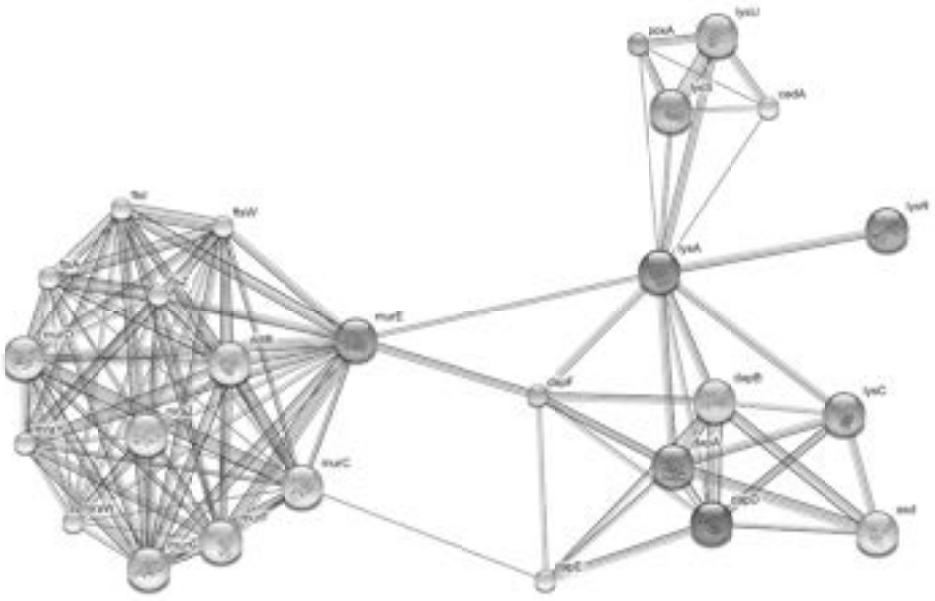


Figure 1.18: Output of String webserver when probed with "LysA" text search.

by different genomes. This field has been greatly developed in recent years and, paradoxically, the extant challenge seems not to be the lack of orthology predictions, but the right choice within the plethora of methods and databases that have been recently implemented [Gabaldon *et al.*, 2009]. The identification of orthologs between two genomes often relies on the so-called bidirectional best-hit (BBH) criterion, a reiteration of the BLAST algorithm [Altschul *et al.*, 1997]: two proteins, **a** and **b**, from genomes **A** and **B** respectively, are orthologs if **a** is the best-hit (i.e. the most similar) of **b** in genome **A** and *vice versa*. For three or more genomes, groups of orthologous sequences can be constructed by extending the BBH relationships with a clustering algorithm. This approach has led to the assembly of pre-compiled databases embedding groups of orthologous proteins, such as COG or KEGG-related systems (KOBAS and KAS). Moreover several others algorithms have been developed to fulfill this tasks, including Ncut [Abascal & Valencia, 2002], Rio, [Zmasek & Eddy, 2002], Outgroup Conditioned Score (OCS) [Cotter *et al.*, 2002] or OrthoParaMap [Cannon & Young, 2003]. Recent advancements showed that clustering techniques applied to matrices storing pair-wise similarities perform quite well [Brilli *et al.*, 2008]. These algorithms work either on the grouping of weakly similar homologs or on the identification of protein domains. The most widespread are: i) orthoMCL [Li *et al.*, 2003] which adopts a Markov Clustering algorithm (previously implemented in tribeMCL [Enright *et al.*, 2002]), Ortholuge [Fulton *et al.*, 2006] that aims at identifying orthologs by comparing proteins and species phylogenetic trees and, lastly, iii) InParanoid [O'Brien *et al.*, 2005] that relies on a similar flowchart. All these ortholog

identification methods have been recently tested on a dataset of proteins from different species previously characterized using functional genomics data, such as expression data and protein interaction data [Hulsen *et al.*, 2006]. Results have shown that InParanoid software seems the best ortholog identification method in terms of identifying functionally equivalent proteins in different species [Hulsen *et al.*, 2006].

1.5.3 Multiple Sequence Alignments

In a phylogenetic analysis workflow (but also when interested, for example, in structure modeling, functional site prediction and sequence database searching), a key step (usually following the correct orthologs retrival procedure) consists in comparing those residues with inferred common evolutionary origin or structural/ functional equivalence in the whole sequence dataset. This task is fulfilled through multiple sequence alignment (MSA), that is arranging homolog protein sequences into a rectangular array with the goal that residues in a given column are homologous (derived a single position in an ancestral sequence), superposable (in a rigid local structural alignment) or play a common functional role. Although these three criteria are essentially equivalent for closely related proteins, sequence, structure and function diverge over evolutionary time and different criteria may result in different alignments [Edgar & Batzoglou, 2006]. Many approximate algorithms have been developed for multiple sequence alignments, including the commonly used progressive alignment technique [Pei, 2008]. This greedy heuristic assembly algorithm involves estimating a guide tree (rooted binary tree) from unaligned sequences and then incorporating the sequences into the MSA with a pairwise alignment algorithm while following the tree topology. The scoring schemes used by the pairwise alignment algorithm are arguably the most influential component of the progressive algorithm. They can be divided in two categories, that is matrix- and consistency-based algorithms. Matrix-based algorithms such as ClustalW [Thompson *et al.*, 2002], MUSCLE [Edgar, 2004], and Kalign [Lassmann & Sonnhammer, 2005] use a substitution matrix to assess the cost of matching two symbols or two profiled columns [Notredame, 2007]. Conversely, consistency-based schemes incorporate a larger share of information into the evaluation. This result is achieved by using an approach initially developed for T-Coffee [Notredame *et al.*, 2000] and inspired by Dialign overlapping weights [Morgenstern *et al.*, 1998; Subramanian *et al.*, 2005]. Its principle is to compile a collection of pairwise global and local alignments (primary library) and to use this collection as a position-specific substitution matrix during a regular progressive alignment. The aim is to deliver a final MSA as consistent as possible with the alignments contained in the library. Many extant algorithms are based on this approach such as PCMA [Pei *et al.*, 2003], ProbCons (adopting a Bayesian framework) [Do *et al.*, 2005], MUMMALS [Pei & Grishin, 2006]. Sequence and structural databases are expanding rapidly owing to genome sequencing projects and structural genomics initiatives, offering helpful sources to further improve multiple protein sequence alignments. Structural additional information, for example known 3-dimensional (3D) structures, can be exploited in some multiple alignment methods. In fact, since structures are generally more conserved than sequences, structural information is also valuable for aligning sequences. Several MS algorithm have started implementing

this source of information, and they include 3DCoffee [Poirot *et al.*, 2004] and FUGUE [Shi *et al.*, 2001]. Recently, the Expresso server [Armougom *et al.*, 2006] extends the 3DCoffee method by automatically identifying highly similar 3D structural templates for target sequences and using structural alignments for consistency-based alignments.

1.5.4 Phylogeny

Understanding microbial evolution is essential for gathering information on the most ancient events in the history of Life on our planet [Gribaldo & Brochier, 2009] as well as on the extant relationships between the whole microbial community. This task implies the use of molecular phylogeny techniques, that is the study of phylogenies and processes of evolution by the analysis of DNA or amino acid sequence data [Whelan *et al.*, 2001]. Although parsimony and distance-based methods are widely used, the most statistically robust approach is to consider the problem in a likelihood framework and use accurate models of evolution [Brilli *et al.*, 2008]. It is known [Whelan *et al.*, 2001], in fact, that disadvantages of distance methods include the inevitable loss of evolutionary information when a sequence alignment is converted to pairwise distances, and the inability to deal with models containing parameters for which the values are not known a priori. Concerning maximum parsimony (MP), this approach selects and outputs the tree (or trees) that require the fewest evolutionary changes and is reasonably confident when the number of changes per sequence position is relatively small [Steel & Penny, 2000]. However, as more-divergent sequences are to be analysed, the degree of homoplasy (i.e. parallel, convergent, reversed or superimposed changes) increases and MP tree reconstruction might be misleading since this method has no adequate means to deal with this [Whelan *et al.*, 2001]. Conversely, Maximum likelihood (ML) approaches take the hypothesis (the tree topology) that maximizes the likelihood of the data (the sequence alignment) in the light of an evolutionary model. A great attraction of this approach is the ability to perform robust statistical hypothesis tests and to use modern statistical techniques such as hidden Markov models, Markov chain Monte Carlo and Bayesian inference [Ewens & Grant, 2001; Shoemaker *et al.*, 1999]. The ML framework also allows each site of the alignment to evolve with different replacement patterns, and with different substitution rates in all branches of the tree [Whelan *et al.*, 2001] as in real proteins, where slowly evolving sites are generally functionally or structurally constrained, while variable sites are likely to be less important for protein function. The ML approach (including its variants as the Bayesian framework) has been included in a number of different packages, such as Phylip (http://evolution.gs.washington.edu/phylip.html) PAUP* (http://paup.csit.fsu.edu/)MEGA http://www.megasoftware.net/mega.html, [Tamura *et al.*, 2007]), PAML (http://abacus.gene.ucl.ac.uk/software/paml.html, [Yang, 1997]), mrBayes [Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003] and phyML [Guindon & Gascuel, 2003].

1.5.5 Networks in Biology

A network is a graphical representation of a set of agents, or vertices , linked by edges that represent the connections or interactions between these agents [Dagan *et al.*, 2008]. Given their conceptual plasticity, recent years saw a great increase in the use of networks for representing and analyzing all major biological topics, including protein-protein interaction, gene regulation, metabolism and, recently, sequence similarity. Hence, depending on the subject under study, nodes may represent sequences, products, enzymes, or protein structures whereas links may stand for sequence similarity relationships, metabolic substrates, metabolic reactions, or protein interactions. Biological network analysis has become a central component of computational and systems biology because such analysis provides a unifying language to describe relations within (and between) complex systems also providing useful hints in understanding physiological function(s) of their components. In Figure reffig:system some examples of major biological systems that recently have been analyzed taking advantage of graph theory are proposed . These large-scale analysis are

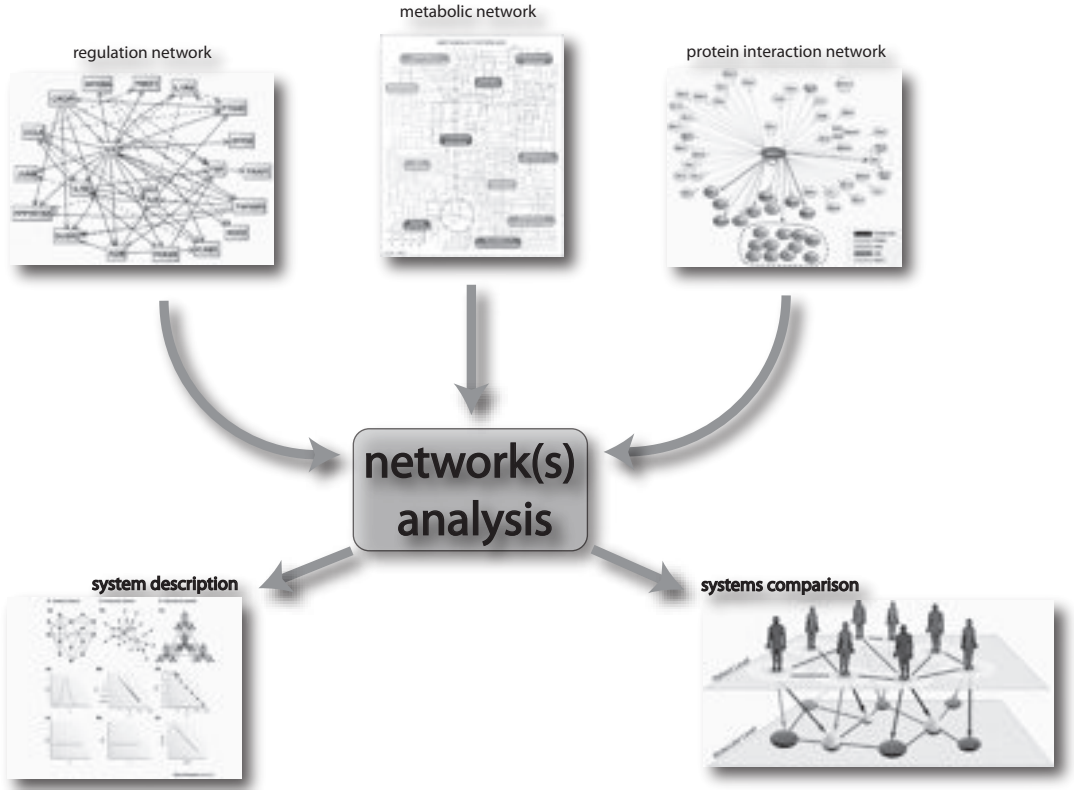


Figure 1.19: The importance of data information in studying complex systems.

beginning to reveal the global organization of the cell. A topological feature often found in large complex networks (both biological or not) is the so-called "scale-free" topology [Barabasi & Albert, 1999]. In networks with such a topology, the vertex connectivity [P(k)] distribution, decays as a power-law [Dwight Kuo *et al.*, 2006], that is $P(k) \approx k^{-\gamma}$, with k representing the number of connections. This indicates a non-random structure of the network and the presence of a few highly connected nodes linking the bulk of poorly

connected ones (Figure 1.20). Recently, scale-free behaviors have been found in many

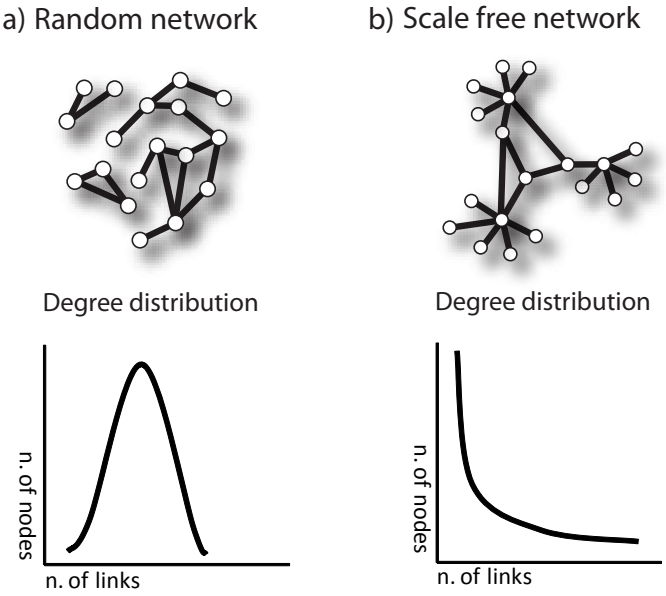


Figure 1.20: Degree distributions of (a) random and (b) scale free networks.

biological networks, including nervous systems [Watts & Strogatz, 1998], metabolic networks [Jeong *et al.*, 2000] protein domains [Wuchty, 2001] and horizontally transferred genes [Dagan *et al.*, 2008]. An important consequence of the power-law connectivity distribution is that a few hubs dominate the overall connectivity of the network (Figure 1.20b), and upon the sequential removal of the most connected nodes the diameter of the network rises sharply, the network eventually disintegrating into isolated clusters that are no longer functional. Scale-free networks also demonstrate unexpected robustness against random errors. In fact, because of the heterogeneity of scale-free networks, random node disruptions do not lead to a major loss of connectivity, but the loss of the hubs causes the breakdown of the network into isolated clusters [Albert *et al.*, 2000]. The validity of these general conclusions for cellular networks can be verified by correlating, for example, the severity of a gene knockout with the number of interactions the gene products participate in. Indeed, as much as 73% of the *S. cerevisiae* genes are non-essential, i.e. their knockout has no phenotypic effects [Giaever *et al.*, 2002]. This might suggest a certain cellular networks robustness in the face of random disruptions. Although the debate is far from being totally resolved (several researchers are questioning this point on the basis of new experimental evidence [Arita, 2004; Przulj *et al.*, 2004]), it is now a commonly accepted fact that biological networks exhibit small-world and scale-free properties and that these collective characteristics are strongly related to the cellular phenotypes observed at the macroscopic level [Grigorov, 2005].

Most of the ideas and the figures presented in this introductory section have appeared on the following peer reviewd published papers and/or book chapters:

Available online at www.sciencedirect.com
ScienceDirect
Physics of Life Reviews 6 (2009) 23–52
PHYSICS of LIFE reviews
www.elsevier.com/locate/plrev

Review

Origin and evolution of metabolic pathways

Renato Fani *, Marco Fondi

Laboratory of Microbial and Molecular Evolution, Department of Evolutionary Biology, Via Romana 17-19, University of Florence, Italy

Received 1 May 2008; received in revised form 27 November 2008; accepted 1 December 2008

Available online 8 January 2009

Communicated by E. Di Mauro

Abstract

The emergence and evolution of metabolic pathways represented a crucial step in molecular and cellular evolution. In fact, the exhaustion of the prebiotic supply of amino acids and other compounds that were likely present in the ancestral environment, imposed an important selective pressure, favoring those primordial heterotrophic cells which became capable of synthesizing those molecules. Thus, the emergence of metabolic pathways allowed primitive organisms to become increasingly less-dependent on exogenous sources of organic compounds.

Comparative analyses of genes and genomes from organisms belonging to Archaea, Bacteria and Eukarya revealed that, during evolution, different forces and molecular mechanisms might have driven the shaping of genomes and the arisal of new metabolic abilities. Among these gene elongations, gene and operon duplications undoubtedly played a major role since they can lead to the (immediate) appearance of new genetic material that, in turn, might undergo evolutionary divergence giving rise to new genes coding for new metabolic abilities. Gene duplication has been invoked in the different schemes proposed to explain why and how the extant metabolic pathways have arisen and shaped. Both the analysis of completely sequenced genomes and directed evolution experiments strongly support one of them, i.e. the patchwork hypothesis, according to which metabolic pathways have been assembled through the recruitment of primitive enzymes that could react with a wide range of chemically related substrates. However, the analysis of the structure and organization of genes belonging to ancient metabolic pathways, such as histidine biosynthesis and nitrogen fixation, suggested that other different hypothesis, i.e. the retrograde hypothesis or the semi-enzymatic theory, may account for the arisal of some metabolic routes.

Keywords: Gene duplication; Patchwork hypothesis; Histidine biosynthesis; Nitrogen fixation

Contents

* Corresponding author. Tel.: +39 055 2288244; fax: +39 055 2288250.
E-mail address: renato.fani@unifi.it (R. Fani).

1571-0645/$ – see front matter
doi:10.1016/j.plrev.2008.12.003

ARTICLE IN PRESS

INSTITUT PASTEUR

ELSEVIER

Research in Microbiology xx (2009) 1–11

www.elsevier.com/locate/resmic

Origin and evolution of operons and metabolic pathways

Marco Fondi [a], Giovanni Emiliani [b], Renato Fani [a,*]

[a] Laboratory of Microbial and Molecular Evolution, Department of Evolutionary Biology, Via Romana 17-19, University of Florence, 50125 Florence, Italy
[b] Department of Environmental and Forestry Sciences, University of Florence, via S. Bonaventura 13, 50145 Florence, Italy

Received 27 March 2009; accepted 8 May 2009

Abstract

The emergence and evolution of metabolic pathways represent a crucial step in molecular and cellular evolution, allowing primitive organisms to become less dependent on exogenous sources of organic compounds. This work will review the main theories accounting for their assembly and for the origin and evolution of prokaryotic operons.

Keywords: Operon origin; Operon evolution; Metabolic pathway origin; Metabolic pathway evolution

1. Origin and evolution of metabolic pathways

1.1. The primordial metabolism

It is widely accepted that ancestral life forms inhabited an environment (the so-called primordial soup) rich in organic compounds spontaneously formed in the prebiotic world. This hypothesis is known as the "Oparin–Haldane theory" [37,49] and predicts that early organisms were heterotrophic and had to perform only a minimum of biosynthesis. If this is so, the increasing number of primordial cells might have led to exhaustion of the prebiotic supply of amino acids and other compounds that were present in the primordial soup. This, in turn, would have imposed a progressively stronger selective pressure, favoring those primordial heterotrophic cells that became capable of synthesizing those molecules whose concentration was decreasing in the primordial soup (Fig. 1). Hence, the emergence of basic metabolic pathways represented a key step in molecular and cellular evolution, since it allowed primitive organisms to become increasingly less dependent on exogenous sources of organic compounds. This also led to the thousands of extant different biochemical reactions and transport processes linked together in pathways (reaction chains) or networks (branched pathways) that can synthesize or catabolize organic compounds and at the same time maximize energy flows through living matter [6]. In a broad sense, they are responsible for the basic metabolic functions that fuel the molecular machinery and inner workings of life [6]. The presence of such highly complex metabolic networks in the extant organisms raises the question of how they appeared starting from ancestral genomes, that were probably composed by only 200–300 genes [45].

In other words, which are the molecular mechanisms that drove the evolution of genes and genomes and the expansion of metabolic abilities of primordial cells?

1.2. Gene duplication and fusion: two main mechanisms for the evolution of metabolic pathways

Different molecular mechanisms may have been responsible for the expansion and shaping of early genomes and metabolic pathways; these include gene elongation, duplication and/or fusion, the modular assembly of new proteins, cell fusion (synology) and horizontal gene transfers (HGTs). The role of horizontal transfer in early cell evolution is discussed by Gribaldo et al. in this issue; thus, the next sections will deal with two other key molecular mechanisms, i.e. gene duplication and fusion.

* Corresponding author.
E-mail addresses: marco.fondi@unifi.it (M. Fondi), giovanni.emiliani@unifi.it (G. Emiliani), renato.fani@unifi.it (R. Fani).

0923-2508/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2009.05.001

Please cite this article in press as: Fondi, M., et al., Origin and evolution of operons and metabolic pathways, Research in Microbiology (2009), doi:10.1016/j.resmic.2009.05.001

S. Casellato, P. Burighel & A. Minelli, eds.
Life and Time: The Evolution of Life and its History. Cleup, Padova 2009.

The primordial metabolism: on the origin and evolution of metabolic pathways and operons

Marco Fondi[1], Giovanni Emiliani[2], Renato Fani[1]*

[1]Laboratorio di Evoluzione Microbica e Molecolare, Dipart. di Biologia Evoluzionistica, Università di Firenze, Via Romana 17-19, I-50125 Firenze, Italy
[2] Consiglio Nazionale delle Ricerche, Istituto per la Valorizzazione del Legno e delle Specie Arboree, via Biasi 75, I 38010 San Michele all'Adige (TN) Italy
*Email: renato.fani@unifi.it

From ancestral to extant genomes

We still do not know when and how life originated (Peretò *et al.* 1997). However, it is commonly assumed that early organisms inhabited an environment rich in organic compounds spontaneously formed in the prebiotic world. This heterotrophic origin of life is frequently referred to as the Oparin-Haldane theory (Oparin 1924, 1936; Lazcano & Miller 1996). If this idea is correct, life evolved from a primordial soup, containing different organic molecules. This soup of nutrient compounds was available for the early heterotrophic organisms, so they had to do only a minimum of biosynthesis. An experimental support to this proposal was obtained in 1953 when Miller (Miller 1953) and Urey showed that amino acids and other organic molecules are formed under atmospheric conditions thought to be representative of those on the early Earth. The first living systems probably did stem directly from the primordial soup and evolved relatively fast up to a common ancestor, usually referred to as LUCA (Last Universal Common Ancestor), an entity representing the divergence starting-point of all the extant life forms on Earth (Fig. 1). According to this view, contemporary genomes are the result of 3.5-4 billions of years of evolution. But how did these ancestral genomes look like? The increasing number of available sequences from organisms belonging to the three domains of life (Bacteria, Archaea and Eukarya) has allowed inferring both the size and the gene content of the genomes of the first living cells that appeared on the Earth. A recent estimate of the minimal gene content of LUCA based on whole-genome phylogenies indicated that ancestral genomes were probably composed by about 1000-1500 genes (Ouzounis *et al.* 2006). However, despite this small gene content, ancestral genomes were probably fairly complex, similar to those of the extant free-living prokaryotes and included

References

Abascal, F. & Valencia, A. (2002). Clustering of proximal sequence space for the identification of protein families. *Bioinformatics*, **18**, 908–21.

Albert, R., Jeong, H. & Barabasi, A.L. (2000). Error and attack tolerance of complex networks. *Nature*, **406**, 378–82.

Alifano, P., Fani, R., Lio, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M.S. & Bruni, C.B. (1996). Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol Rev*, **60**, 44–69.

Alm, E.J., Huang, K.H., Price, M.N., Koche, R.P., Keller, K., Dubchak, I.L. & Arkin, A.P. (2005). The microbesonline web site for comparative genomics. *Genome Res*, **15**, 1015–22.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.

Arita, M. (2004). The metabolic world of escherichia coli is not small. *Proc Natl Acad Sci U S A*, **101**, 1543–7.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V. & Notredame, C. (2006). Expresso: automatic incorporation of structural information in multiple sequence alignments using 3d-coffee. *Nucleic Acids Res*, **34**, W604–8.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Camara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepere, G., Malinsky, S., Nowacki, M., Nowak, J.K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Betermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J. & Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. *Nature*, **444**, 171–8.

Babbitt, P.C. & Gerlt, J.A. (1997). Understanding enzyme superfamilies. chemistry as the fundamental determinant in the evolution of new catalytic activities. *J Biol Chem*, **272**, 30591–4.

Barabasi, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–12.

Beadle, G.W. & Tatum, E.L. (1941). Genetic control of biochemical reactions in neurospora. *Proc Natl Acad Sci U S A*, **27**, 499–506.

Belfaiza, J., Parsot, C., Martel, A., de la Tour, C.B., Margarita, D., Cohen, G.N. & Saint-Girons, I. (1986). Evolution in biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region. *Proc Natl Acad Sci U S A*, **83**, 867–71.

Bork, P. & Rohde, K. (1990). Sequence similarities between tryptophan synthase beta subunit and other pyridoxal-phosphate-dependent enzymes. *Biochem Biophys Res Commun*, **171**, 1319–25.

Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O. & Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol*, **5**, R35.

Brilli, M. & Fani, R. (2004). The origin and evolution of eucaryal his7 genes: from metabolon to bifunctional proteins? *Gene*, **339**, 149–60.

Brilli, M., Fani, R. & Lio, P. (2008). Current trends in the bioinformatic sequence analysis of metabolic pathways in prokaryotes. *Brief Bioinform*, **9**, 34–45.

Brown, J.R. (2003). Ancient horizontal gene transfer. *Nat. Rev. Genet.*, **4**, 121–32, brown, James R Research Support, Non-U.S. Gov't Review England Nature reviews. Genetics Nat Rev Genet. 2003 Feb;4(2):121-32.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. & Emili, A. (2005). Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, **433**, 531–7.

Cannon, S.B. & Young, N.D. (2003). Orthoparamap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics*, **4**, 35.

Cassan, M., Parsot, C., Cohen, G.N. & Patte, J.C. (1986). Nucleotide sequence of lysc gene encoding the lysine-sensitive aspartokinase iii of escherichia coli k12. evolutionary pathway leading to three isofunctional enzymes. *J Biol Chem*, **261**, 1052–7.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. & Cesareni, G. (2009). Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res*.

Chain, P.S., Grafham, D.V., Fulton, R.S., Fitzgerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., Cole, J.R., Ding, Y., Dugan, S., Field, D., Garrity, G.M., Gibbs, R., Graves, T., Han, C.S., Harrison, S.H., Highlander, S., Hugenholtz, P., Khouri, H.M., Kodira, C.D., Kolker, E., Kyrpides, N.C., Lang, D., Lapidus, A., Malfatti, S.A., Markowitz, V., Metha, T., Nelson, K.E., Parkhill, J., Pitluck, S., Qin, X., Read, T.D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R.L., Sutton, G., Thomson, N.R., Tiedje, J.M., Weinstock, G., Wollam, A. & Detter, J.C. (2009). Genomics. genome project standards in a new era of sequencing. *Science*, **326**, 236–7.

Clarke, P. (1974). *The evolution of enzymes for the utilization of novel substrates. Evolution in the microbial world.*. Cambridge University Press, Cambridge.

Conant, G. & Wolfe, K. (2007). Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology*, **3**, 129.

Copley, R.R. & Bork, P. (2000). Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J Mol Biol*, **303**, 627–41.

Copley, S.D. (2000). Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends Biochem Sci*, **25**, 261–5.

Cotter, P.J., Caffrey, D.R. & Shields, D.C. (2002). Improved database searches for orthologous sequences by conditioning on outgroup sequences. *Bioinformatics*, **18**, 83–91.

Dagan, T. & Martin, W. (2006). The tree of one percent. *Genome Biol.*, **7**, 118.

Dagan, T. & Martin, W. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. USA*, **104**, 870–5.

Dagan, T., Artzy-Randrup, Y. & Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A*, **105**, 10039–44.

Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, **23**, 324–8.

Daubin, V. & Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of orfans in e. coli. *Genome Res*, **14**, 1036–42.

de Daruvar, A., Collado-Vides, J. & Valencia, A. (2002). Analysis of the cellular functions of escherichia coli operons and their conservation in bacillus subtilis. *J Mol Evol*, **55**, 211–21.

de Rosa, R. & Labedan, B. (1998). The evolutionary relationships between the two bacteria *Escherichia coli* and *Haemophilus influenzae* and their putative last common ancestor. *Molecular Biology and Evolution*, **15**, 17–27.

Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D., Huang, K.H., Keller, K., Novichkov, P.S., Dubchak, I.L., Alm, E.J. & Arkin, A.P. (2009). Microbesonline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res*.

Delaye, L., Becerra, A. & A, L. (2005). The last common ancestor: Whats in a name? *Origin of Life and Evolution of Biosphere*, **35**, 537–54.

Demerec, M. & Demerec, Z. (1956). Analysis of linkage relationships in salmonella by transduction techniques. *Brookhaven Symp. Biol*, **8**, 7584.

Do, C.B., Mahabhashyam, M.S., Brudno, M. & Batzoglou, S. (2005). Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, **15**, 330–40.

Dwight Kuo, P., Banzhaf, W. & Leier, A. (2006). Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems*, **85**, 177–200.

Edgar, R.C. (2004). Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar, R.C. & Batzoglou, S. (2006). Multiple sequence alignment. *Curr Opin Struct Biol*, **16**, 368–73.

Eklund, H. & Fontecave, M. (1999). Glycyl radical enzymes: a conservative structural basis for radicals. *Structure*, **7**, R257–62.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, **30**, 1575–84.

Ewens, W. & Grant, G. (2001). *Statistical Methods in Bioinformatics: AnIntroduction*. Springer, New York.

Eyre-Walker, A. (1995). The distance between escherichia coli genes is related to gene expression levels. *J Bacteriol*, **177**, 5368–9.

Fani, R. (2004). Gene duplication and gene loading. In *Microbial evolution: gene establishment, survival, and exchange.*, ASM Press, Washington DC.

Fani, R., Lio, P., Chiarelli, I. & Bazzicalupo, M. (1994). The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisa and hisf genes. *J Mol Evol*, **38**, 489–95.

Fani, R., Lio, P. & Lazcano, A. (1995). Molecular evolution of the histidine biosynthetic pathway. *J Mol Evol*, **41**, 760–74.

Fani, R., Mori, E., Tamburini, E. & Lazcano, A. (1998). Evolution of the structure and chromosomal distribution of histidine biosynthetic genes. *Orig Life Evol Biosph*, **28**, 555–70.

Fani, R., Gallo, R. & Lio, P. (2000). Molecular evolution of nitrogen fixation: the evolutionary history of the nifd, nifk, nife, and nifn genes. *J Mol Evol*, **51**, 1–11.

Fani, R., Brilli, M. & Lio, P. (2005). The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. *J Mol Evol*, **60**, 378–90.

Fischer, D. & Eisenberg, D. (1999). Finding families for genomic orfans. *Bioinformatics*, **15**, 759–62.

Fondi, M., Brilli, M., Emiliani, G., Paffetti, D. & Fani, R. (2007). The primordial metabolism: an ancestral interconnection between leucine, arginine, and lysine biosynthesis. *BMC Evol Biol*, **7 Suppl 2**, S3.

Fondi, M., Emiliani, G. & Fani, R. (2009). Origin and evolution of operons and metabolic pathways. *Res Microbiol*, **160**, 502–12.

Forterre, P. & Gribaldo, S. (2007). The origin of modern terrestrial life. *HFSP J*, **1**, 156–68.

Frost, L.S., Leplae, R., Summers, A.O. & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–32.

Fulton, D.L., Li, Y.Y., Laird, M.R., Horsman, B.G., Roche, F.M. & Brinkman, F.S. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, **7**, 270.

Gabaldon, T., Dessimoz, C., Huxley-Jones, J., Vilella, A.J., Sonnhammer, E.L. & Lewis, S. (2009). Joining forces in the quest for orthologs. *Genome Biol*, **10**, 403.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balazsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M.V., Grechkin, Y., Mseeh, F., Fonstein, M.Y., Overbeek, R., Barabasi, A.L., Oltvai, Z.N. & Osterman, A.L. (2003). Experimental determination and system level analysis of essential genes in escherichia coli mg1655. *J Bacteriol*, **185**, 5673–84.

Gerlt, J.A. & Babbitt, P.C. (1998). Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Curr Opin Chem Biol*, **2**, 607–12.

Gevers, D., Vandepoele, K., Simillon, C. & Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol*, **12**, 148–54.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A.P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.D., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kotter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D., Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.Y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W. & Johnston, M. (2002). Functional profiling of the saccharomyces cerevisiae genome. *Nature*, **418**, 387–91.

Glansdorff, N. (1999). On the origin of operons and their possible role in evolution toward thermophily. *J Mol Evol*, **49**, 432–8.

Gogarten, J.P. & Townsend, J.P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.*, **3**, 679–87.

Gogarten, J.P., Doolittle, W.F. & Lawrence, J.G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, **19**, 2226–38.

Granick, S. (1957). Speculations on the origins and evolution of photosynthesis. *Ann N Y Acad Sci*, **69**, 292–308.

Granick, S. (1965). Evolution of heme and chlorophyll. In F. Neidhardt, R. Curtiss III, J. Ingraham, E. Lin, K. Low, B. Magasanik, W. Reznikoff, M. Schaechter, H. Umbarger & M. Riley, eds., *Evolving genes and proteins*, 67–88, Academic Press, New York.

Gribaldo, S. & Brochier, C. (2009). Phylogeny of prokaryotes: does it exist and why should we care? *Res Microbiol*, **160**, 513–21.

Gribaldo, S. & Brochier-Armanet, C. (2006). The origin and evolution of archaea: a state of the art. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **361**, 1007–22.

Grigorov, M.G. (2005). Global properties of biological networks. *Drug Discov Today*, **10**, 365–72.

Guglierame, P., Pasca, M.R., De Rossi, E., Buroni, S., Arrigo, P., Manina, G. & Riccardi, G. (2006). Efflux pump genes of the resistance-nodulation-division family in burkholderia cenocepacia genome. *BMC Microbiol*, **6**, 66.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**, 696–704.

Gupta, R.S. & Singh, B. (1992). Cloning of the hsp70 gene from halobacterium marismortui: relatedness of archaebacterial hsp70 to its eubacterial homologs and a model for the evolution of the hsp70 gene. *J Bacteriol*, **174**, 4594–605.

Hall, B. & Zuzel, T. (1980). Evolution of a new enzymatic function by recombination within a gene. *Proc Natl Acad Sci USA*, **77**, 352933.

Hazkani-Covo, E. & Graur, D. (2005). Evolutionary conservation of bacterial operons: does transcriptional connectivity matter? *Genetica*, **124**, 145–66.

He, X. & Zhang, J. (2006). Transcriptional reprogramming and backup between duplicate genes: Is it a genomewide phenomenon? *Genetics*, **172(2)**, 13631367.

Hegeman, G.D. & Rosenberg, S.L. (1970). The evolution of bacterial enzyme systems. *Annu Rev Microbiol*, **24**, 429–62.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. & Apweiler, R. (2004). Intact: an open source molecular interaction database. *Nucleic Acids Res*, **32**, D452–5.

Horowitz, N. (1965). The evolution of biochemical syntheses retrospect and prospect. In F. Neidhardt, R. Curtiss III, J. Ingraham, E. Lin, K. Low, B. Magasanik, W. Reznikoff, M. Schaechter, H. Umbarger & M. Riley, eds., *Evolving genes and proteins*, 1523, Academic Press, New York.

Horowitz, N.H. (1945). On the evolution of biochemical syntheses. *Proc Natl Acad Sci U S A*, **31**, 153–7.

Huang, J. & Gogarten, J.P. (2006). Ancient horizontal gene transfer can benefit phylogenetic reconstruction. *Trends Genet.*, **22**, 361–6.

Huelsenbeck, J.P. & Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–5.

Hulsen, T., Huynen, M.A., de Vlieg, J. & Groenen, P.M. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*, **7**, R31.

Huynen, M., Snel, B., Lathe, r., W. & Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, **10**, 1204–10.

Itoh, T., Takemoto, K., Mori, H. & Gojobori, T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol*, **16**, 332–46.

Jacob, F. & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, **3**, 318–56.

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. & von Mering, C. (2009). String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, **37**, D412–6.

Jensen, R. (1996). Evolution of metabolic pathways in enteric bacteria. In *Escherichia coli and Salmonella typhimurium, Cellular and Molecular Biology*, 26492662, ASM Press, Washington DC.

Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, **30**, 409–25.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–4.

Kanehisa, M. & Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27–30.

Keseler, I.M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R.P., Johnson, D.A., Krummenacker, M., Nolan, L.M., Paley, S., Paulsen, I.T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A.G. & Karp, P.D. (2009). Ecocyc: a comprehensive view of escherichia coli biology. *Nucleic Acids Res*, **37**, D464–70.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. & Pandey, A. (2009). Human protein reference database–2009 update. *Nucleic Acids Res*, **37**, D767–72.

Klotz, M.G. & Norton, J.M. (1998). Multiple copies of ammonia monooxygenase (amo) operons have evolved under biased at/gc mutational pressure in ammonia-oxidizing autotrophic bacteria. *FEMS Microbiol Lett*, **168**, 303–11.

Koonin, E. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Review in Microbiology*, **1**, 127–36.

Koonin, E. & Martin, W. (2002). On the evolution of cells. *Proc Natl Acad Sci U S A*, **99**, 8742–7.

Koonin, E. & Martin, W. (2005). On the origin of genomes and cells within inorganic compartments. *Trends in Genetcis*, **12**, 647–54.

Labedan, B. & Riley, M. (1995). Widespread protein sequence similarities: Origin of *Escherichia coli* genes. *Journal of Bacteriology*, **16**, 15.

Langer, D., Hain, J., Thuriaux, P. & Zillig, W. (1995). Transcription in archaea: similarity to that in eucarya. *Proc Natl Acad Sci U S A*, **92**, 5768–72.

Lassmann, T. & Sonnhammer, E.L. (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.

Lawrence, J. (1999). Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev*, **9**, 642–8.

Lawrence, J.G. & Roth, J.R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–60.

Lawrence, M.C., Barbosa, J.A., Smith, B.J., Hall, N.E., Pilling, P.A., Ooi, H.C. & Marcuccio, S.M. (1997). Structure and mechanism of a sub-family of enzymes related to n-acetylneuraminate lyase. *J Mol Biol*, **266**, 381–99.

Lazcano, A. & Miller, S.L. (1994). How long did it take for life to begin and evolve to cyanobacteria? *Journal of Molecular Evolution*, **39**, 546–54.

Lazcano, A. & Miller, S.L. (1996). The origin and early evolution of life: prebiotic chemistry, the pre-rna world, and time. *Cell*, **85**, 793–8.

Lazcano, A., Fox, G. & Or, J. (1992). Life before dna: the origin and evolution of early archean cells. In R. Mortlock & M. Gallo, eds., *Experiments in the evolution of catabolic pathways using modern bacteria, the evolution of metabolic functions*, 1–13, CRC Press, Boca Raton, FL.

Lazcano, A., Diaz-Villagomez, E., Mills, T. & Oro, J. (1995). On the levels of enzymatic substrate specificity: implications for the early evolution of metabolic pathways. *Adv Space Res*, **15**, 345–56.

Lewis (1951). Pseudoallelism and gene evolution. *Spring Harb Symp Quant Biol*, **16**, 15.

Li, L., Stoeckert, J., C. J. & Roos, D.S. (2003). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, **13**, 2178–89.

Li, W. & Graur, D. (1991). *Fundamentals of molecular evolution.*. Sinauer Associates, Inc, Sunderland, MA, USA.

Lio', P., Brilli, M. & Fani, R. (2007). *Phylogenetics and computational biology of multigene families.*. Springer, Berlin.

Lynch, M. & Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 11515.

Lynch, M. & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–73.

Ma, J., Campbell, A. & Karlin, S. (2002). Correlations between shine-dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol*, **184**, 5733–45.

Maas, W.K. (1964). Studies on the mechanism of repression of arginine biosynthesis in escherichia coli. ii. dominance of repressibility in diploids. *J Mol Biol*, **8**, 365–70.

Maeder, D.L., Weiss, R.B., Dunn, D.M., Cherry, J.L., Gonzalez, J.M., DiRuggiero, J. & Robb, F.T. (1999). Divergence of the hyperthermophilic archaea pyrococcus furiosus and p. horikoshii inferred from complete genomic sequences. *Genetics*, **152**, 1299–305.

Makarova, K.S., Ponomarev, V.A. & Koonin, E.V. (2001). Two c or not two c: recurrent disruption of zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol*, **2**, RESEARCH 0033.

Martin, R. (1971). Enzymes and intermediates of histidine biosynthesis in *Salmonella typhimurium. Methods Enzymol B*, **17**, 3–44.

Mathews, C.K. (1993). The cell-bag of enzymes or network of channels? *J Bacteriol*, **175**, 6377–81.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L. & D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, **37**, D619–22.

McLachlan, A. (1991). Gene duplication and the origin of repetitive protein structures. *Cold Spring Harb Symp Quant Biol*, **52**, 411–20.

Melendez-Hevia, E., Waddell, T.G. & Cascante, M. (1996). The puzzle of the krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *J Mol Evol*, **43**, 293–303.

Miller, S.L. (1953). Production of amino acids under possible primitive earth conditions. *Science*, **117**, 528–9.

Mira, A., Ochman, H. & Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet*, **17**, 589–96.

Moreno-Hagelsieb, G. & Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18 Suppl 1**, S329–36.

Morgenstern, B., Frech, K., Dress, A. & Werner, T. (1998). Dialign: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–4.

Mortlock, R. & Gallo, M. (1992). Experiments in the evolution of catabolic pathways using modern bacteria. In R. Mortlock & M. Gallo, eds., *The evolution of metabolic functions*, 1–13, CRC Press, Boca Raton, FL.

Mushegian, A.R. & Koonin, E.V. (1996). Gene order is not conserved in bacterial evolution. *Trends Genet*, **12**, 289–90.

Nadeau, J. & Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution paramecium. *Genetics*, **147**, 1259–66.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, **3**, e123.

Notredame, C., Higgins, D.G. & Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205–17.

Nyunoya, H. & Lusty, C.J. (1983). The carb gene of escherichia coli: a duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase. *Proc Natl Acad Sci U S A*, **80**, 4629–33.

O'Brien, K.P., Remm, M. & Sonnhammer, E.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, **33**, D476–80.

Ochman, H., Lerat, E. & Daubin, V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proc. Natl. Acad. Sci. USA*, **102**, 65956599.

Ohno, S. (1972a). Simplicity of mammalian regulatory systems. *Dev Biol*, **27**, 131–6.

Ohno, S. (1972b). Simplicity of mammalian regulatory systems. *Developmental Biology*, **27**, 131–6.

Ohno, S. (1980). Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics*, **95**, 237–258.

Ohte, T. (2000). Evolution of gene families. *Gene*, **259**, 45–52.

Omelchenko, M.V., Makarova, K.S., Wolf, Y.I., Rogozin, I.B. & Koonin, E.V. (2003). Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol*, **4**, R55.

Oparin (1936). *The origin of life*. Dover, New York.

Oparin (1967). The origin of life. In *Translation: Appendix in Bernal JD.*, World Publishers, Cleveland, Ohio.

Ourisson, G. & Nakatani, Y. (1994). The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. *Chem Biol*, **1**, 11–23.

Ouzounis, C., Kunin, V., Darzentas, N. & L, G. (2006). A minimal estimate for the gene content of the last universal common ancestor exobiology from a terrestrial perspective. *Research in Microbiology*, **157**, 57–68.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, **96**, 2896–901.

Pal, C. & Hurst, L.D. (2004). Evidence against the selfish operon theory. *Trends Genet*, **20**, 232–4.

Papaleo, M.C., Russo, E., Fondi, M., Emiliani, G., Frandi, A., Brilli, M., Pastorelli, R. & Fani, R. (2009). Structural, evolutionary and genetic analysis of the histidine biosynthetic "core" in the genus burkholderia. *Gene*, **448**, 16–28.

Parsot, C. (1986). Evolution of biosynthetic pathways: a common ancestor for threonine synthase, threonine dehydratase and d-serine dehydratase. *EMBO J*, **5**, 3013–9.

Parsot, C., Cossart, P., Saint-Girons, I. & Cohen, G.N. (1983). Nucleotide sequence of thrc and of the transcription termination region of the threonine operon in escherichia coli k12. *Nucleic Acids Res*, **11**, 7331–45.

Pei, J. (2008). Multiple protein sequence alignment. *Curr Opin Struct Biol*, **18**, 382–6.

Pei, J. & Grishin, N.V. (2006). Mummals: multiple sequence alignment improved by using hidden markov models with local structural information. *Nucleic Acids Res*, **34**, 4364–74.

Pei, J., Sadreyev, R. & Grishin, N.V. (2003). Pcma: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–8.

Pereto, J., Fani, R., Leguina, J. & Lazcano, A. (2000). Enzyme evolution and the development of metabolic pathways. In *New beer in an old bottle: Eduard Buchner and the growth of biochemical knowledge*, Cornish-Bowden, A, editor, Valencia: Universitat de Valencia.

Pertea, M., Ayanbule, K., Smedinghoff, M. & Salzberg, S.L. (2009). Operondb: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res*, **37**, D479–82.

Poirot, O., Suhre, K., Abergel, C., O'Toole, E. & Notredame, C. (2004). 3dcoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res*, **32**, W37–40.

Price, M.N., Huang, K.H., Alm, E.J. & Arkin, A.P. (2005a). A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, **33**, 880–92.

Price, M.N., Huang, K.H., Arkin, A.P. & Alm, E.J. (2005b). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res*, **15**, 809–19.

Price, M.N., Arkin, A.P. & Alm, E.J. (2006). The life-cycle of operons. *PLoS Genet*, **2**, e96.

Price, M.N., Dehal, P. & Arkin, A.P. (2007). Orthologous transcription factors in bacteria have different functions and regulate different genes. *Plos Computational Biology*, **3**, 1739–50.

Przulj, N., Corneil, D.G. & Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–15.

Reizer, J. & Saier, J., M. H. (1997). Modular multidomain phosphoryl transfer proteins of bacteria. *Curr Opin Struct Biol*, **7**, 407–15.

Rieder, G., Merrick, M.J., Castorph, H. & Kleiner, D. (1994). Function of hisf and hish gene products in histidine biosynthesis. *J Biol Chem*, **269**, 14386–90.

Rocha, E.P. (2006). Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol*, **23**, 513–22.

Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A. & Koonin, E.V. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*, **30**, 2212–23.

Ronquist, F. & Huelsenbeck, J.P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–4.

Rubin, R.A., Levy, S.B., Heinrikson, R.L. & Kezdy, F.J. (1990). Gene duplication in the evolution of the two complementing domains of gram-negative bacterial tetracycline efflux proteins. *Gene*, **87**, 7–13.

Sabatti, C., Rohlin, L., Oh, M.K. & Liao, J.C. (2002). Co-expression pattern from dna microarray experiments as a tool for operon prediction. *Nucleic Acids Res*, **30**, 2886–93.

Sharov, A. (2006). Genome increase as a clock for the origin and evolution of life. *Biology Direct*, **1**, 17.

Shi, J., Blundell, T.L. & Mizuguchi, K. (2001). Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**, 243–57, shi, J Blundell, T L Mizuguchi, K Research Support, Non-U.S. Gov't England Journal of molecular biology J Mol Biol. 2001 Jun 29;310(1):243-57.

Shi, T. & Falkowski, P. (2008). Genome evolutuon in cyanobacteria: The stable core and the variable shell. *Proc. Natl. Acad. Sci. USA*, **107**, 2510–2515.

Shoemaker, J.S., Painter, I.S. & Weir, B.S. (1999). Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet*, **15**, 354–8.

Srere, P.A. (1987). Complexes of sequential metabolic enzymes. *Annu Rev Biochem*, **56**, 89–124.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, **34**, D535–9.

Steel, M. & Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol*, **17**, 839–50.

Subramanian, A.R., Weyer-Menkhoff, J., Kaufmann, M. & Morgenstern, B. (2005). Dialign-t: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.

Swain, P.S. (2004). Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J Mol Biol*, **344**, 965–76.

Takiguchi, M., Matsubasa, T., Amaya, Y. & Mori, M. (1989). Evolutionary aspects of urea cycle enzyme genes. *Bioessays*, **10**, 163–6.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). Mega4: Molecular evolutionary genetics analysis (mega) software version 4.0. *Mol Biol Evol*, **24**, 1596–9.

Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2002). Multiple sequence alignment using clustalw and clustalx. *Curr Protoc Bioinformatics*, **Chapter 2**, Unit 2 3.

Vicente, M., Gomez, M.J. & Ayala, J.A. (1998). Regulation of transcription of cell division genes in the escherichia coli dcw cluster. *Cell Mol Life Sci*, **54**, 317–24.

Walsh, J. (1995). How often do duplicated genes evolve new functions? *Genetics*, **139**, 421–8.

Watanabe, H., Mori, H., Itoh, T. & Gojobori, T. (1997). Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol*, **44 Suppl 1**, S57–64.

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–2.

Whelan, S., Lio, P. & Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet*, **17**, 262–72.

Wilmanns, M., Hyde, C.C., Davies, D.R., Kirschner, K. & Jansonius, J.N. (1991). Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry*, **30**, 9161–9.

Woese, C. (1998). The universal ancestor. *Proc Natl Acad Sci U S A*, **95**, 6854–9.

Woese, C. (2000). Interpreting the universal phylogenetic tree. *Proc. Natl. Acad. Sci. USA*, **97**, 8392–6.

Woese, C. (2002). On the evolution of cells. *Proc. Natl. Acad. Sci. USA*, **99**, 8742–8747.

Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S. & Koonin, E.V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res*, **11**, 356–72.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Mol Biol Evol*, **18**, 1694–702.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. & Eisenberg, D. (2002). Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, **30**, 303–5.

Xie, G., Keyhani, N.O., Bonner, C.A. & Jensen, R.A. (2003). Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol Mol Biol Rev*, **67**, 303–42, table of contents.

Yanai, I., Wolf, Y.I. & Koonin, E.V. (2002). Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*, **3**, research0024.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, **13**, 555–6.

Ycas, M. (1974). On earlier states of the biochemical system. *J Theor Biol*, **44**, 145–60.

Yu, J.S., Madison-Antenucci, S. & Steege, D.A. (2001). Translation at higher than an optimal level interferes with coupling at an intercistronic junction. *Mol Microbiol*, **42**, 821–34.

Zmasek, C.M. & Eddy, S.R. (2002). Rio: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.

Chapter 2

Aims and presentation of the work

This section is an overview of all the results presented in this dissertation and that will be discussed separately in each of the following chapters. The whole body of data embedded in this thesis can be subdivided into two different major areas (Figure 2.1): the first (namely "Origin and Evolution of Metabolic Pathways", Part I) deals with evolutionary events that likely played a key role in the assembly and in the shaping of modern biosynthetic routes. Events presented in these chapters span through several evolutionary phases, ranging from early events (likely soon after the emergence of LUCA) up to more recent ones. The second part of the work (Part II) deals with comparative evolutionary genomics (Figure 2.1), and data presented in the corresponding chapters generally refer to more recent evolutionary events.

2.1 Origin and Evolution of Metabolic Pathways: a summary

The analysis of histidine biosynthetic route, one of the best characterized anabolic pathways, is reported in *Chapter 3*. In order to depict a comprehensive scenario of its evolution, three different aspects of this route were taken into account, that is i) the role of gene fusion in the assembly and shaping of this pathway, ii) the evolution of histidine biosynthesis in Archaea and, finally, iii) the structure, the organization and the regulation of the histidine biosynthetic *core* in the genus *Burkholderia*. After histidine, we analyzed the lysine biosynthetic route, another interesting case study in the context of metabolism origin and evolution (*Chapter 4*). In particular, we analyzed two important evolutionary features of this pathway: i) the presence of two (apparently) unrelated biosynthetic routes for the biosynthesis of the aminoacid lysine and ii) its evolutionary interconnnections with two other metabolic pathways, namely methionine and threonine. Another key point of bacterial metabolism evolution is likely represented by the building up of nitrogen fixation. We analyzed the molecular mechanisms associated with the appearance of this important metabolic innovation in *Chapter 5*. In the last chapter of the "Origin and Evolution of Metabolic Pathways" section (*Chapter 6*), we faced another key step towards the development of modern terrestrial ecosystems, that is appearance of land plants. In particular we

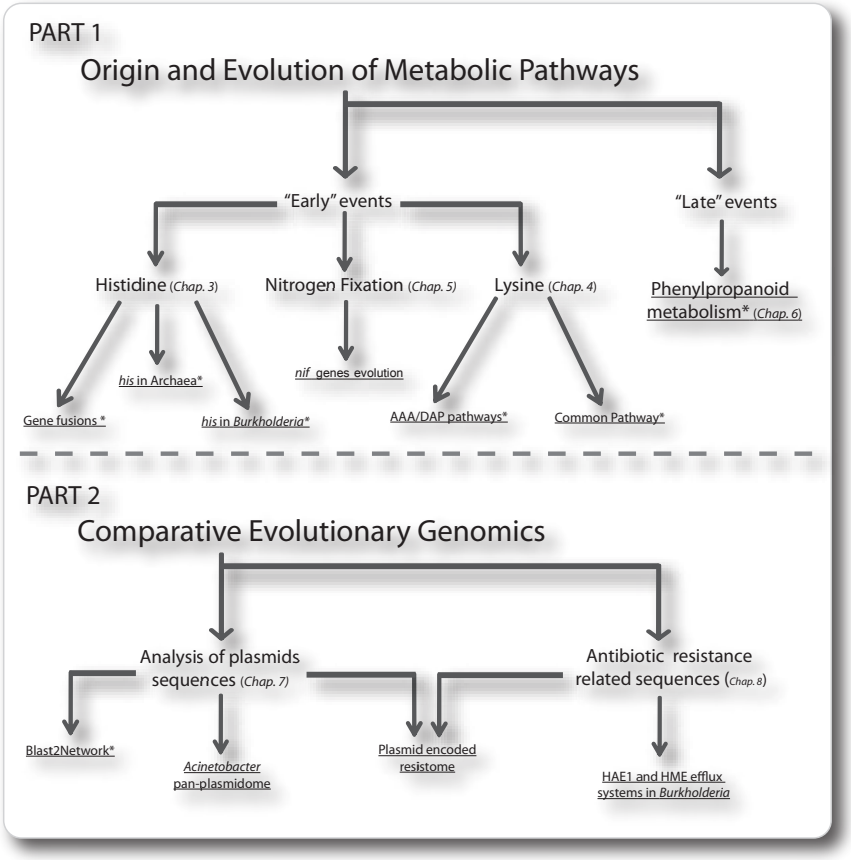


Figure 2.1: Schematic representation of the overall organization of the work. Asterisks indicate works published on *peer reviewed* journals.

focused on the appearance of the phenylpropanoid metabolism, a ubiquitous and specific trait of land plants that, nowadays, provides vital compounds such as lignin [essential for vascularization (xylem) and stem rigidity out of water], flavonoids [essential for reproductive biology (flower and fruit colors)], protection against UV (pigments), microbial attack (phytoalexins), and plant-microbe interaction (flavonoids). Our results highlight a possible crucial role of HGT from soil bacteria in the assembly of phenylpropanoid metabolism and, in turnm, in the path leading to land colonization by plants and their subsequent evolution.

2.2 Comparative Evolutionary Genomics: a summary

In this section different bioinformatic tools are used to compare genes and genomes from different microorganisms in order to gain insights into the mechanisms of evolution. In

this part of the thesis, a particular attention is reserved to plasmid molecules (a class of Mobile Genetic Elements, MGE) and particularly to their role in prokaryotic evolution, such as their evolutionary cross-talking with chromosomes and the spreading of antibiotic resistance. In particular (*Chapter 7*, Blast2Network (B2N), a newly developed bioinformatic package allowing the automatic phylogenetic profiling and the visualization of homology relationships in a large number of plasmid sequences is presented (together with its first application to decipher the evolutionary steps of the whole set of plasmids belonging to *Enterobacteriaceae* subdivision). Furthermore, in *Chapter 8*, computational tools were used for reconstructing the reticulate evolution (mainly guided by HGT and recombination events) of a larger set of sequences, that is all the plasmids and the chromosomes of microorganisms belonging to the γ-proteobacterial genus *Acinetobacter*. In *Chapter 9*, the B2N package was implemented with other *ad hoc* developed Perl modules in order to perform a comprehensive analysis aiming at describing i) the horizontal flow of antibiotic resistance coding genes (the resistome) across the microbial community and ii) to identify those ecological niches (if any) whose inhabitants mostly contribute to their mobilization. Still in the context of bacterial antibiotic resistance issue, *Chapter 10* reports a comprehensive computational analysis concerning both the distribution and the phylogeneny of the HAE1 and HME efflux systems in the genus *Burkholderia*, providing a i) deeper knowledge of the presence, the structure and the distribution of RND proteins in these species and ii) an evolutionary model accounting for their appearance and maintenance in this genus. Interestingly, data presented in this work may serve as a basis for future experimental tests, focused especially on HAE1 proteins, aimed at the identification of novel targets in antimicrobial therapy against *Burkholderia* species.

Part of the data presented in this dissertation have been published on *peer-reviewed* journals. In these cases results will be presented with the journal paper format and inserted as a whole in the corresponding chapter.

Part I

Origin and Evolution of Metabolic Pathways

Chapter 3

Histidine biosynthesis evolution

Histidine biosynthesis represents an excellent model for the analysis of the molecular mechanisms and the forces that have driven the origin and evolution of metabolic pathways. Indeed, it is one of the best characterized anabolic pathways and a large body of genetic and biochemical information is available, including gene structure, organization and expression. For over 40 years this pathway has been the subject of extensive studies, mainly in the enterobacterium *Escherichia coli* and its close relative *Salmonella typhimurium*, for both of which details of histidine biosynthesis appear to be identical. As shown in Figure 3.1, in these two enterobacteria the pathway is unbranched, and includes a number of complex and unusual biochemical reactions. It consists of twelve intermediates, all of which have been described, produced by ten enzymes. There are several independent evidences for the antiquity of the histidine biosynthetic pathway. It is generally accepted that histidine is present in the active sites of enzymes because of the special properties of the imidazole group. The apparently universal phylogenetic distribution of the *his* genes suggests that histidine synthesis was already part of the metabolic abilities of the last common ancestor of the three extant cell domains. The chemical synthesis of histidine, of prebiotic analogues of histidine, and of histidyl-histidine under primitive conditions has been reported, as well as the role of the latter in the enhancement of some possible prebiotic oligomerization reactions involving amino acids and nucleotides. Since its biosynthesis requires a carbon and a nitrogen equivalent from the purine ring of ATP, it has also been suggested that histidine may be the molecular vestige of a catalytic ribonucleotide from an earlier biochemical stage in which RNA played a major role in catalysis. If primitive catalysts required histidine, then the eventual exhaustion of the prebiotic supply of histidine and histidine-containing peptides must have imposed an important pressure favoring those organisms capable of synthesizing histidine. Histidine biosynthesis plays also an important role in cellular metabolism, since four of the *his* genes (*hisBHAF*), forming the so-called *core* of the pathway (Figure 3.1), represent a metabolic cross-point interconnecting histidine biosynthesis to both nitrogen metabolism and de novo synthesis of purines. The connection with purine biosynthesis results from an enzymatic step catalyzed by imidazole glycerol phosphate synthase, an enzyme which has been shown to be a dimeric protein composed of one subunit each of the *hisH* and *hisF* genes product. This heterodimeric enzyme catalyzes the transformation of PRFAR into

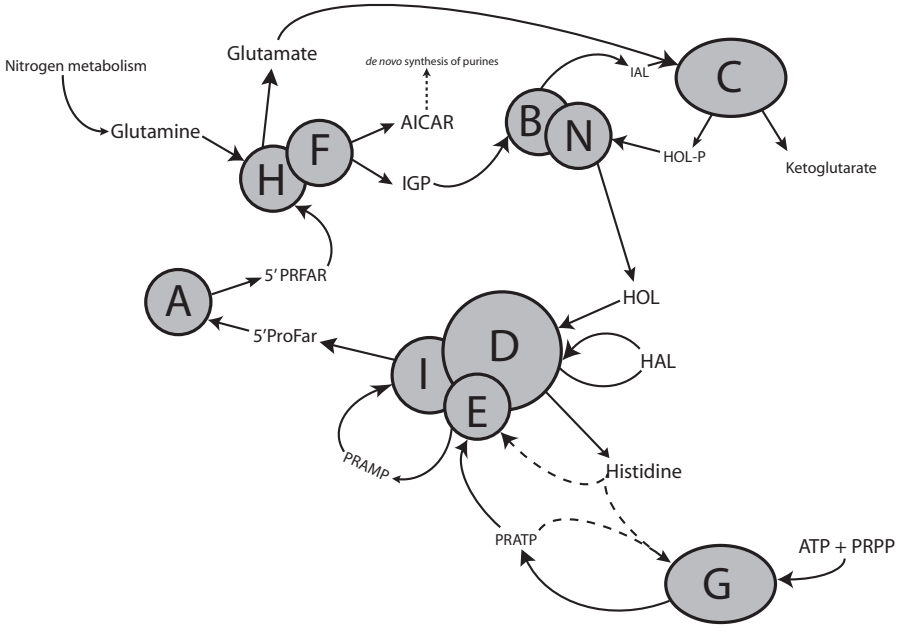


Figure 3.1: The histidine biosynthetic pathway.

AICAR, which is then recycled into the de novo purine biosynthetic pathway, and imidazole glycerol phosphate (IGP), which in turn is then transformed into histidine (Figure 3.1). Histidine biosynthesis is connected to nitrogen metabolism by a glutamine molecule, which is believed to be the source of the final nitrogen atom of the imidazole ring of IGP. The important role played by histidine biosynthesis in cellular metabolism is in fact underscored by the considerable energy (41 ATP molecules) that is required for the synthesis of each histidine molecule. The analysis of several completely sequenced genomes have disclosed many examples of elongation, duplication and/or fusion events involving different his genes. Interestingly, in some species, more than one enzymatic function is encoded by the same bi- or multifunctional cistron, such as *hisD*, *hisNB*, *hisHF*, and *hisIE* in some prokaryotes, HIS4 and HIS7 in eukaryotes. These multifunctional genes very likely are the outcome of fusion events. It has also been demonstrated that gene duplication also played a key role in shaping histidine biosynthesis. Indeed, *hisA* and *hisF* are the outcome of a cascade of gene elongation (i.e., an in-tandem gene duplication followed by the fusion of the two copies) and duplication events, and *hisH* was very likely recruited from other metabolic pathways. Noteworthy, after the assembly of the entire pathway, the structure and/or organization of his genes underwent major rearrangements in the three domains, generating a wide variety of structural and/or clustering strategies of his genes. Thus, the analysis of the structure and organization of his genes could help investigating the general problem of the origin and evolution of operons. The whole body of data available led to the assumption that the entire biosynthetic pathway was assembled long before the appearance of LUCA. However, it is still not clear how these genes were organized in the genome of the LUCA community, which was their structure and how many functions they performed. This is mainly due to the fact that the analysis of the structure and or-

ganization of *his* genes has been focused on bacterial genomes, especially proteobacteria. Thus, the aim of this part of the work was to give a further insight into the molecular mechanisms that have played a major role in shaping the histidine biosynthetic pathway; in this context we evaluated:

1. The role of gene fusions in the evolution of the histidine metabolic pathway.

2. The organization of histidine genes in prokaryotes in order to try to infer the structure and organization of histidine genes in the LUCA, and to try to understand the forces driving the organization of *his* genes in the different phylogenetic lineages; we approached this issue by analyzing the structure and organization of *his* genes in the third domain of life, Archaea.

3. The degree of conservation of *his* genes structure and organization within a bacterial genus. This issue was fulfilled in order to check whether a different lifestyle might have influence the structure, organization and regulation of *his* genes. To this purpose, we performed a structural, evolutionary and genetic analysis of histidine biosynthetic *core* in the genus *Burkholderia*, since this genus is a complex taxonomic unit embedding strains/species from different origins (environmental, clinical, etc.).

3.1 The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case

One of the major routes of gene evolution is the fusion of independent cistrons leading to bi- or multifunctional proteins. Gene fusions provide a mechanism for the physical association of different protein domains that might be catalytic or regulatory. It is widely accepted that this molecular mechanisms played a key role during the evolution and the assembly of genes and genomes although, a clear picture of its impact on the evolution of entire metabolic routes has been provided only in few cases (e.g. tryptophan).The aim of this work is to evaluate the overall role that gene fusion(s) might have had in the context of the assembly and evolution of histidine biosynthetic route, and to understand the biological significance of each fusion. For this purpose we performed a detailed analysis of *his* gene fusions in available genomes to understand the role of gene fusions in shaping histidine pathway. Our analyses on HisA structures across different lineages revealed that several gene elongation events are at the root of this protein family: internal duplication have been identified by structural superposition of the modules composing the TIM-barrel protei. Moreover several other *his* gene fusions happened in distinct taxonomic lineages; *hisNB* originated within γ-proteobacteria and after its appearance it was transferred to *Campylobacter* species (δ-proteobacteria) and to some Bacteria belonging to the CFB group. The transfer involved the entire *his* operon. The *hisIE* gene fusion was found in several taxonomic lineages and our results suggest that it probably happened several times in distinct lineages. Gene fusions involving *hisIE* and *hisD* genes (*HIS4*) and *hisH* and *hisF* genes (*HIS7*) took place in the Eukarya domain; the latter has been transferred to some δ-proteobacteria. In conlcusion, although gene duplication is probably the

most widely known mechanism responsible for the origin and evolution of metabolic pathways we showed that, several other mechanisms might concur in the process of pathway assembly and gene fusion appeared to be one of the most important and common.

BMC Evolutionary Biology

BioMed Central

Research

Open Access

The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case

Renato Fani*[1], Matteo Brilli[1], Marco Fondi[1] and Pietro Lió[2]

Address: [1]Dept. of Animal Biology and Genetics, via Romana 17, 50125 Florence, Italy and [2]Computer Laboratory, University of Cambridge, CB3 0FD, Cambridge, UK

Email: Renato Fani* - renato.fani@unifi.it; Matteo Brilli - matteo.brilli@dbag.unifi.it; Marco Fondi - marco.fondi@unifi.it; Pietro Lió - pl219@cam.ac.uk

* Corresponding author

from Second Congress of Italian Evolutionary Biologists (First Congress of the Italian Society for Evolutionary Biology)
Florence, Italy. 4–7 September 2006

Published: 16 August 2007

BMC Evolutionary Biology 2007, **7**(Suppl 2):S4 doi:10.1186/1471-2148-7-S2-S4

This article is available from: http://www.biomedcentral.com/1471-2148/7/S2/S4

Abstract

Background: Histidine biosynthesis is one of the best characterized anabolic pathways. There is a large body of genetic and biochemical information available, including operon structure, gene expression, and increasingly larger sequence databases. For over forty years this pathway has been the subject of extensive studies, mainly in *Escherichia coli* and *Salmonella enterica*, in both of which details of histidine biosynthesis appear to be identical. In these two enterobacteria the pathway is unbranched, includes a number of unusual reactions, and consists of nine intermediates; *his* genes are arranged in a compact operon (*hisGDC [NB]HAF [IE]*), with three of them (*hisNB*, *hisD* and *hisIE*) coding for bifunctional enzymes. We performed a detailed analysis of *his* gene fusions in available genomes to understand the role of gene fusions in shaping this pathway.

Results: The analysis of HisA structures revealed that several gene elongation events are at the root of this protein family: internal duplication have been identified by structural superposition of the modules composing the TIM-barrel protein.

Several *his* gene fusions happened in distinct taxonomic lineages; *hisNB* originated within γ-proteobacteria and after its appearance it was transferred to *Campylobacter* species (ε-proteobacteria) and to some Bacteria belonging to the CFB group. The transfer involved the entire *his* operon. The *hisIE* gene fusion was found in several taxonomic lineages and our results suggest that it probably happened several times in distinct lineages.

Gene fusions involving *hisIE* and *hisD* genes (*HIS4*) and *hisH* and *hisF* genes (*HIS7*) took place in the Eukarya domain; the latter has been transferred to some δ-proteobacteria.

Conclusion: Gene duplication is the most widely known mechanism responsible for the origin and evolution of metabolic pathways; however, several other mechanisms might concur in the process of pathway assembly and gene fusion appeared to be one of the most important and common.

Background

Histidine biosynthesis is one of the best characterized anabolic pathways. There is a large body of genetic and biochemical information available, mainly for *Escherichia coli* and *Salmonella enterica*, including operon structure, gene expression, and growing sequence data [1]. In these two enterobacteria, the pathway is the same, unbranched, includes a number of unusual reactions, and consists of nine intermediates; *his* genes are arranged in a compact operon (*hisGDC [NB]HAF [IE]*), with three of them (*hisNB*, *hisD* and *hisIE*) coding for bifunctional enzymes (Figure 1) [2,3].

Histidine biosynthesis is a metabolic cross-road and plays an important role in cellular metabolism being interconnected to both the *de novo* synthesis of purines and to nitrogen metabolism. The connection to purine biosynthesis results from an enzymatic step catalyzed by imidazole glycerol phosphate (IGP) synthase, a heterodimeric protein composed by one subunit each of the *hisH* and *hisF* products [2]. This heterodimeric enzyme catalyzes the transformation of N'-(5'-phosphoribosyl)-formimino-5-aminoimidazol-4-carboxamide ribonucleotide (PRFAR) into 5'-(5-aminoimidazole-4-carboxamide) ribonucleotide (AICAR), which is recycled into the *de novo* purine biosynthetic pathway, and IGP, which leads to histidine. The important connection to nitrogen metabolism is due to a glutamine molecule, the source of the final nitrogen atom of the imidazole ring of IGP. Chemical and biological evidences suggest that histidine was formed during the long period of chemical abiotic synthesis of organic compounds and the monophyly of the three cell domains in phylogenetic trees of concatenated His proteins, suggests that this biosynthetic route is ancient. The chemical syntheses of histidine [4], prebiotic analogues of histidine [5], and of histidyl-histidine under primitive conditions has been reported [6], as well as the role of the latter in the enhancement of some possible prebiotic oligomerization

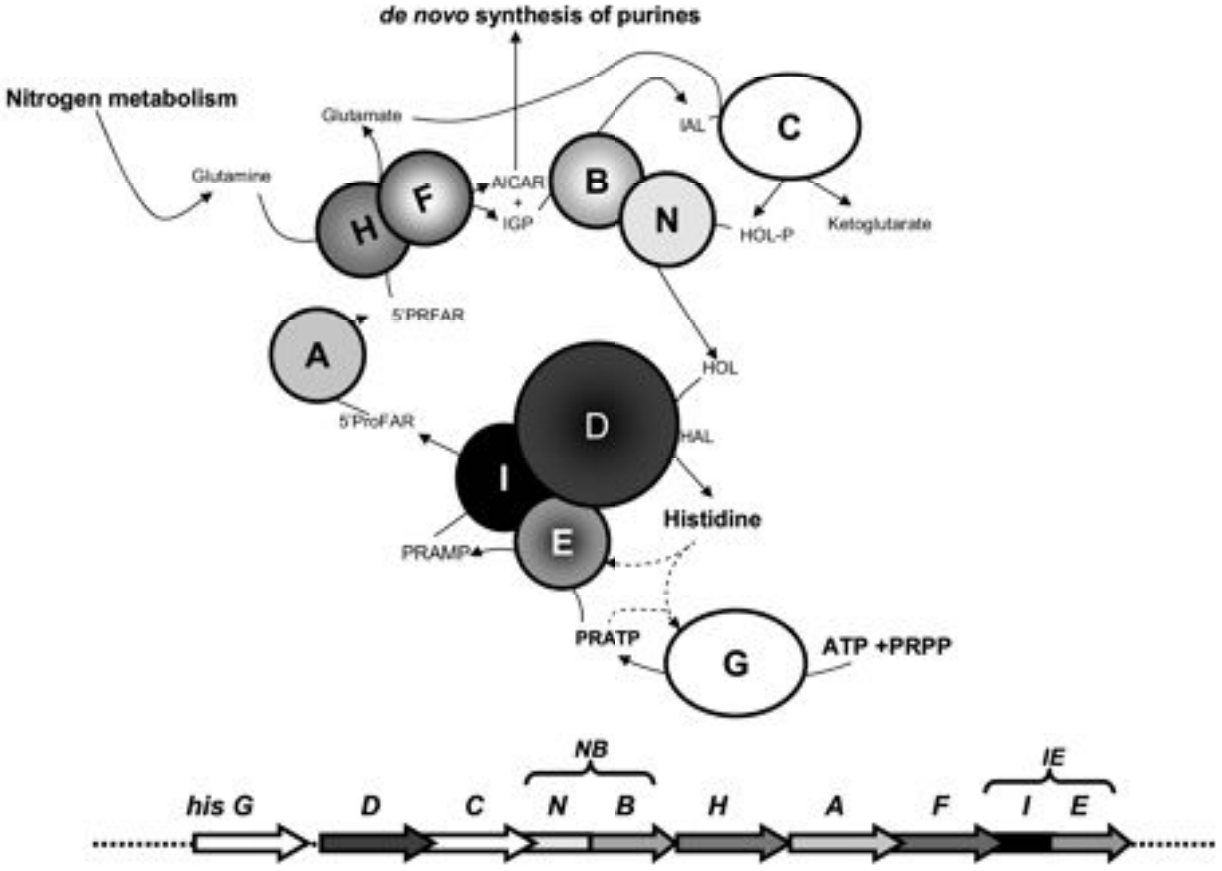


Figure 1
Summary of Histidine biosynthesis. Schematic representation of the histidine biosynthetic pathway and the organization of *his* gene in *Escherichia coli*. Genes and proteins in color are those involved in fusion events.

reactions involving amino acids [7] and nucleotides [8]. It is therefore reasonable to assume that His-containing small peptides could have been involved in the prebiotic formation of other peptides and nucleic acid molecules, once these monomers accumulated in primitive tidal lagoons or ponds. If primitive catalysts required histidine, then the eventual exhaustion of the prebiotic supply of histidine and histidine-containing peptides [4,6,8] imposed a selective pressure favoring those organisms capable of synthesizing histidine. Hence this metabolic pathway might have been assembled long before the appearance of the Last Universal Common Ancestor (LUCA) and the wet-lab and bioinformatics work carried out by our group in the last 15 years strongly supported this thesis [2,9-14]. A wide variety of clustering strategies of *his* genes have been documented [10]; moreover, an impressive series of well characterized duplication, elongation, and fusion events has shaped this pathway. Therefore, the histidine biosynthetic pathway represents a very good model for understanding the molecular mechanisms driving the assembly and refining of metabolic routes.

It is worth noting that at least seven genes, namely *hisD*, *hisB*, *hisN*, *hisI*, *hisE*, *hisF*, and *hisH* underwent fusion events in different phylogenetic lineages [11-13,15]. Gene fusions provide a mechanism for the physical association of different protein domains that might be catalytic or regulatory [16]. Moreover, fusions frequently involve genes coding for proteins that function in a concerted manner, such as enzymes catalyzing sequential steps within a metabolic pathway [17]. Fusion of such catalytic centers likely facilitates the channeling of intermediates [16]; the high fitness of gene fusions can also rely on the tight regulation of the expression of the fused domains.

Besides, a special case of gene fusion has played a key role in the evolution of ancestral genes: several proteins have been shown to be the outcome of coupled "duplication and fusion" events (gene elongation). The outcome of such an event is a gene with two paralogous moieties (modules) that might undergo further duplication events, leading to a gene with several internal repetitions. Gene elongation events allow improving a protein's function by increasing the number of active sites and/or the acquisition of an additional function by modifying a redundant segment. The most documented example pertains two *his* genes, *hisA* and *hisF*, encoding two $(\beta/\alpha)_8$ barrel (TIM-barrel) proteins [18].

The aim of this work is to evaluate the overall role that gene fusion(s) might have had in the context of the assembly and evolution of histidine biosynthetic route, and to understand the biological significance of each fusion. For this purpose, the structure and organization of all the available *his* genes that underwent fusion event(s) were analyzed using statistical and bioinformatics methods.

Results and discussion

A cascade of gene elongations and duplications: hisA *and* hisF

The two genes *hisA* and *hisF* code for a [N-(5'-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) isomerase and a cyclase, respectively, which catalyze two central and sequential reactions (the fourth and fifth ones) of the pathway (Figure 1) and belong to the TIM-barrel family of proteins. The comparative analysis of the HisA and HisF proteins from different archaeal, bacterial, and eukaryotic (micro)organisms revealed that they are paralogous and share a similar internal organization into two paralogous modules half the size of the entire sequence [18]. Comparison of these modules led to the suggestion that *hisA* and *hisF* are the result of two ancient successive duplications, the first one involving an ancestral module half the size of the present-day *hisA* gene and leading (by a gene elongation event) to the ancestral *hisA* gene, which in turn underwent a duplication that gave rise to the ancestor of *hisF* [18].

The barrel structure is composed by eight concatenated (βstrand)-loop-(α-helix) units. The βstrands are located in the interior of the protein, forming the staves of a barrel, whereas the α-helices pack around them facing the exterior. According to the model proposed [18,19] the ancestral half-barrel gave a functional enzyme by homodimerization. The elongation event leading to the ancestor of *hisA*/*hisF* genes resulted in the covalent fusion of two half-barrels producing a protein whose function was refined and optimized by mutational changes; once assembled, the "whole-barrel gene" underwent gene duplication, leading to the ancestor of *hisA* and *hisF* [18,19].

The structural symmetry of the TIM barrel has prompted us to investigate the possibility of an even older gene elongation event involving (β/α)-mers smaller than the $(\beta/\alpha)_4$ units of the ancestral "half-barrel" precursor. To this purpose an extensive analysis of all the available HisA and HisF sequences was carried out. This analysis was performed by splitting each HisA sequence into four modules (HisA1-HisA4) following secondary structures succession in the corresponding protein from *Thermotoga maritima*, whose three-dimensional structure is available (we will refer to these four regions as the "quarters"). The alignments concerning *Methanocaldococcus jannaschii* are shown in Figure 2 (identity and similarity values are in Table 1). The degree of sequence similarity is not very high but it might be used to support an evolutionary model suggesting that the present day situation could have been reached after two gene elongation events, each one dou-

```
                  10        20        30        40        50        60
                   |         |         |         |         |         |
mjHisA1     MIIIPAVDLKDKKCVQLIQGDPNKKHLELNNVE-VAKKFVDEGAE-YLHIIDLDAAFG-TGNN
mjHisA2     RDVIKN--IIKEVNVPVEVGGGIRNLEIAKELISLGVDRVIVGTKAILEPKFIDDLNKEIGKD
mjHisA3     KIVLA----VECKEGKVVIKG-WKEKVDKTPIEVIKEFEDKVGYILFTN-VDVEGLLKGIN--
mjHisA4     VDIIK--ELIEKTDIPIIYSGGITTLEDIKALKELGIYGVVIGSALYKGLIDLKKALEIVKN-
              ::    :::: : ::  *  ::  : : :  :    * :*:  :   :*::   : :

mjHisA1     MIIIPAVDLKDKKCVQLIQGDPNKKHLELNNP-VEVAKKFVDEGAE-YLHIIDLDAAFG-TGNN
mjHisA2     RDVIKNI-IKEVNVPVEVGG--GIRNLEIAKELISLGVDRVIVGTKAILEPKFIDDLNKEIGKD
              :*  : :*: :    : *    ::**: :  : :    *  *::  *    :*     *::

mjHisA1     MIIIPAVDLKDKKCVQLIQGDPNKKHLELNNPVEVAKKFVDE-GAEYLHIIDLDAAFGTGNN
mjHisA3     -KIVLAVECKEGKVV--IKG--WKEKVD-KTPIEVIKEFEDKVGYILFTNVDVEGLL-KGIN
              *: **: *: * *  *:*   *:::: : *:** *:* *:     :  :*::  :  * *

mjHisA1     MIIIPAVDLKDKKCVQLIQGDPNKKHLE-LNNPVEVAKKFVDEGAE-YLHIIDLDAAFG-TGNN
mjHisA4     VDIIK--ELIEKTDIPIIYSG-GITTLEDIKALKELGIYGVVIGSALYKGLIDLKKALEIVKN-
            : **   :* :*  : :*        ** ::   *:    *  *:  *  :***  *:    *

mjHisA2     RDVIKNI-IKEVNVPVEVGG--GIRNLEIAKELISLGVDRVIVGTKAILEPKFIDDLNKEIGKD
mjHisA3     -KIVLAVECKEGKVVIK-GWKEKVDKTPI--EVIKEFEDK--VG-YILFTNVDVEGLLKGIN--
              ::  :  ** :* :: *    : :  *  *:*    *:  **   ::    :: * * *

mjHisA2     RDVIKNIIKEVNVPVEVGGGIRNLEIAKELISLGVDRVIVGTKAILEPKFIDDLNKEIGKD
mjHisA4     VDIIKELIEKTDIPIIYSGGITTLEDIKALKELGIYGVVIGS-ALYKGLIDLKKALEIVKN
             *:**::*:: ::*:   ***  **  * *  **:  *::*: *: :  :      ** *:

mjHisA3     -KIVLAVECK-EGKVVIKGWKEKV-DKTPIEVIKEFEDKVGYILFTN-VDVEGLLKGIN-
mjHisA4     VDIIKELIEKTDIPIIYSGGITTLEDIKALKELGIYGVVIGSALYKGLIDLKKALEIVKN
              *:  :  * :  ::  *    : *   :: :  :   :*  *:   :*::  *: ::
```

Figure 2
HisA "quarters" alignments. Pairwise alignments of *Methanocaldococcus jannaschii* (mj) HisA subregions corresponding to "quarters" of barrels (named HisA1, HisA2, HisA3, HisA4). Symbols: *,: correspond to identical or similar aminoacids, respectively.

bling the length of the ancestral gene and the number of $(\beta\alpha)$-modules in the product. Thus the HisA TIM-barrel would be the result of a cascade of two consecutive gene elongations (Figure 3). This model is supported by the structures shown in Figure 3 (upper left panel). We used the *T. maritima* HisA structure (1Q02) to obtain the coordinates of each atom composing the "quarters" of barrel and then performed a structural superposition using swiss PDB viewer. Lang et al., [19] compared the structure of the two HisA half-barrels and obtained a *root-mean-square* deviation (rms) ranging from 1.5 to 2.0 Å using all main chain, non hydrogen atoms. Our results concerning "quarters" structural superpositions showed for the first and the second quarters an average rms of 1.21 Å using all alpha carbons, strongly supporting the model proposed. We showed examples from these two organism, because *M. jannaschii* with several other Archaea showed the best overall degree of conservation of internal repetitions, while the choice of *T. maritima* followed the availability of HisA tridimensional structure [19].

Table 1: Identity and similarity values for the HisA quarters comparisons.

	HisA quarters comparisons			
	hisA1	*hisA2*	*HisA3*	*hisA4*
HisA1		15.8	29	25.4
HisA2	42.9		20.6	32.2
HisA3	58.1	42.9		13
HisA4	42.9	58.1	41.9	

Identity (upper diagonal) and similarity percentages calculated for the alignments shown in Figure 2.

The symmetry of the TIM-barrel structure suggests to test a further ancestral duplication in which the original gene coded for a single (β/α)-module, capable of forming a homo-octamer to form a complete barrel. Although the alignments constructed from the eight single (β/α)-modules are very short, they still contain a non negligible amount of sequence and secondary structure similarities not expected from random distribution of amino acids and by visual comparison of their structure is in agreement with this hypothesis (Figure 3, lower left panel). Thereby, the ancestral forms of life might have expanded their coding abilities and their genomes by duplicating a small number of mini-genes, i.e. the "starter types". We are completely aware that the evidence of the ancient duplications involving an ancestral mini-gene encoding the "quarter" of barrel and the single (β/α)-module is based on very limited amount of sequence and structural similarities; in spite of this, in our opinion, the hypothesis mentioned above remains valid. Moreover, further studies are needed to clarify the presence of these additional gene

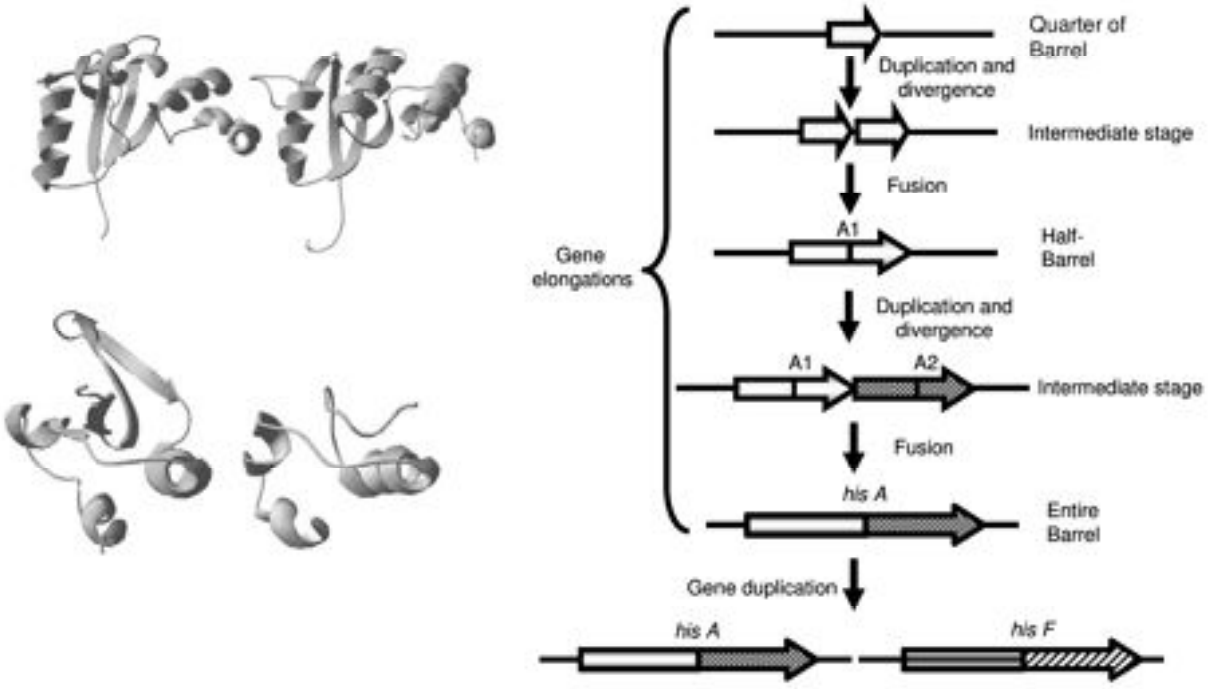


Figure 3
An evolutionary model for *hisA* and *hisF* genes. Right-most panel is the evolutionary model that we propose and discuss in the text concerning *hisA* and *hisF* origin and evolution. Panels on the left are: the first and the second quarters (top) and two single (*βα*) modules of the HisA protein from *Thermotoga maritima* which illustrates the structural similarities from which we derived our model. The quarters have a structural alignment with only 1.2 Å of RMS on 104 alpha carbons.

elongation events and to integrate them into a more general picture of the evolution of the very diversified TIM-barrel family of proteins.

The hisNB *fusion*

The eighth and sixth steps of histidine biosynthesis are catalyzed by histidinol-phosphate phosphatase (EC 3.1.3.15) (HOL-Pase) and IGP dehydratase (EC 4.2.1.19), respectively [1]. Distinct HOL-Pases have been characterized in different organisms, whereas IGP dehydratase is the same in all known histidine synthesizing organisms. In *E. coli* the two activities are coded for by a single gene, referred to as *hisNB* [11]: the N-terminal domain (HisN) is a phosphatase belonging to the DDDD family [11] and the C-terminal domain is responsible for IGP dehydratase activity. The evolutionary history of the *hisNB* gene has been recently reported by [11] who showed that *hisNB* gene fusions are present in most γ-proteobacteria and in the ε-proteobacterium *Campylobacter jejuni;* phylogenetic analysis allowed to trace the fusion event in an ancestor of the γ-subdivision and its later horizontal transfer to *C. jejuni*. Moreover, *hisN* is paralogous to *gmhB* (*E. coli* nomenclature), catalyzing the dephosphorylation of D-α-D-heptose 1,7-PP for surface Lipopolysaccharide production [20,21].

Since the former *hisNB* evolutionary model was based on a limited number of genomes, we update it including all available genomes (April, 1: 41 Archaea, 759 Bacteria and 135 Eukarya).

By combining results obtained with several queries, we retrieved 131 orthologous bifunctional HisNB sequences: 104 come from γ-proteobacteria, 9 from ε-proteobacteria, 1 to a α-proteobacterium and 17 from the CFB group (Table 2 and Additional file 1). No archaeal and eukaryo-

Table 2: Phylogenetic distribution of *HisIE* and *HisNB* genes.

Domain	Phylum	Class	# His+	% HisIE	% HisNB
Bacteria	Acidobacteria	Acidobacteria	1	0	0
		Solibacteres	1	0	0
	Actinobacteria	Actinobacteria	36	2.8	0
	Aquificae	Aquificae	1	100	0
	Bacteroidetes	Bacteroidetes	2	100	50
		Flavobacteria	1	100	100
		Sphingobacteria	1	100	100
	Chlorobi	Chlorobia	4	0	0
	Chloroflexi	Dehalococcoidetes	2	0	0
	Cyanobacteria	Chroococcales	10	90	0
		Gloeobacteria	1	100	0
		Nostocales	2	100	0
		Oscillatoriales	1	100	0
		Prochlorales	10	100	0
	Firmicutes	Bacilli	37	48.7	0
		Clostridia	10	80	0
	Planctomycetes	Planctomycetacia	1	0	0
	Proteobacteria	α	52	1.9	1.85
		β	39	0	0
		δ	12	16.7	0
		ε	7	85.7	82
		γ	89	68.5	95
	Spirochaetes	Spirochaetes	4	0	0
	TD group	Deinococci	4	100	0
	Thermotogae	Thermotogae	1	100	0
Archaea	Crenarchaeota	Thermoprotei	6	0	0
	Euryarchaeota	Archaeoglobi	1	0	0
		Halobacteria	4	0	0
		Methanobacteria	2	0	0
		Methanococci	3	0	0
		Methanomicrobia	8	0	0
		Methanopyri	1	0	0
		Thermococci	2	100	0
		Thermoplasmata	1	100	0

Phylogenetic distribution analysis of the fusions HisIE and HisNB, along with the percentage of species possessing one of them (two right – most columns) on the number of histidine producing organisms (as assessed by the presence of his genes in their genome). Eukarya are treated separately (see HIS4 section) and never possess the *his*NB gene fusion.

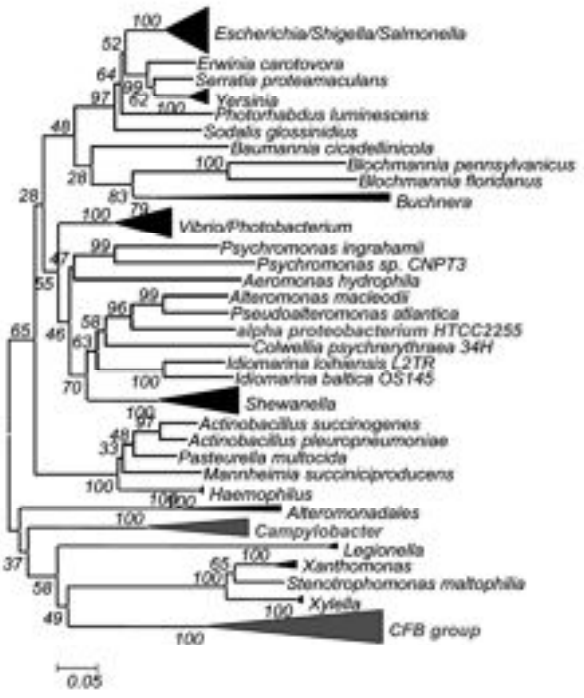


Figure 4
HisNB phylogenetic analysis. HisNB Phylogenetic tree. Organisms (groups) in red are bacteria not belonging to γ-proteobacteria and harboring the HisNB fusion.

tic bifunctional sequence was retrieved although they possess genes encoding DDDD-type phosphatases, or, more generally HAD hydrolases [11]. These data confirmed the narrow phylogenetic distribution of the *hisNB* fusion, which is mostly present in γ-proteobacteria. However, the occurrence of a fused *hisNB* gene in other lineages enlarged its distribution raising the question of the origin of this fusion in these phyla, i.e. if it is either the outcome of convergent evolution or a horizontal gene transfer event (HGT). To discern between these two different scenarios, a phylogenetic analysis of HisNB sequences was carried out. A phylogenetic tree obtained using a representative set of HisNB sequences is reported in Figure 4, which shows that HisNB sequences from α- and ε-proteobacteria, and CFB bacteria are intermixed with γ-proteobacterial sequences and do not reflect the 16S rDNA phylogeny. This result strongly suggests that the *hisNB* gene has been horizontally transferred from some γ-proteobacteria to the other microorganisms. It is also quite possible that the transfer event might have involved the entire *his* operon or part thereof, as evidenced by phylogenetic trees of other *his* genes (e.g. Additional File 2). This statement relies on the analysis of the organization of *his* genes in the 131 genomes harboring the *hisNB* fusion, which revealed (Additional file 3) that all of them are localized within more or less compact operons.

The hisIE *fusion*

The *hisI* and *hisE* genes code for a phosphoribosyl-ATP phosphohydrolase and a phosphoribosyl-AMP cyclohydrolase that are responsible for the third and second steps in histidine biosynthesis, respectively. In *E. coli* and *S. enterica* the two genes are fused to form the last gene of the *his* operon (Figure 1).

The phylogenetic distribution of *hisIE* genes was obtained by retrieving homologous sequences using the *E. coli* HisIE amino acid sequence as a query to probe genome database. The data are reported in Table 2 and can be summarized as follows (see also Additional file 4):

1. The *hisIE* fusion is not universally distributed;

2. Bifunctional *hisIE* genes were found in all eukaryotes (see section regarding *HIS4*);

3. Most of the archaeal genomes harbor monofunctional *hisI* and *hisE* genes; the occurrence of *hisIE* in Thermococci and Thermoplasmata is very likely the outcome of a HGT event from a bacterium donor [15]. Moreover, when the *hisI* and *hisE* genes are not fused in Archaea, they do not belong to operons and are separated on the chromosome. The only exceptions are represented by *Sulfolobus* species, where the two genes are in a compact operon but separated by the *hisH* gene;

4. The *hisIE* gene fusion is present in 100% of the histidine producing organisms belonging to Aquificae, Deinococci, Bacteroidetes, Cyanobacteria, and Thermotogae. Moreover, a bifunctional *hisIE* gene was found in all γ-proteobacteria that branched off after the separation of Pseudomonadales from the main branch. The presence of the fusion in ε-proteobacteria can be explained by means of a HGT of the entire operon [13] from γ-proteobacteria; the same appears to be true for Bacteria belonging to the CFB group and possessing the *hisNB* gene fusion (see the corresponding paragraph). In spite of the high number of genomes sequenced (39), no *hisIE* fusion was found in β-proteobacteria, which represent the key-point for the compacting of *his* genes during the construction of proteobacterial *his* operon [13].

5. Firmicutes show a complex scenario: we have found fused and stand-alone genes in very closely related species of *Bacillus* (i.e. *Bacillus subtilis* possesses the gene fusion while *B. thuringiensis* and *B. anthracis* do not; in these cases *hisI* and *hisE* are contiguous and very close on the chromosome); the same is true for Clostridia. The presence of the fusion in model organisms such as *Bacillus subtilis* but not

in some of the recently sequenced genomes of the same genus suggests that sequencing artifacts, probably favored by the gene organization of these two genes, might explain this situation.

6. Actinobacteria lack this gene fusion.

7. Apparently there is no correlation between the occurrence of *hisIE* fusion and *his* genes organization. However, it is interesting that during Proteobacteria evolution, we witness a progressive approaching of two initially far *hisI* and *hisE* cistrons, starting from δ- and ε-proteobacteria. The distance between them decreases in bacteria belonging to the α-branch; then, they partially overlap in β-proteobacteria, a molecular event which is coincident with the formation of a complete *his* operon, which very often includes genes apparently not involved in histidine biosynthesis. Finally, the two genes fused in the ancestor of γ-proteobacteria, where the compactness of *his* operon is very high [13].

Despite of the proposed HGTs, the current phylogenetic distribution of the HisIE bifunctional enzyme evokes a scenario of convergent evolution and independent gene fusions/splitting of the two cistrons in different lineages. However, phylogenetic analyses are not of great help to confirm this view (data not shown) because these proteins are very short (less than 100 residues each) and the informative sites in a multialignment of sequences coming from complete genomes are extremely reduced bringing to unreliable trees (i.e. very low bootstrap support, star topologies, data not shown). On a different perspective, the analysis of the linker region connecting the two domains might help in understanding the evolutionary history of these fusions, but we have found that they are very short, giving no information on this issue (data not shown). If the idea on convergent evolution of these gene fusions will hold future works, it might turn out that there is a strong selective pressure favoring their appearance in different lineages. Lastly, the finding that the gene order in all the bifunctional genes is always *hisI*, *hisE*, suggests that a different arrangement of the two domains should be disadvantageous for the enzymatic activity and that structural and/or functional constraints might be responsible for the extant arrangement.

From metabolons to multifunctional enzymes: the eukaryotic HIS7 and HIS4 fusions

Two fusions involving *his* genes were disclosed in the yeast *Saccharomyces cerevisiae*: *HIS7*, corresponding to the bacterial *hisH* and *hisF* genes [12], and *HIS4*. The latter gene codes for a tetrafunctional enzyme consisting of about 800 residues and containing three regions homologous to the prokaryotic *hisI*, *hisE* and *hisD* genes, arranged in this order in the yeast gene, whose products perform the second, the third and the last two steps of histidine biosynthesis, respectively.

The IGP synthase coding gene: HIS7

The bifunctional enzyme IGP synthase catalyzes the fifth step of histidine biosynthesis, generating the imidazole ring of the histidine precursor, IGP. The overall reaction is the conversion of PRFAR to IGP and AICAR [2] *via* a glutamine molecule, and with no free intermediate (Figure 1). This represents the central step of the pathway, which connects histidine biosynthesis to nitrogen metabolism and the *de novo* synthesis of purines, through AICAR. The active form of the *E. coli* IGP synthase is a heterodimer of the *hisH* and *hisF* products i.e. a glutamine amidotransferase (GAT) and a cyclase, respectively [22]. The requirement for a direct interaction between GAT and the cyclase was confirmed by the discovery of the structure of the *S. cerevisiae* HIS7 gene coding for IGP synthase [23]: the analysis of the encoded enzyme demonstrated that it is constituted by two domains, an N-terminal and a C-terminal one, sharing a high degree of sequence similarity with known HisH and HisF, respectively. Previous works suggested that the eukaryotic *HIS7* gene is the outcome of a fusion event involving two monofunctional, bacterial-like genes [12]. According to the model proposed, the eukaryotic lineage inherited two monofunctional genes, *hisH* and *hisF*, that underwent gene fusion. The alternative scenario, that is the possibility of a HGT event from a prokaryote harboring a fused *hisHF* gene to eukaryotes was excluded on the basis of the absence of the HisHF fusion in prokaryotes. However, the increasing number of sequence in databases opens the possibility to modify this model. To this purpose the *S. cerevisiae HIS7* aminoacid sequence was used in a BLASTP search [24], allowing to retrieve 21 bifunctional sequences. Whilst most (18) come from Eukarya, three of them belonged to two δ-proteobacteria, raising the possibility that the bacterial and the eukaryotic HisHF sequences share a common ancestry, even though a phenomenon of convergent evolution could not be ruled out. Thus, all the available HisHF bifunctional sequences and a set of bacterial and archaeal concatenated HisH and HisF sequences were aligned using the program ClustalW [25]. The multialignment (partially shown in Figure 5) allowed to detect several conserved insertions that distinguish bifunctional proteins. This speaks towards a common origin of the eukaryotic and bacterial *hisHF* genes rather than a phenomenon of convergent evolution. The phylogenetic tree obtained using the above mentioned multialignment showed in Figure 6 supports this view: it can be splitted into two main clusters, one containing all the monofunctional sequences including the one coming from the δ-proteobacterium *Geobacter sulfurreducens*, the second one comprising all the bifunctional eukaryotic and bacterial sequences; the three bifunctional bacterial sequences clus-

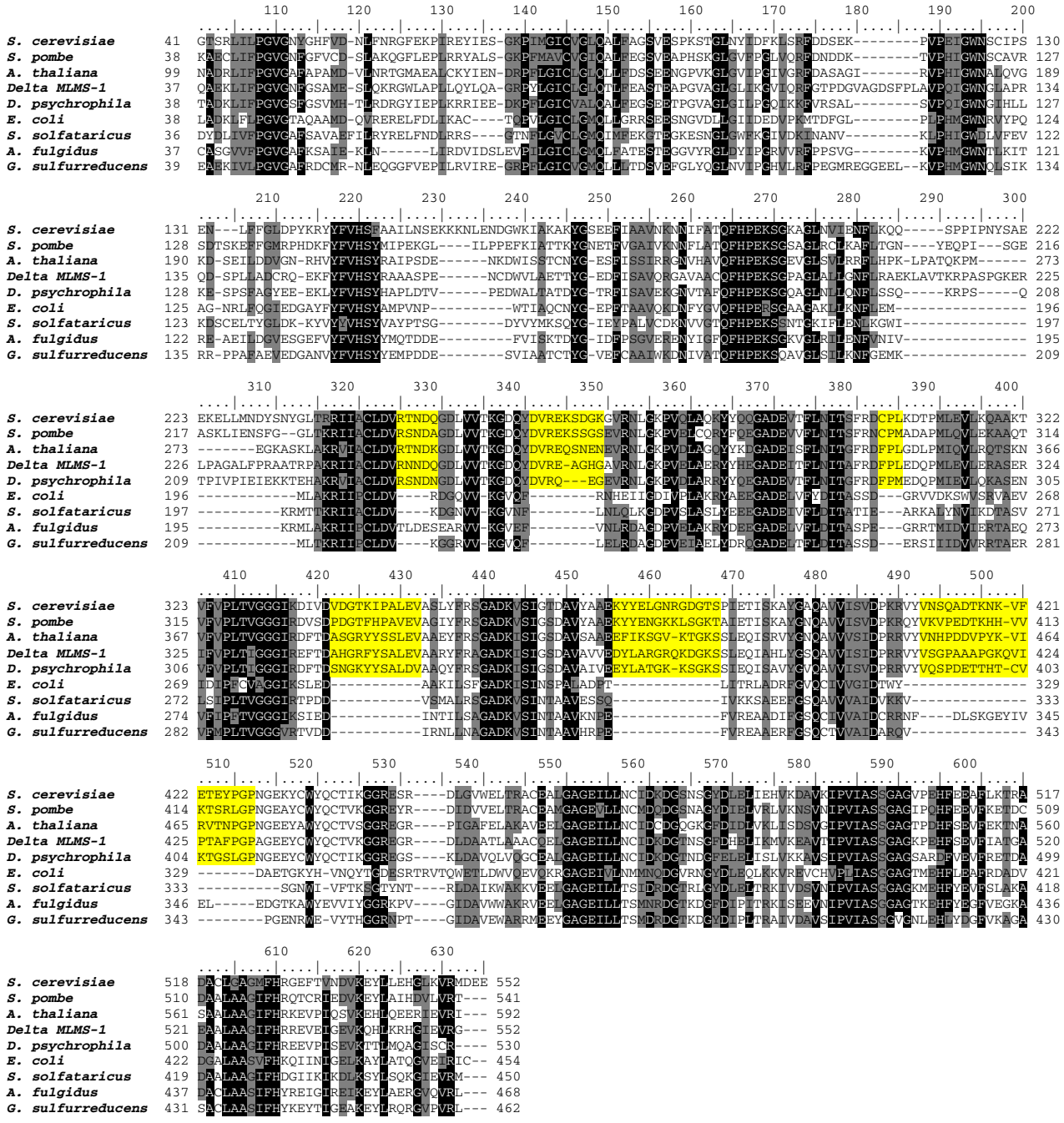


Figure 5

His7 multialignment. A multialignment of HIS7, HisHF and a set of representative concatenated HisH and HisF sequences from *E. coli*, *S. solfataricus*, *A. fulgidus* and *G. sulfurreducens*. Yellow shading represent insertions shared only by bifunctional HIS7 and HisHF proteins. Shading of the multialignment has been made with PAM250 matrix.

Figure 6
His7 phylogenetic analysis. Phylogenetic tree of HIS7, HisHF and concatenated HisH and HisF sequences. See Methods for details on phylogenetic tree construction.

tered with Plants sequences. This body of data suggests that a bifunctional *hisHF* gene might have been transferred from Plants to some δproteobacteria.

A tetrafunctional gene: HIS4

The *S. cerevisiae HIS4* gene codes for a tetrafunctional enzyme and consists of three regions sharing a high degree of sequence similarity with prokaryotic HisI, HisE, and HisD, thus the activities performed by the *HIS4* enzyme are, from the N-terminal end, a phosphoribosyl-ATP pyrophosphohydrolase, a phosphoribosyl-AMP cyclohydrolase and the two-step histidinol dehydrogenase. The first one (HisI) uses N'-5'-phosphoribosyl-ATP (PR-ATP) to produce N'-5'-phosphoribosyl-AMP (PR-AMP), whose purine ring is subsequently opened by the second (HisE) to give 5'-ProFAR. This compound, in turn, undergoes seven additional enzymatic reactions leading to histidine, the last two of which are catalyzed by histidinol-dehydrogenase (HisD) (Figure 1) i.e. the double oxidation of histidinol to histidine, through the intermediate histidinal, concomitant to the reduction of two NAD+ molecules, with a Bi-Uni-Uni-Bi kinetic mechanism [26,27].

Sequence retrieval and hisI, hisE and hisD gene structure in Eukarya

The eukaryotic complete genomes database of protein sequences was probed using the HisIE and HisD domains of the *S. cerevisiae HIS4* enzyme (residues 134–329 and 351–795, respectively). The BLASTP [24] search returned the eukaryotic sequences listed in Table 3 and revealed that the two sequences used as queries retrieved two identical sequences, with the exception of *Aspergillus nidulans* (where an additional protein with a putative HisD function was retrieved), *Schizosaccharomyces pombe* and Plants, where genes corresponding to the prokaryotic *hisIE* and *hisD* counterparts were detected.

A multialignment of all the retrieved sequences with a representative set of archaeal and bacterial HisIE and HisD sequences (Additional file 5) revealed that the *HIS4*-like genes can be subdivided into four portions (Figure 7): i) an N-terminal region of variable length, followed by ii) the *hisIE* moiety, which in turn precedes iii) a linker region of variable sequence and length connecting the *hisIE* region to the last domain, iv) the *hisD* region. Plants sequences have an N-terminal region which is unrelated to any histidine biosynthetic enzymes and that might represent a signal sequence for chloroplast localization, an idea which is in agreement with a (at least partial) compartmentalization of histidine biosynthesis into these organelles, as previously suggested [28]. A similar search performed on prokaryotic databases did not allow retrieving any HIS4-like protein.

Analysis of the HIS4 N-terminal region

The fungal HIS4 sequences have an N-terminal domain (whose length ranges from 160 to 220 residues) that is much longer than that found in HisIE from Plants; moreover, we detected no significant homology neither with the signal sequence found in Plants nor with any other signal sequence of Fungi or other organisms (data not shown). This sequence has no known conserved domains, as appeared by the absence of hits in the Conserved Domain Database (data not shown). Moreover, a psi-blast search did not permit to obtain any statistically significant hit, if we exclude other HIS4 proteins (data not shown). However, both the presence of the corresponding sequence in mRNAs (see GenBank entry NM_212387.1 from *Ashbya gossypii*) and the molecular weight of the isolated *S. cerevisiae* HIS4 enzyme (95000 ± 500 Da) [29] speak toward the presence of this N-terminal sequence in the "final polypeptide". A structural rather than a catalytic role of this region can be suggested on the basis of alignment of the isolated N-terminal regions of the fungal HIS4 enzymes, which revealed that the degree of sequence similarity between them is quite low and significantly less than that shared by catalytic domains (Additional file 6 and 7).

Phylogenetic analyses

A phylogenetic analysis was performed to check whether HisIE and HisD proteins/domains listed in Table 3 experienced the same evolutionary history and whether the phylogenetic trees were congruent with the phylogeny of

Table 3: Phylogenetic distribution of *HIS4* genes.

Taxonomy			Organism	Strain	Protein GI	Length (aa)	Gene structure			
							*	I	E	D
Ascomycota	**Pezizomycotina**	**Eurotiomycetes**	*Aspergillus nidulans*	FGSC A4	40743835	438				X
					40746471	867	N	X	X	X
		Sordariomycetes	*Gibberella zeae*	PH-1	42547615	854	N	X	X	X
			Magnaporthe grisea	70–15	38109852	865	N	X	X	X
			Neurospora crassa		32420263	870	N	X	X	X
	S. mycotina; S. mycetales	**Dipodascaceae mitosporic S. mycetales**	*Yarrowia lipolytica*	CLIB99	50545145	855	N	X	X	X
			Candida albicans		3757752	838	N	X	X	X
			Candida glabrata	CBS138	50285163	802	N	X	X	X
		Saccharomycetaceae	*Debaryomyces hansenii*	CBS767	50424339	861	N	X	X	X
			Eremothecium gossypii		45199222	806	N	X	X	X
			Kluyveromyces lactis	NRRL Y-1140	50304609	795	N	X	X	X
			Pichia pastoris		3203	844	N	X	X	X
			Saccharomyces cerevisiae		10383761	799	N	X	X	X
		Schizosaccharomycetes	*Schizosaccharomyces pombe*		19112678	439				X
					19112622	417	N	X	X	
Basidiomycota	**Hymenomycetes**	**Heterobasidiomycetes**	*Cryptococcus neoformans*		50258877	852	N	X	X	X
			Hebeloma cylindrosporum		31095443	843	N	X	X	X
Viridiplantae	**Ustilaginomycetes**	**Brassicaceae**	*Ustilago maydis*	521	46099735	896	N	X	X	X
			Arabidopsis thaliana		10177677	467				X
					21554409	281	S	X	X	
			Brassica oleracea	var. capitata	99844	469				X
			Thlaspi goesingense		3982577	464				X
	Magnoliophyta	**Poaceae**	*Oryza sativa*	japonica	34904356	459				X
					34903270	202	S	X	X	

Summary of results obtained for HIS4 genes. Right-most columns indicate the presence of a given domain in the corresponding protein; column marked with an asterisk corresponds to the N–terminal region found in Fungi (N), of unknown function, and Plants (S), a signal peptide for chloroplast localization.

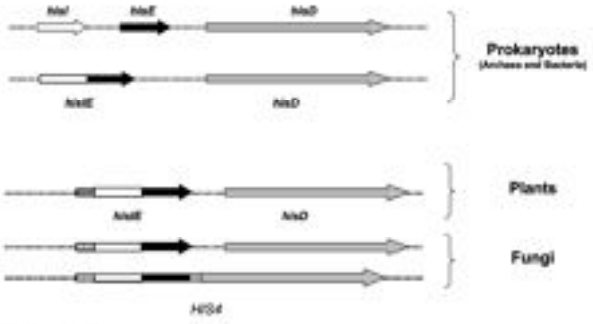


Figure 7
***HIS4*, *hisIE* and *hisD* genes.** Schematic representation of the archaeal, bacterial and eukaryotic genes coding for phosphoribosyl-ATP pyrophosphohydrolase (*hisI*), phosphoribosyl-AMP cyclohydrolase (*hisE*) and histidinol dehydrogenase (*hisD*). The HisI and HisE proteins are coded by a bifunctional gene or two separate cistrons in both Archaea and Bacteria and by a bifunctional gene in all Eukarya. Homologus regions are represented by the same hatching.

Eukarya based on other molecular markers, such as 18S rDNA. According to the eukaryotic phylogeny based on 18S rDNA [30] Viridiplantae branched off from the major eukaryotic line earlier than Fungi, that represent a sister group of Metazoa. Moreover, within Fungi, Basidiomycota appear to branch off earlier than Ascomycota; the latter are rooted by Archiascomycotina, including Schizosaccharomycetes. However, discordant phylogenies have been obtained using different sequences and this might be mainly due to HGTs or to the transfer of organellar genes to the nucleus [31,32]. The analysis of the HisD phylogenetic tree obtained using all the available eukaryotic sequences with their prokaryotic counterparts revealed that it is consistent with the eukaryotic species phylogeny and supported by very high bootstrap values (Figure 8a), suggesting a vertical inheritance for *hisD* in Fungi and Plants. The phylogenetic analyses also revealed that the second *Aspergillus nidulans hisD* gene has been acquired from a prokaryote, probably a Cyanobacterium. The interpretation of the phylogenetic tree obtained using the HisIE sequences is quite different (Figure 8b): two distinct clusters can be recognized, the first one comprising sequences from Fungi, and the second one including prokaryotic and Plants sequences. HisIE from Fungi now is separated from Plants, and Eukarya are not monophyletic. A possible explanation is that *hisIE* genes of Fungi and Plants have not been vertically inherited from a common ancestor; however HisIE proteins often gave rise to 'strange' phylogenetic trees for their short length.

Concerning the splitting of the two moieties (HisIE and HisD) in *S. pombe*, the presence of the N-terminal unknown sequence in the *hisIE* gene product (Table 3) and the phylogenetic trees reported in Figures 7 suggest a *HIS4* gene fission event rather than its primary absence. Moreover, the fission yeast belongs to the Ascomycota, all of which possess the canonical *S. cerevisiae*-like HIS4 enzyme, and the same (Table 3) is true for species which branch off earlier from the fungal lineage (as the Basidiomycota *Cryptococcus neoformans* and other).

An evolutionary model for the origin and evolution of HIS4 *gene*

A possible evolutionary model for the origin and evolution of *HIS4* predicts that (at least) one copy of *hisD* was donated from prokaryotes to the ancestor of Eukarya and this sequence was vertically inherited by Fungi and Plants. Concerning *hisIE*, it is quite possible that the ancestor of eukaryotes received a bifunctional *hisIE* gene from prokaryotes rather than two monofunctional genes that then underwent a gene fusion. These ideas are consistent with both the structure of *hisIE* genes in known Eukarya and with the phylogenetic trees shown in Figure 8. However, the ancestor of Fungi and Plants might have received the HisIE gene from different prokaryotes, as shown by the HisIE phylogenetic tree (Additional file 8). After the divergence of Fungi from Plants the fusion between the two bifunctional genes (*hisIE* and *hisD*) occurred in Fungi leading to the extant *HIS4* tetrafunctional gene, which was maintained during the evolution with the exception of *S. pombe*, where it was split.

Conclusion

In this paper we have analyzed the fusions involving histidine biosynthetic genes. At least eight out of ten *his* genes, i.e., *hisA, B, D, E, F, H, I,* and *N* underwent different fusion events strongly supporting a major role of this mechanism in both the assembly and evolution of histidine biosynthesis. Each of the five *his* fusions detected so far, i.e. *hisA/hisF, hisIE, hisHF* (*HIS7*), *hisNB*, and *hisIED* (*HIS4*) has been analyzed for: i) gene structure, ii) phylogenetic distribution, iii) timing of appearance, iv) horizontal gene transfer, v) correlation with gene organization, and vi) biological significance. Our results might be summarized as follows:

1. The only "universal" gene fusion concerns *hisA* and *hisF* genes, which are the outcome of a cascade of (at least) two gene elongation events followed by a paralogous gene duplication. The structure of *hisA* and *hisF*, that is the presence of two paralogous modules half the size of the entire gene, is the same in all histidine-synthesizing organisms and no correlation with *his* genes organization exists, in the sense that the two genes maintain the same structure independently from *his* gene organization (complete or partial clustering or scattering). It is also interesting that the traces of the two elongation events are detectable at the primary sequence level in only a few species, mainly Archaea.

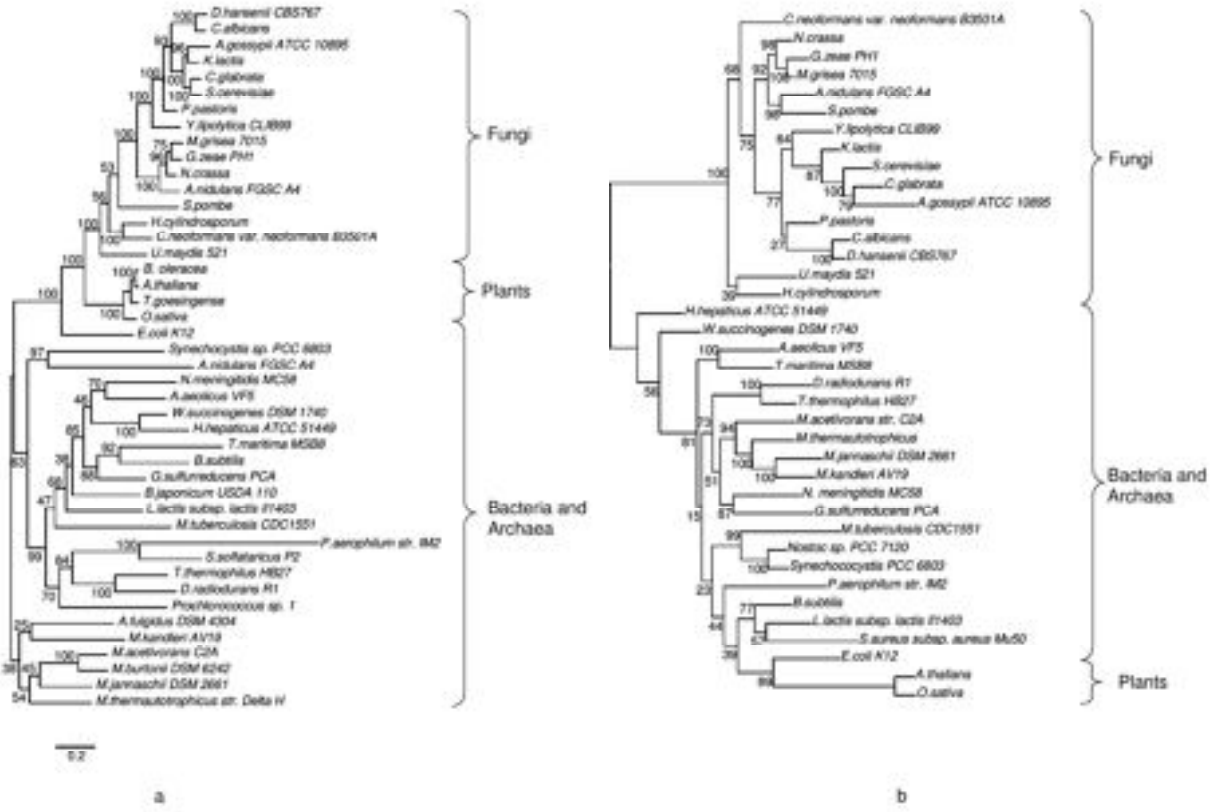


Figure 8

HisD and HisIE phylogenetic analysis. Phylogenetic tree obtained using a multialignment of HisD (a) and HisIE (b) proteins and domains from HIS4 proteins and MrBayes program. Values above nodes are posterior probabilities*100 (for details on phylogenetic tree construction see Methods). See text for details.

This suggests that the two elongation events are very ancient i.e. they date before LUCA. The analysis of the sequence and structure of HisA and HisF depicts a likely scenario for divergent evolution of (at least) some of the proteins belonging to the TIM-barrel family: interestingly HisA is the only one maintaining an almost perfect subdivision in two modules half the size of the entire gene and sharing a high degree of sequence similarity. In other TIM-barrels, such as HisF and TrpF, the common origin of the two halves has been obscured by point mutations and/or larger rearrangements due to functional and/or structural constraints. Therefore it is possible that HisA might resembles the ancestral TIM-barrel enzyme. By structural comparisons of fragments of the *T. maritima* HisA protein we obtained indications on the paralogy between quarters of the barrel (each correspondingto a $(\beta/\alpha)_2$ module).

2. No fusion involving *his* genes has been disclosed in Archaea, with the exception of *hisIE* in some Euryarchaeota. However, the *hisIE* bifunctional genes are very likely not native of those Archaea, but are the outcome of a HGT event involving an entire bacterial *his* operon [15].

3. The fusion between *hisI* and *hisE* occurred more than once in Bacteria, speaking towards a phenomenon of convergent evolution; in many cases it has been preceded by the progressive approaching and overlapping of the genes (e.g. in proteobacteria). Sometimes, the fusion was concomitant with the formation of compact operons. Moreover, this gene might have been horizontally transferred.

4. The *hisNB* fusion is a relatively recent evolutionary event that happened in the ε-branch of proteobacteria. This fusion was parallel to the introgression of *hisN* into an already formed and more or less compact *his* operon. Once occurred, the fusion was fixed and transferred to other proteobacteria and/or CFB group along with the entire operon or part thereof.

5. The fusions involving *hisH* and *hisF*, *hisIE* and *hisD* occurred in the eukaryotic lineage. Whilst the fusion leading to the tetrafunctional gene *HIS4* is peculiar of eukaryotes, a *hisHF* fusion was found also in two bacteria, probably as a result of a HGT from an eukaryote.

The more or less narrow phylogenetic distribution of these fusions raises the question of the structure of *his* genes in the LUCA. On the basis of the available data, we suggest (Figure 9) that LUCA possessed all monofunctional histidine biosynthetic genes.

1
The whole body of data reported above suggests that the fusion(s) of histidine biosynthetic genes has been driven by different selective pressures. In the case of the elongation events leading to the extant *hisA* and *hisF*, a structural/functional significance can be invoked. Indeed, the elongation events were very likely positively selected in order to optimize the structure and the function of the ancestral TIM-barrel.

The fusion of HOL-P phosphatase and IGP dehydratase might have been selected to ensure a fixed ratio of gene products that function in the same biochemical pathway. Concerning the *hisHF* (*HIS7*) fusion, its biological significance is clear; whilst in prokaryotes the two proteins encoded by *hisH* and *hisF* must interact in a 1:1 ratio to give the active form of IGP synthase, in the eukaryotic bifunctional protein, the two entities are fused allowing their immediate interaction and the substrate tunnelling. A similar "substrate channeling" and/or "fixed ratio of gene products" might be invoked for the fusion involving the prokaryotic *hisIE* genes, which code for enzymes performing consecutive steps of histidine biosynthesis.

Independently from their case-by-case biological significance, such associations might be responsible for a more specific commitment of intermediates in a given pathway by means of the spatial co-localization of enzymes. Operons might allow Bacteria to reach the same target: the translation of polycistronic mRNAs favors protein-protein interactions or the spatial segregation of a pathway. Indeed, genes coding for interacting proteins are often organized in operons [33]; in this context, it has been suggested that the bacterial IGP synthase might be part of a complex metabolon whose entities are encoded by the four genes *hisBHAF*, constituting the so-called "core" of histidine biosynthesis [9,12]. Data presented here might suggest that the polypeptides coded for by *hisI*, *hisE*, and *hisD* are part of another metabolon.

This scenario can clarify the biological significance and the evolutionary advantages of the fusion leading to the HIS7 and HIS4 proteins and their prokaryotic counterparts. Indeed, the cytoplasm of a prokaryotic or eukaryotic cell represents an extremely complex and crowded environment, where al lot of macromolecular structures might represent an important barrier to the free diffusion of (even small) polypeptides; it is plausible that the stochastic movement of proteins that have to interact in the bacterial cytoplasm is a rate-limiting step for a pathway. It has been observed that the diffusion coefficient of many molecules in prokaryotic and eukaryotic cells is less than in water [34]. Accordingly, the intracellular concentration of proteins in *E. coli* cells has been measured to range between 300 and 400 mg/ml ([35] and references therein) revealing the bacterial cell interior to be a very crowded environment. The greater the volume and complexity of the cell the greater is the obstacle to the free diffusion of intermediates or signal molecules inside the cell. The problems related to the crowding of the intracellular milieu have been proposed to be greater for eukaryotic cells, not only for the distances an intermediate have to override to reach a given catalytic site, but especially for the presence of a number of physical obstacles, as the cytoskeleton, multi-enzymatic complexes and organelles. One of the major effects of the crowding is the reduced mobility of molecules, an effect directly related to the properties and the translational ray of a molecule. Moreover many of the intermediates of metabolic pathways can be sequestered by aspecific binding to intracellular structures or be consumed by unwanted catalytic activities, reducing the overall rate of production of the end-product and augmenting its average energetic cost. If a similar view is correct, and the diffusion problem is rate limiting for at least some of the metabolic pathways performed by the cell, then the substrate channeling can permit to bypass the problem and the loss of intermediates by collateral pathways, which might result in a waste of energy. In eukaryotic cells, where the operons are absent and whose inside is more complex than that of prokaryotes, this obstacle might be bypassed by the fusion of functional domains that permits an immediately active product after the translation of a single mRNA.

Methods

Sequence retrieval

Nucleotide and amino acid sequences were obtained from the NCBI complete genomes database. The BLASTp option of the BLAST program [24] was used to retrieve His proteins used in this work.

Alignment and phylogenetic analyses

The ClustalW program [25] with standard parameters was used for pairwise and multiple amino acid sequences alignments, followed by careful visual inspection. Phylogenetic trees were obtained by using the software Mega3, the Neighbor-Joining method [39] with 1000 bootstrap replicates and the JTT [38] evolutionary model. We

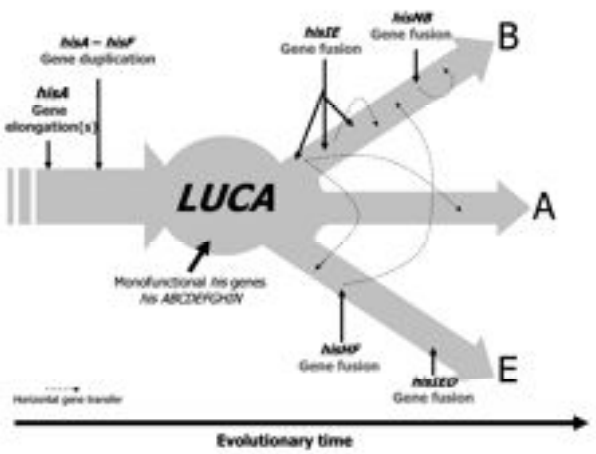


Figure 9
A global view of *his* gene fusions appearance. Schematical representation of *his* gene fusion appearance and horizontal transfer. Abbreviations used: A, B, E correspond to Archaea, Bacteria and Eukarya, respectively. LUCA stands for the Last Universal Common Ancestor.

obtained different topologies for the tree corresponding to HIS4 domains and we compared them with those obtained with bayesian inference, i.e. MrBayes [36]. We defined the following parameters (not cited if default settings were used): evolutionary models of amino acid sequences were the WAG [37] and JTT [38] with character frequencies estimated from dataset (-F); topologies obtained were identical for the two model; we report the shortest trees (in both cases corresponding to the one obtained with WAG); we used heterogeneous rates among sites, distributed as a Gamma distribution with shape parameter free of variate from 0 to 50 (average obtained for these datasets 0.85); MCMC settings were as follows: 2,500,000 and 1,500,000 generations, respectively for HisIE and HisD domains, and five chains. No starting trees were used, with the idea in mind that convergence to very similar values of the five chains would have been more significant than starting each chain from the same tree. Trees were sampled every 250 generations, for a total set of 10,000 and 6,000 trees. Burnin was used to exclude from following analysis those trees which were sampled before convergence of the chains; this was assessed case-by-case by calculating average, standard deviation and variance. Convergence was reached after about 50,000 generations (200 trees discarded). The resulting datasets were used to obtain the shown trees with the allcompat option.

Structural alignments

Structural alignments and Root-means-squares calculations were performed using Swiss pdb viewer [40]. We isolated modules corresponding to $(\beta/\alpha)_2$ and (β/α) accordingly to the analysis performed by [19] showing secondary elements belonging and not belonging to the barrel structure. These modules were then used as independent molecules and structurally aligned by using the 'magic fit' option.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally to the work and manuscript preparation.

Additional material

Additional file 1

Phylogenetic distribution of* hisNB *genes. *Histogram showing the percentage of organisms possessing a* hisNB *gene for taxonomic groups represented in NCBI genomes database and taking into account only histidine producing organisms.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S1.pdf]

Additional file 2

***HisD phylogenetic tree of organisms possessing* hisNB**. *A NJ phylogenetic tree (evolutionary model: Dayhoff, 500 bootstrap replicates) obtained from a HisD multialignment. The topology is congruent with those obtained with other His proteins; it illustrates that* hisNB *has been probably transferred together with a complete histidine biosynthetic operon. See also Additional File 3 concerning gene organization. Red: ε-proteobacteria possessing (upper group) and not possessing (bottom group) the* hisNB *gene fusion; Green: CFB group bacteria possessing (upper group) and not possessing (bottom group) the* hisNB *gene fusion.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S2.pdf]

Additional file 3

HisNB and gene organization. *Several features concerning HisNB proteins from Bacteria belonging to species not available at the time of our previous analysis concerning* hisNB *genes [11].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S3.pdf]

Additional file 4

Phylogenetic distribution of* hisIE *genes. *Histogram showing the percentage of organisms possessing a* hisIE *gene for taxonomic groups represented in NCBI genomes database and taking into account only histidine producing organisms.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S4.pdf]

Additional file 5

***HIS4* multialignments with concatenated prokaryotic sequences.** *A multialignment of fungal HIS4 sequences and the corresponding proteins from Plants and from a selected number of Prokaryotes.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S5.pdf]

Additional file 6

Entropy plot of a multialignment of HIS4 sequences. *Entropy plot of the multialignment of HIS4 proteins. The regions of the protein are also indicated showing their different degree of conservation. Entropy was calculated with the following formula: H(l) = S [f(b, l) * ln(f(b, l))], where b is a residue found in in column l and f(b, l) its frequency in l and the summation extends over all residues in column l.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S6.pdf]

Additional file 7

Pairwise identity values within HIS4 domains. *Identity values for the pairwise comparison of the different domains composing HIS4 proteins (the standard deviation is also shown).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S7.pdf]

Additional file 8

Identifiers of sequences used in this work.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-S2-S4-S8.txt]

Acknowledgements

This article has been published as part of *BMC Evolutionary Biology* Volume 7 Supplement 2, 2007: Second Congress of Italian Evolutionary Biologists (First Congress of the Italian Society for Evolutionary Biology). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2148/7?issue=S2

References

1. Winkler ME: **Biosynthesis of histidine.** In *Escherichia coli and Salmonella typhimurium: cellular and molecular biology Volume 1*. Edited by: *Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Humberger HD. Washington DC: ASM Press*; 1987:395-411.
2. Alifano P, Fani R, Lió P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB: **Histidine biosynthetic pathway and genes: structure, regulation and evolution.** *Microbiol Rev* 1996, **60:**44-69.
3. Carlomagno MS, Chiarotti L, Alifano P, Nappo AG, Bruni CB: **Structure of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons.** *J Mol Biol* 1988, **203:**585-606.
4. Shen C, Yang L, Miller SL, Oró J: **Prebiotic synthesis of histidine.** *J Mol Evol* 1990, **31:**167-74.
5. Maurel MC, Ninio J: **Catalysis by a prebiotic nucleotide analog of histidine.** *Biochimie* 1987, **69:**551-553.
6. Shen C, Mills T, Oró J: **Prebiotic synthesis of histidyl-histidine.** *J Mol Evol* 1990, **31:**175-9.
7. White DH, Erickson JC: **Catalysis of peptide bond formation by histidyl-histidine in a fluctuating clay environment.** *J Mol Evol* 1980, **16:**279-290.
8. Shen C, Lazcano A, Oró J: **The enhancement activities of histidyl-histidine in some prebiotic reactions.** *J Mol Evol* 1990, **31:**445-52.
9. Fani R, Lió P, Lazcano A: **Molecular evolution of the histidine biosynthetic pathway.** *J Mol Evol* 1995, **41:**760-774.
10. Fani R, Mori E, Tamburini E, Lazcano A: **Evolution of the structure and chromosomal distribution of histidine biosynthetic genes.** *Orig Life Evol Biosph* 1998, **28:**555-570.
11. Brilli M, Fani R: **Molecular evolution of *hisB* genes.** *J Mol Evol* 2004, **58:**225-237.
12. Brilli M, Fani R: **Origin and evolution of eucaryal *HIS7* genes: from metabolons to bifunctional proteins?** *Gene* 2004, **339:**149-160.
13. Fani R, Brilli M, Lió P: **The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon.** *J Mol Evol* 2005, **60:**378-390.
14. Fani R: **Gene duplication and gene loading.** In *Microbial evolution: gene establishment, survival and exchange* Edited by: *Miller RV, Day MJ. Washington DC: ASM Press*; 2004:67-81.
15. Fani R, Brilli M, Lió P: **Inference from proteobacterial operons shows piecewise organization: a reply to price et Al.** *J Mol Evol* 2006, **63:**577-580.
16. Jensen R: **Evolution of metabolic pathways in enteric bacteria In *Escherichia coli* and *Salmonella typhimurium*.** In *Escherichia coli and Salmonella typhimurium: cellular and molecular biology Volume 1*. Edited by: *Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Humberger HD. Washington DC: ASM Press*; 1987:2649-2662.
17. Yanai I, Wolf YI, Koonin EV: **Evolution of gene fusions: horizontal transfer versus independent events.** *Genome Biol* 2002, **3:**research0024-.
18. Fani R, Lió P, Chiarelli I, Bazzicalupo M: **The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes.** *J Mol Evol* 1994, **38:**489-495.
19. Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M: **Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion.** *Science* 2000, **289:**1546-1550.
20. Kneidinger B, Marolda C, Graninger M, Zamyatina A, McArthur F, Kosma P, Valvano MA, Messner P: **Biosynthesis pathway of ADP-L-glycero-beta-D-manno-heptose in *Escherichia coli*.** *J Bacteriol* 2002, **184:**363-9.
21. Kneidinger B, Graninger M, Puchberger M, Kosma P, Messner P: **Biosynthesis of nucleotide-activated D-glycero-D-manno-heptose.** *J Biol Chem* 2001, **276:**20935-44.
22. Klem TJ, Davisson VJ: **Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis.** *Biochemistry* 1993, **32:**5177-5186.
23. Kuenzler M, Balmelli T, Egli CM, Paravicini G, Braus GH: **Cloning, primary structure, and regulation of the *HIS7* gene encoding a bifunctional glutamine amidotransferase: cyclase from *Saccharomyces cerevisiae*.** *J Bacteriol* 1993, **175:**5548-5558.
24. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25:**3389-3402.
25. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
26. Bürger E, Görisch H: **Evidence for an essential lysine at the active site of L-Histidinol:NAD + oxidoreductase; a bifunctional dehydrogenase.** *Eur J Biochem* 1981, **118:**125-130.
27. Kheirolomoom A, Mano J, Nagai A, Ogawa A, Iwasaki G, Ohta D: **Steady-state kinetics of cabbage Histidinol dehydrogenase.** *Arch Biochem Biophys* 1994, **312:**493-450.
28. Fujimori K, Ohta D: **Isolation and characterization of a histidine biosynthetic gene in *Arabidopsis* encoding a polypeptide with two separate domains for phosphoribosyl-ATP pyrophosphohydrolase and phosphoribosyl-AMP cyclohydrolase.** *Plant Physiology* 1998, **118:**275-283.
29. Keesey JK Jr, Bigelis R, Fink GR: **The product of the *his4* gene cluster in *Saccharomyces cerevisiae*. A trifunctional polypeptide.** *J Biol Chem* 1979, **254:**7427-7433.
30. Van de Peer Y, De Wachter R: **Evolutionary relationships among the eucaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA.** *J Mol Evol* 1997, **45:**619-630.
31. Brandvain Y, Barker MS, Wade MJ: **Gene co-inheritance and gene transfer.** *Science* 2007, **315:**1685.
32. Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D: **Cyanobacterial contribution to algal nuclear genomes is pri-**

marily limited to plastid functions. *Curr Biol* 2006, **16:**2320-2325.
33. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
34. Dauty E, Verkman AS: **Molecular crowding reduces to a similar extent the diffusion of small solutes and macromolecules: measurement by fluorescence correlation spectroscopy.** *J Mol Recognit* 2004, **17:**441-447.
35. Zimmerman SB, Minton AP: **Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of *Escherichia coli*.** *J Mol Biol* 1991, **222:**599-620.
36. Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
37. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Mol Biol Evol* 2001, **18:**691-699.
38. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition.** *Nature* 1992, **358:**86-89.
39. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.
40. Kaplan W, Littlejohn TG: *Swiss-PDB Viewer (Deep View). Brief Bioinform* 2001, **2:**195-197.

3.2 The evolution of histidine biosynthesis in Archaea: insights into the *his* genes structure and organization in LUCA

The available sequences of genes encoding the enzymes associated with histidine biosynthesis suggest that this is an ancient metabolic pathway that was assembled prior to the diversification of Bacteria, Archaea, and Eucaryabefore that is before (or in concomitance with) the appearance of LUCA. Paralogous duplication, gene elongation, and fusion events of several different his genes have played a major role in shaping this biosynthetic route. However, it is still not clear how these genes were organized in the genome of the LUCA community, which was their structure and how many functions they performed. This is mainly due to the fact that the analysis of the structure and organization of *his* genes has been focused on bacterial genomes (especially proteobacterial ones). Very little is known about this issue in Archaea. Therefore, in this work, we have analyzed the structure and organization of histidine biosynthetic genes (*his*) from 55 complete archaeal genomes and combined it with phylogenetic inference in order to investigate the mechanisms responsible for the assembly of the *his* pathway and the origin of his operons. We show that a wide variety of different organizations of *his* genes exists in Archaea and that some *his* genes or entire *his* (sub-)operons have been likely transferred horizontally between Archaea and Bacteria. However, we show that, in most Archaea, *his* genes are monofunctional (except for *hisD*) and scattered throughout the genome, suggesting that *his* operons might have been assembled multiple times during evolution and that in some cases they are the result of recent evolutionary events. An evolutionary model for the structure and organization of *his* genes in LUCA is proposed. Lastly, our analysis also reinforce the idea that *his* biosynthesis is an ancient metabolic pathway that was assembled prior to the diversification of Bacteria, Archaea, and Eucarya.

J Mol Evol
DOI 10.1007/s00239-009-9286-6

The Evolution of Histidine Biosynthesis in Archaea: Insights into the *his* Genes Structure and Organization in LUCA

Marco Fondi · Giovanni Emiliani · Pietro Liò · Simonetta Gribaldo · Renato Fani

Received: 22 July 2009 / Accepted: 18 September 2009

Abstract The available sequences of genes encoding the enzymes associated with histidine biosynthesis suggest that this is an ancient metabolic pathway that was assembled prior to the diversification of Bacteria, Archaea, and Eucarya. Paralogous duplication, gene elongation, and fusion events of several different *his* genes have played a major role in shaping this biosynthetic route. We have analyzed the structure and organization of histidine biosynthetic genes from 55 complete archaeal genomes and combined it with phylogenetic inference in order to investigate the mechanisms responsible for the assembly of the *his* pathway and the origin of *his* operons. We show that a wide variety of different organizations of *his* genes exists in Archaea and that some *his* genes or entire *his* (sub-)operons have been likely transferred horizontally between Archaea and Bacteria. However, we show that, in most Archaea, *his* genes are monofunctional (except for *hisD*) and scattered throughout the genome, suggesting that *his* operons might have been assembled multiple times during evolution and that in some cases they are the result of recent evolutionary events. An evolutionary model for the structure and organization of *his* genes in LUCA is proposed.

Keywords Operon evolution · Histidine biosynthesis · Paralogous genes · Gene fusion · Evolution of metabolic pathways · Operon origin · Metabolic pathway origin

Abbreviations

HisG	ATP phosphoribosyl transferase (EC 2.4.2.17)
HisD	Histidinol dehydrogenase (EC 1.1.1.23)
HisC	Histidinol-phosphate aminotransferase (EC 2.6.1.9)
HisN	Histidinol-phosphate phosphatase (EC 3.1.3.15)
HisB	Imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19)
HisH	G-type glutamine amidotransferase
HisA	[*N*-(5_-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide isomerase (EC 5.3.1.16)
HisF	Imidazole glycerol phosphate synthase subunit HisF (EC 4.1.3.-)
HisI	Phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19)
HisE	Phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31)
LUCA	Last universal common ancestor
HGT	Horizontal gene transfer
PRFAR	*N*-(5′phospho-D-1′-ribulosylformimino)-5-amino-1-(5-phosphoribosyl)-4-imidazolecarboxamide
AICAR	5-Aminoimidazole-4-carboxamide ribonucleoside

M. Fondi · R. Fani (✉)
Department of Evolutionary Biology, University of Florence, Via Romana 17-19, 50125 Florence, Italy
e-mail: renato.fani@unifi.it

M. Fondi
e-mail: marco.fondi@unifi.it

G. Emiliani
Tree and Timber Institute, National Research Council, Via Biasi, 75, 38010 San Michele all'Adige, Trento, Italy

P. Liò
Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

S. Gribaldo
Institut Pasteur, Department of Microbiology, BMGE Unit, Paris, France

Published online: 03 November 2009

Introduction

It is widely accepted that ancestral life forms inhabited an environment (the so-called primordial soup) rich in organic compounds spontaneously formed in the prebiotic world (Lazcano et al. 1992). This hypothesis is known as the "Oparin–Haldane theory" and predicts that early organisms were heterotrophic and had to perform only a minimum of biosynthesis. If this is so, the increasing number of primordial cells might have led to the exhaustion of the prebiotic supply of amino acids and other compounds that were present in the primordial soup. This, in turn, would have imposed a progressively stronger selective pressure favoring those primordial heterotrophic cells that became capable of synthesizing those molecules whose concentration was becoming limiting in the primordial soup. Hence, the emergence of basic biosynthetic pathways was one of the major events during the early evolution of life, because their appearance allowed ancient organisms to become increasingly less dependent on exogenous sources of compounds present in the primitive environment as a result of prebiotic syntheses. Different molecular mechanisms may have been responsible for shaping the early metabolic pathways, including gene elongation, duplication and/or fusion, modular assembly of new proteins, cell fusion (synology) and HGT (xenology). How the major biosynthetic pathways actually originated is still an open question, but several different hypotheses have been formulated to explain the establishment of anabolic routes. These hypotheses include the patchwork theory, according to which metabolic routes are the result of the serial recruitment of relatively small, inefficient enzymes endowed with broad-specificity that could react with a wide range of chemically related substrates (Jensen 1976; Ycas 1974).

Clues on the evolutionary history of metabolic pathways can be derived from both experimental approaches, based on the so-called "directed evolution experiments" (Fani and Fondi 2009; Fondi et al. 2009) and by the comparative analysis of the structure and organization of genes in microorganisms belonging to the three domains of life (Archaea, Bacteria, Eucarya). Histidine biosynthesis represents an excellent model for the analysis of the molecular mechanisms and the forces that have driven the origin and evolution of metabolic pathways. Indeed, it is one of the best characterized anabolic pathways and a large body of genetic and biochemical information is available, including gene structure, organization and expression. For over 40 years this pathway has been the subject of extensive studies, mainly in the enterobacterium *Escherichia coli* and its close relative *Salmonella typhimurium*, for both of which details of histidine biosynthesis appear to be identical (Winkler 1987). As shown in Fig. 1, in these two enterobacteria the pathway is unbranched, and includes a number of complex and unusual biochemical reactions. It consists of nine intermediates, all of which have been described, produced by eight distinct enzymes (Alifano et al. 1996).

There are several independent evidences for the antiquity of the histidine biosynthetic pathway. It is generally accepted that histidine is present in the active sites of enzymes because of the special properties of the imidazole group (Weber and Miller 1981). The apparently universal

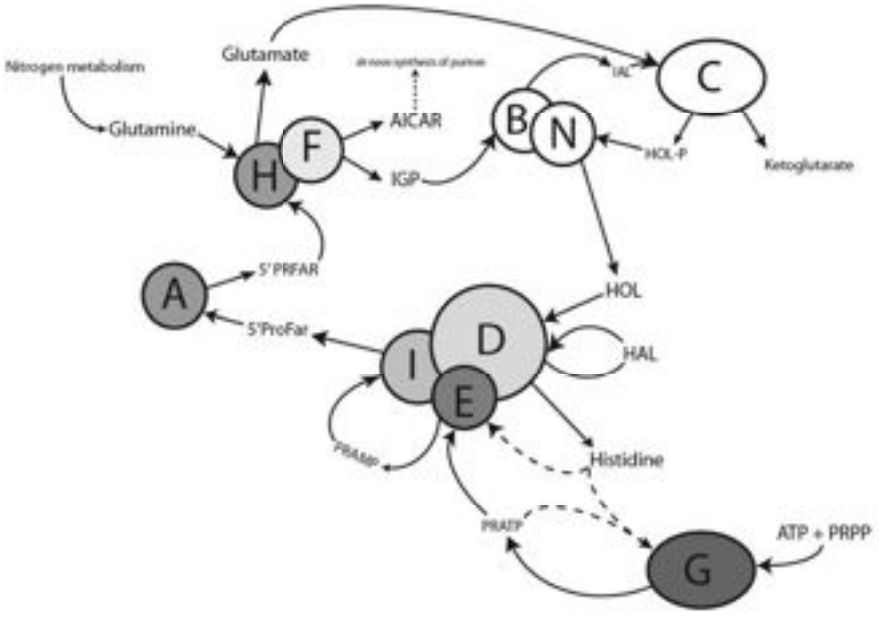


Fig. 1 Schematic representation of the histidine biosynthetic pathway

phylogenetic distribution of the *his* genes (Fani et al. 2007) suggests that histidine synthesis was already part of the metabolic abilities of the last common ancestor of the three extant cell domains (Lazcano et al. 1992). The chemical synthesis of histidine (Shen et al. 1990c), of prebiotic analogues of histidine (Maurel and Ninio 1987), and of histidyl-histidine under primitive conditions has been reported (Shen et al. 1990b), as well as the role of the latter in the enhancement of some possible prebiotic oligomerization reactions involving amino acids (White and Erickson 1980) and nucleotides (Shen et al. 1990a). Since its biosynthesis requires a carbon and a nitrogen equivalent from the purine ring of ATP, it has also been suggested that histidine may be the molecular vestige of a catalytic ribonucleotide from an earlier biochemical stage in which RNA played a major role in catalysis (White 1976). If primitive catalysts required histidine, then the eventual exhaustion of the prebiotic supply of histidine and histidine-containing peptides (Shen et al. 1990a, b, c) must have imposed an important pressure favoring those organisms capable of synthesizing histidine.

Histidine biosynthesis plays also an important role in cellular metabolism, since four of the *his* genes (*hisBHAF*), forming the so-called "core" of the pathway (Fig. 1), represent a metabolic cross-point interconnecting histidine biosynthesis to both nitrogen metabolism and de novo synthesis of purines. The connection with purine biosynthesis results from an enzymatic step catalyzed by imidazole glycerol phosphate synthase, an enzyme which has been shown to be a dimeric protein composed of one subunit each of the *hisH* and *hisF* genes product (Klemm and Davisson 1993). This heterodimeric enzyme catalyzes the transformation of PRFAR into AICAR, which is then recycled into the de novo purine biosynthetic pathway, and imidazole glycerol phosphate (IGP), which in turn is then transformed into histidine (Fig. 1). Histidine biosynthesis is connected to nitrogen metabolism by a glutamine molecule, which is believed to be the source of the final nitrogen atom of the imidazole ring of IGP. The important role played by histidine biosynthesis in cellular metabolism is in fact underscored by the considerable energy (41 ATP molecules) that is required for the synthesis of each histidine molecule (Brenner and Ames 1971).

The analysis of several completely sequenced genomes have disclosed many examples of elongation, duplication and/or fusion events involving different *his* genes. Interestingly, in some species, more than one enzymatic function is encoded by the same bi- or multifunctional cistron, such as *hisD, hisNB, hisHF*, and *hisIE* in some prokaryotes (Carlomagno et al. 1988), *HIS4* and *HIS7* in eukaryotes (Donahue et al. 1982; Kuenzler et al. 1993). These multifunctional genes very likely are the outcome of fusion events (Fani et al. 2007).

It has also been demonstrated that gene duplication also played a key role in shaping histidine biosynthesis. Indeed, *hisA* and *hisF* are the outcome of a cascade of gene elongation (i.e., an in-tandem gene duplication followed by the fusion of the two copies) and duplication events (Fani 2004), and *hisH* was very likely recruited from other metabolic pathways (Fani et al. 1998).

Noteworthy, after the assembly of the entire pathway, the structure and/or organization of *his* genes underwent major rearrangements in the three domains, generating a wide variety of structural and/or clustering strategies of *his* genes (Fani et al. 1998; Fani et al. 2005). Thus, the analysis of the structure and organization of *his* genes could help investigating the general problem of the origin and evolution of operons (Fani et al. 2005; Price et al. 2006).

The whole body of data available led to the assumption that the entire biosynthetic pathway was assembled long before the appearance of LUCA. However, it is still not clear how these genes were organized in the genome of the LUCA community, which was their structure and how many functions they performed. This is mainly due to the fact that the analysis of the structure and organization of *his* genes has been focused on bacterial genomes, especially proteobacteria (Fani et al. 2007). Very little is known about this issue in Archaea. Therefore, the aim of this work was to use all archaeal genome information (a dataset of 55 completely sequenced genomes) to carry out a comparative analysis of the structure and organization of *his* genes in Archaea.

Materials and Methods

Sequence Retrieval

On 30 June 2009, a total of 55 complete genomes belonging to the archaeal lineage were available in the GenBank database. Histidine biosynthetic genes were retrieved from this dataset according to the blast bidirectional best hit (BBH) criterion, using *Escherichia coli, Anabaena variabilis*, and *Lactococcus lactis* sequences as seeds. Moreover, when possible, gene organization (i.e., inside or outside histidine operons) of the putative orthologs was checked for a further validation of orthology relationships.

Sequence Alignment and Phylogenetic Tree Construction

Multiple sequences alignments were conducted using Muscle 3.6 software (Edgar 2004). Maximum Likelihood analysis was carried out using Phyml (Guindon and Gascuel 2003), with a WAG model of amino acid substitution, including a gamma function with 6 categories to take into account differences in evolutionary rates at sites. Statistical

support at nodes was obtained by non-parametric bootstrapping on 1000 resampled datasets by using Phyml.

Results and Discussion

The Histidine Genes and Their Organization in Archaea

The 55 archaeal completely sequenced genomes were probed by using the *E. coli* HisGDCNBHAFIE sequences, the *L. lactis* HisN and the *A. variabilis* HisZ as queries. In this way a (almost) complete set of histidine biosynthetic genes was retrieved from 45 genomes (Table 1). Indeed, ten strains are auxotrophic for histidine. All the retrieved sequences were analyzed for both structure and organization to investigate the mechanisms responsible for their assembly into cluster and/or operons and the extent of HGT of *his* genes between organisms of the same or different phylogenetic lineages. Figure 2 shows the structure and organization of *his* biosynthetic genes correlated both to the archaeal phylogeny and to thermophilic/hyperthermophilic lifestyles. Three different *his* genes arrays exist among the archaeal lineages:

(1) In Euryarchaeota:
   (a) All *his* genes are scattered throughout the genome in Archaeoglobales, Halobacteriales, Methanobacteriales, Methanococcales, and Methanopyrales.
   (b) A partial cluster is found in Methanomicrobiales and Methanosarcinales. In these Archaea, three genes are clustered in a putative sub-operon in the order *hisGBA*, whereas the others are scattered in the genome.
   (c) In Thermococcales and Thermoplasmatales the *his* genes are arranged in a compact operon exhibiting the same relative order, *hisGDBHAFIE*, resembling the *E. coli*-type operon. However, the two genomes also harbor *hisC* and a *hisZ* gene, even though in a different position.

(2) In Crenarchaeota:
   (a) Most of *his* genes are scattered in the *Ignicoccus hospitalis* (Desulforococcales) genome.
   (b) In the Thermoproteales the *his* genes are arranged in different (putative) mini-opeons (*hisCG*, *hisBA*, *hisFDE*).
   (c) In all the Sulfolobales the nine genes are arranged in a compact operon in the order *hisCGABFDEHI*.

(3) In Thaumarchaeota:
   (a) The two organisms belonging to this domain possess the same *his* genes organization. All the *his* genes are embedded in a compact operon (*hisGDCXBHAI*), whose gene order is very similar to the enterobacterial *his* operon. Interestingly, the ORF located between *hisC* and *hisB* (ORFX) occupies the same position of *hisN* in the *E. coli* operon. However, the aminoacid sequence of the encoded protein by did not retrieve any DDDD-type or PHP-like phosphatase and thus does not correspond to the "canonical" *hisN* gene (see below). Besides, no homologs of the *hisE* gene were retrieved from these two genomes.

The Structure of *his* Gene in Archaea

Previous works have revealed that most of *his* genes, i.e., *hisN*, *hisB*, *hisA*, *hisF*, *hisI*, *hisE*, and *hisG* are of particular importance from both an evolutionary and genome organization viewpoint (Brilli and Fani 2004; Fani et al. 1994, 1995), because they underwent different molecular rearrangements in diverse phylogenetic lineages. Thus, a detailed inspection of all these gene products was carried out.

The Archaeal hisN and hisNB Genes

The sixth and eighth steps of histidine biosynthesis are catalyzed by histidinol-phosphate phosphatase (EC 3.1.3.15) (HOL-Pase) and IGP dehydratase (EC 4.2.1.19), respectively. It has been reported that different phosphatases perform the dephosphorylation of HOL-P, whereas the dehydratation of IGP is carried out by the same enzyme (coded for by *hisB*) in all known histidine-synthesizing organisms. In the model organism *E. coli*, the two activities are encoded by a bifunctional gene, referred to as *hisNB* (Brilli and Fani 2004). The accepted model for their association predicts the existence of two independent domains in the gene, i.e., a proximal domain (*hisN*) encoding the phosphatase moiety of the DDDD-type (Brilli and Fani 2004), and a distal one (*hisB*) encoding the dehydratase activity. A model has been proposed to explain the evolution of the bifunctional *hisNB* gene (Brilli and Fani 2004). It posits that it is the outcome of a gene fusion event involving two cistrons (*hisN* and *hisB*) coding for HOL-Pase and IGP-ase activities, respectively (Brilli and Fani 2004). It also predicts that *hisN* likely originated by duplication of a pre-existing gene encoding a DDDD-type phosphatase with a broad range of specificity. The duplication gave rise to two copies, one of which became *hisN* and the other one evolved toward *gmhB* (which is involved in the biosynthesis of a precursor of the inner core of the outer membrane lipopolysaccharides) (Brilli and Fani 2004). According to this model, *hisN* would have joined an already formed *his* operon, and its

Table 1 Phylogenetic distribution of *his* biosynthetic genes (GDCHAFBEINZ) in the archaeal domain. Grey and empty squares indicate the presence or the absence of the corresponding *his* gene, respectively

Organism	G	D	C	H	A	F	B	E	I	N	Z
Aeropyrum pernix											
Archaeoglobus fulgidus											
Caldivirga maquilingensis IC-167											
Candidatus Caenarchaeum symbiosum A											
Candidatus Methanoregula boonei 6A8											
Desulfurococcus kamchatkensis 1221n											
Haloarcula marismortui ATCC 43049											
Halobacterium salinarum R1											
Haloquadratum walsbyi											
Halorubrum lacusprofundi ATCC 49239											
Hyperthermus butylicus											
Ignicoccus hospitalis KIN4 I											
Metallosphaera sedula DSM 5348											
Methanobrevibacter smithii ATCC 35061											
Methanococcus jannaschii											
Methanococcoides burtonii DSM 6242											
Methanococcus aeolicus Nankai-3											
Methanococcus maripaludis S2											
Methanococcus vannielii SB											
Methanococcus voltae											
Methanocorpusculum labreanum Z											
Methanoculleus marisnigri JR1											
Methanopyrus kandleri											
Methanosaeta thermophila PT											
Methanosarcina acetivorans											
Methanosarcina barkeri fusaro											
Methanosarcina mazei											
Methanosphaera stadtmanae											
Candidatus Methanosphaerula palustris E1 9c											
Methanospirillum hungatei JF-1											
Methanobacterium thermoautotrophicum											
Nanoarchaeum equitans											
Natrialba magadii ATCC 43099											
Natronomonas pharaonis											
Nitrosopumilus maritimus SCM1											
Picrophilus torridus DSM 9790											
Pyrobaculum aerophilum											
Pyrobaculum arsenaticum DSM 13514											
Pyrobaculum calidifontis JCM 11548											
Pyrobaculum islandicum DSM 4184											
Pyrococcus abyssi											
Pyrococcus furiosus											
Pyrococcus horikoshii											
Staphylothermus marinus F1											
Sulfolobus acidocaldarius DSM 639											
Sulfolobus islandicus											
Sulfolobus solfataricus											
Sulfolobus tokodaii											
Thermococcus gammatolerans EJ3											
Thermococcus kodakaraensis KOD1											
Thermococcus onnurineus NA1											
Thermofilum pendens Hrk 5											
Thermoplasma acidophilum											
Thermoplasma volcanium											
Thermoproteus neutrophilus V24Sta											

Black squares represent gene fusions

introgression was co-incident to its fusion to *hisB* to produce a bifunctional enzyme. The *hisNB* gene has a narrow phylogenetic distribution and the fusion event was traced back to have occurred in the ancestor of a γ-proteobacterial group. Once arisen, the *hisNB* gene would have been horizontally transferred, as part of the *his* operon, to other Bacteria (Brilli and Fani 2004).

We have analyzed the structure of the *hisB* and *hisN* genes in all the archaeal genomes available. A total of 45 monofunctional HisB sequences were retrieved, one for each of the histidine-synthesizing Archaea. We did not retrieve at a significant degree of sequence similarity neither bifunctional HisNB nor proteins homologous to the *E. coli* HisN or *L. lactis* HisN-like.

These data suggest that the eighth step of histidine biosynthesis is catalyzed by a monofunctional IGP-dehydratase in the Archaea. Concerning the sixth step, the archaeal scenario is completely different from the bacterial one. In most Bacteria the HOL-P dephosphorylation is carried out by a phosphatase either of the DDDD- or PHP-type (Brilli and Fani 2004), whose representatives do not share a significant degree of sequence similarity. To our knowledge, no biochemical, genetic, or functional data concerning the archaeal HOL-Pase is available; furthermore, the analyses carried out in this work indicate that in these microorganisms other phosphatases very likely might perform HOL-P dephosphorylation. These phosphatases might have a broad substrate specificity or might have been recruited through duplication from other metabolic pathways. In this context, particularly interesting is the finding that in Thaumarchaeota, where the genes are organized in a putative operon whose gene order is very similar to the enterobacterial one, the gene located between *hisC* and *hisB* (*hisX*) codes for a putative phosphatase belonging to another group of phosphatases which do not exhibit a significant degree of sequence similarity neither with DDD-type nor PHP-type phosphatases. It is likely that the enzyme coded for by *hisX* catalyzes the HOL-P dephosphorylation in Thaumarchaeota; this gene has not apparently been recruited from other metabolic routes; indeed it has no paralog within the same genome (data not shown). Moreover, probing the 45 archaeal genomes with the Thaumarchaeota aminoacid sequence of the protein encoded by *hisX* retrieved a homologous sequence from few of them (data not shown), suggesting that at least in those cases the HOL-P dephosphorylation might be carried out by a protein homologous to HisX. It is not clear from the available data if *hisX* introgressed an already formed *his* operon, as in the enterobacterial *his* operon. Finally, it must at least be mentioned the hypothesis that the *hisX* sequence retrieved in Archaea might represent a bacterial DDD- or PHP-type phosphatase that has diverged beyond recognition.

The Structure of hisA and hisF: A Cascade of Gene Elongations and Duplications

The *hisA* and *hisF* genes code for a [*N*-(5′-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR)-isomerase and an Imidazole-glycerol-Phosphate (IGP) synthase, respectively, which catalyze two central and sequential reactions (the fourth and fifth ones) in histidine biosynthesis (Fig. 1). The comparative analysis of the HisA and HisF proteins revealed that they are paralogous and share a similar internal organization into two paralogous modules half the size of the entire sequence (Fani et al. 1994). The sequence and predicted secondary structure comparison of these modules led to the suggestion that *hisA* and *hisF* are the result of two ancient successive duplications, the first one involving an ancestral module half the size of the present-day *hisA* gene and leading (by a gene elongation event) to an ancestral gene very likely performing two sequential steps in histidine biosynthesis and resembling the structure of the extant *hisA* gene. This, in turn, underwent a duplication that gave rise to the ancestor of *hisF*. The biological significance of this construction relies in structure of the proteins coded for by *hisA* and *hisF*; indeed, both of them are TIM-barrel proteins. The barrel structure is composed by eight concatenated (β-strand)-loop-(α-helix) units. The β-strands are located in the interior of the protein, forming the staves of a barrel, whereas the α-helices pack around them facing the exterior. According to the model proposed the ancestral half-barrel gave a functional enzyme by homo-dimerization (see Fani and Fondi 2009 and references therein). The elongation event leading to the ancestor of *hisA*/*hisF* genes resulted in the covalent fusion of two half-barrels producing a protein whose function was refined and optimized by mutational changes; once assembled, the "whole-barrel gene" underwent gene duplication, leading to the ancestor of *hisA* and *hisF*. The analysis of all the 45 archaeal gene pair *hisA*–*hisF* revealed that they encode proteins sharing a high degree of sequence similarity and the same internal organization, i.e., they can be easily split into two paralogous modules half the size of the entire sequence (data not shown). Hence, since the overall structure of the *hisA* and *hisF* genes is the same in all the (micro)organisms where they have been identified, it is likely that they were part of the genome of the last common ancestor and that the two successive duplication events leading to the extant *hisA* and *hisF* took place in the early stages of molecular evolution (Fani et al. 2007).

The Structure of hisI and hisE

It is known that the *hisI*–*hisE* genes exhibit a different structure and organization in the diverse bacterial lineages.

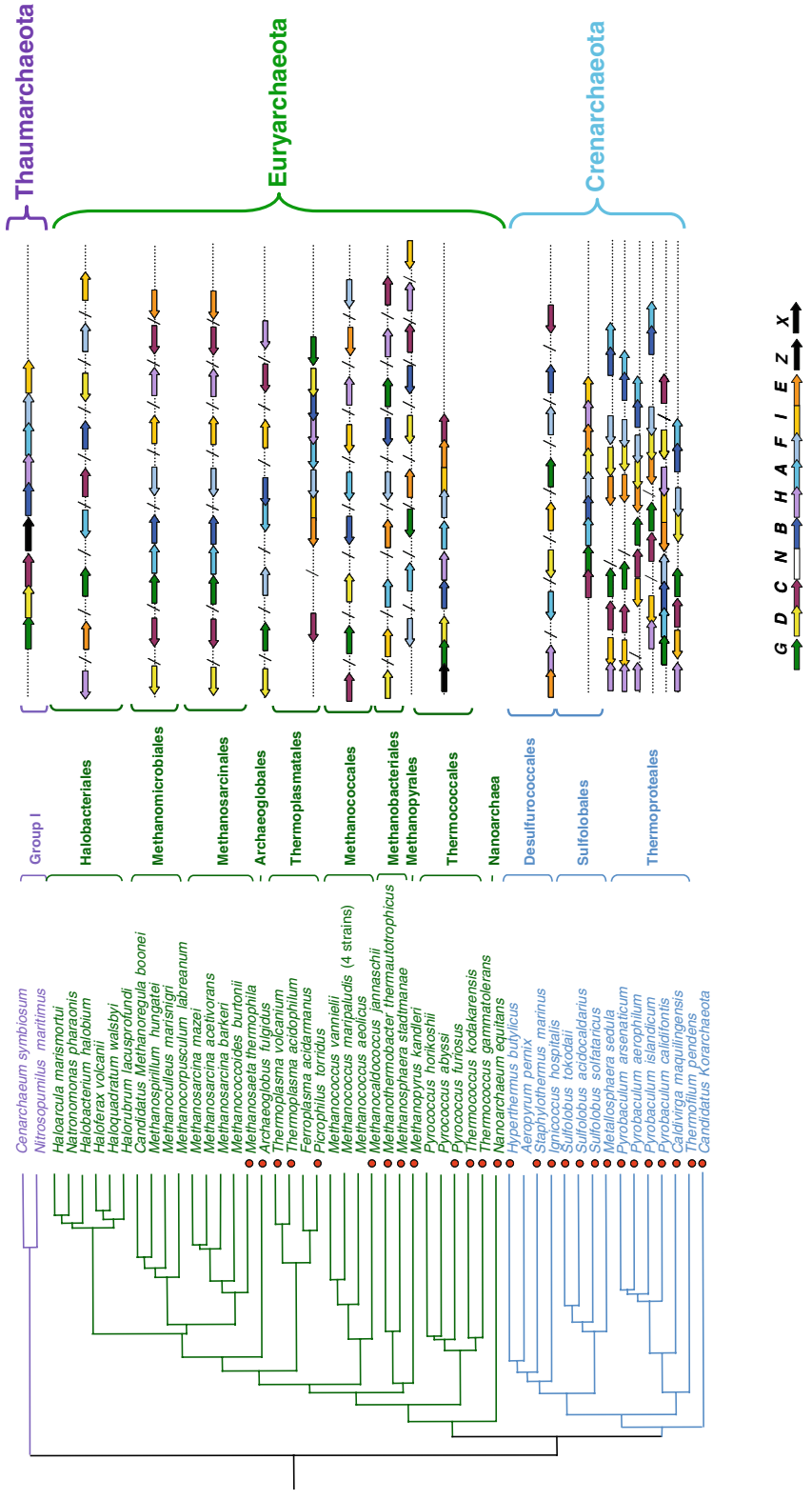


Fig. 2 A consensus archaeal phylogeny based on recent data showing the relationship between the main archaeal lineages together with their histidine biosynthetic genes organization. Dots near taxa indicate their thermophilic/hyperthermophilic lifestyle

In some bacteria, the two genes are fused in a bifunctional one encoding a protein endowed with both phosphoribosyl-ATP-pyrophosphatase (HisE) and phosphoribosyl-AMP-cyclohydrolase (HisI) activities that catalyze the second and third steps of histidine biosynthesis. In other bacteria they overlap, but are not fused, whereas in others they are separated, close to each other or scattered throughout the genome. When *hisI* and *hisE* are fused in a bifunctional gene, they are always arranged in the same relative order, with the *hisI* moiety located upstream of *hisE*. A gene with the two moieties arranged in the opposite order has not been found up to now, suggesting the existence of constraints in this gene fusion. The scattered phylogenetic distribution of *hisIE* fusion as well as the analysis of the organization of *his* genes has led to the suggestion that the fusion between *hisI* and *hisE* cistrons occurred more than once in the bacterial branch, speaking toward a phenomenon of convergent evolution (Fani et al. 2007). Once occurred, the fusion was fixed and vertically inherited. Moreover, this gene might have undergone multiple HGT events between different domains, or within the same domain, in the same or different lineages (Fani et al. 2007). The analysis of the presence/absence pattern of *hisI* and *hisE* in Archaea revealed that in most of them they encode a monofunctional enzyme (Table 1; Fig. 2) and only in some genomes (Thermococcales, Thermoplasmatales, and *Caldivirga maquillingensis*) they are fused in a single bifunctional cistron, *hisIE* (Fig. 2). Interestingly, in all these latter genomes, the bifunctional *hisIE* is embedded in a (compact) operon (see below and Fig. 2). Surprisingly, three archaeal genomes, *Archaeoglobus fulgidus*, *Caenarchaeum symbiosum* A, and *Nitrosopumilus maritimus* SCM1, lack the HisE coding gene. The lack of this gene in these three Archaea is obscure. Since *hisE* has been found in the vast majority of histidine-synthesizing organisms, it can be assumed that it codes for an essential enzyme. If this is true, the loss of this gene in a given organism should be paralleled by the concomitant replacement by another (and still unknown) gene performing the same function. However, the possibility that *hisE* code for a non-essential enzyme cannot be a priori ruled out.

The Structure of Archaeal hisG (and hisZ)

The regulation of the expression of the *his* operon has been particularly studied in *E. coli* and *S. typhimurium*, where the general mechanisms and the molecular details of the process are known (Alifano et al. 1996). In these two model organisms the biosynthetic pathway is under the control of distinct regulatory mechanisms operating at different levels and finely tuning the expression of the *his* genes. One of them is the feed-back inhibition by histidine of the activity of the first enzyme of the pathway, *N*-1-(5′-phosphoribosyl)-ATP transferase (ATP-PRT), which is coded for by *hisG*, allowing the almost instantaneous adjustment of the flow of intermediates along the pathway to the availability of exogenous histidine. The possibility of controlling the ATP-PRT activity relies on the existence of a C-terminal domain in the *E. coli* enzyme which binds to histidine, thus inhibiting the catalytic activity. A different mechanism responsible for the feedback inhibition of histidine biosynthesis has been identified in *L. lactis* (Sissler et al. 1999). In this bacterium, whose HisG lacks the C-terminal domain, the activity of ATP-PRT is controlled by the product of another gene, called *hisZ*. This mechanism is based on the interaction between HisZ and a shorter version of the HisG protein (referred to as $HisG_s$), which is per se unable to bind histidine, a function performed by HisZ, which then turns off HisG. Thus, HisZ regulates HisG and it is not present in γ-proteobacteria (Bovee et al. 2002; Kleeman and Parsons 1977; Lohkamp et al. 2004; Vega et al. 2005). The presence of *hisZ* in a genome is parallel to the presence of $hisG_s$. Probing the 55 archaeal genomes using the *A. variabilis* HisZ sequence returned four HisZ sequences from four microorganisms (Table 1; Fig. 2). Interestingly, all of them belong to Thermococcales where *his* genes are arranged in a (compact) operon with the same gene order and also harboring a bifunctional *hisIE* gene. Accordingly, the four corresponding HisG enzymes are shorter than the others. Besides, in a phylogenetic tree constructed with all the archaeal HisG sequences and with a set of $HisG_L$ (long HisG, *E. coli*-type) and $HisG_S$ sequences representatives of the bacterial domain, all the Thermococcales and Thermoplasmatales $HisG_s$ sequences do not cluster together with their archaeal counterparts but are instead placed inside the clade comprising the bacterial $HisG_S$ sequences (Fig. 3). It is worth of note that this archaeal clade comprises also the HisG sequence retrieved from *Picrophilus torridus*, although no HisZ-like sequence was retrieved in its genome (Table 1). Hence, the situation found in *P. torridus* (presence of $HisG_s$ and absence of HisZ) might represent an interesting exception to the histidine feedback inhibition system previously described and, in turn, may reveal the absence of such system acting in this microorganism.

Phylogeny of Concatenated His Proteins

The presence in histidine-synthesizing Thermococcales, Thermoplasmatales, and Thaumarchaeota of an operon whose gene order is identical to that of some bacteria, as well as the presence of a bifunctional *hisIE* gene and *hisZ* in some of them, raised the possibility that some *his* genes, entire operons or parts thereof might have undergone HGT between bacteria and Archaea. In order to check this

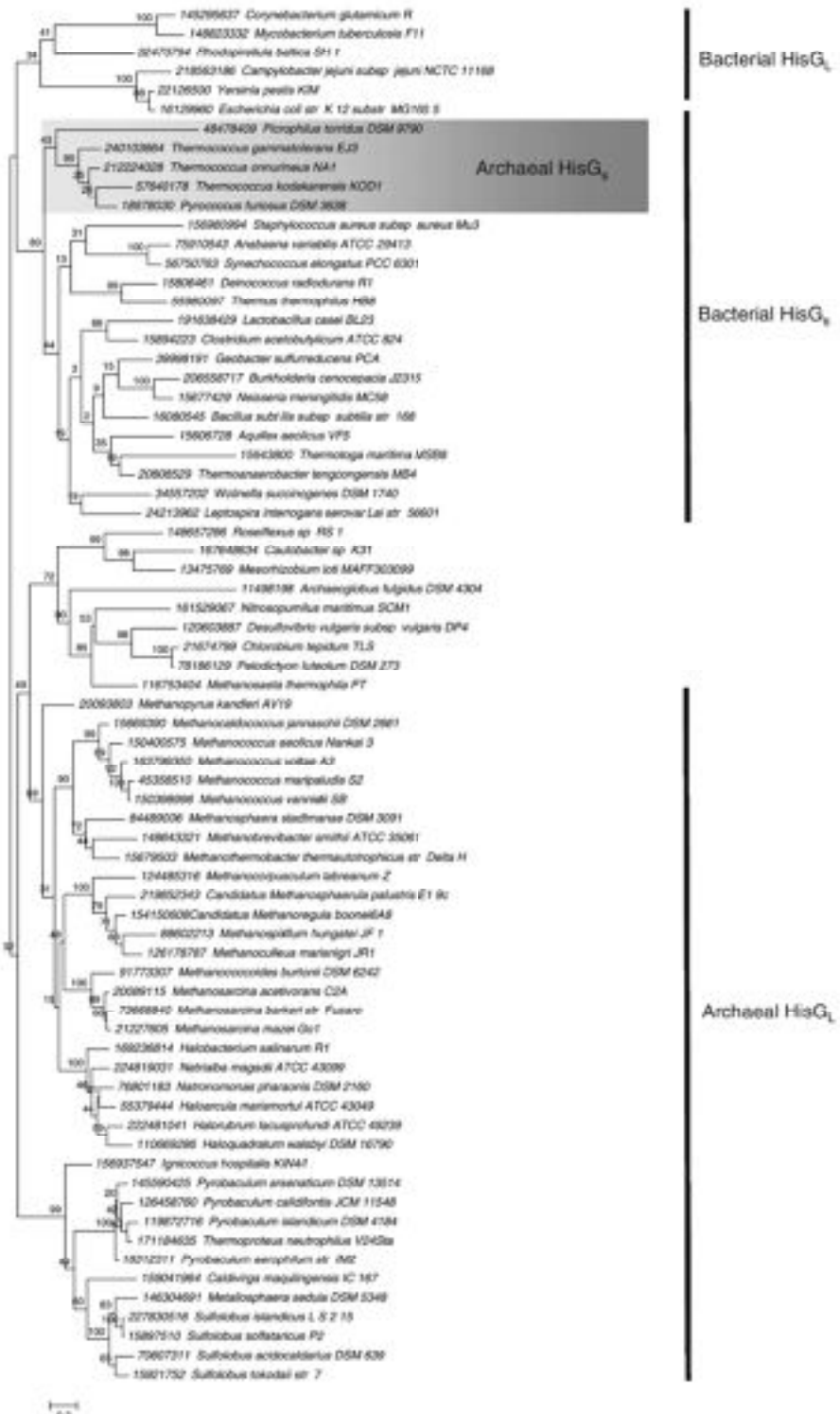


Fig. 3 Phylogenetic tree of all the HisG$_L$ sequences retrieved from the archaeal genomes and a set of bacterial counterparts (including both HisG$_L$ and HisG$_E$)

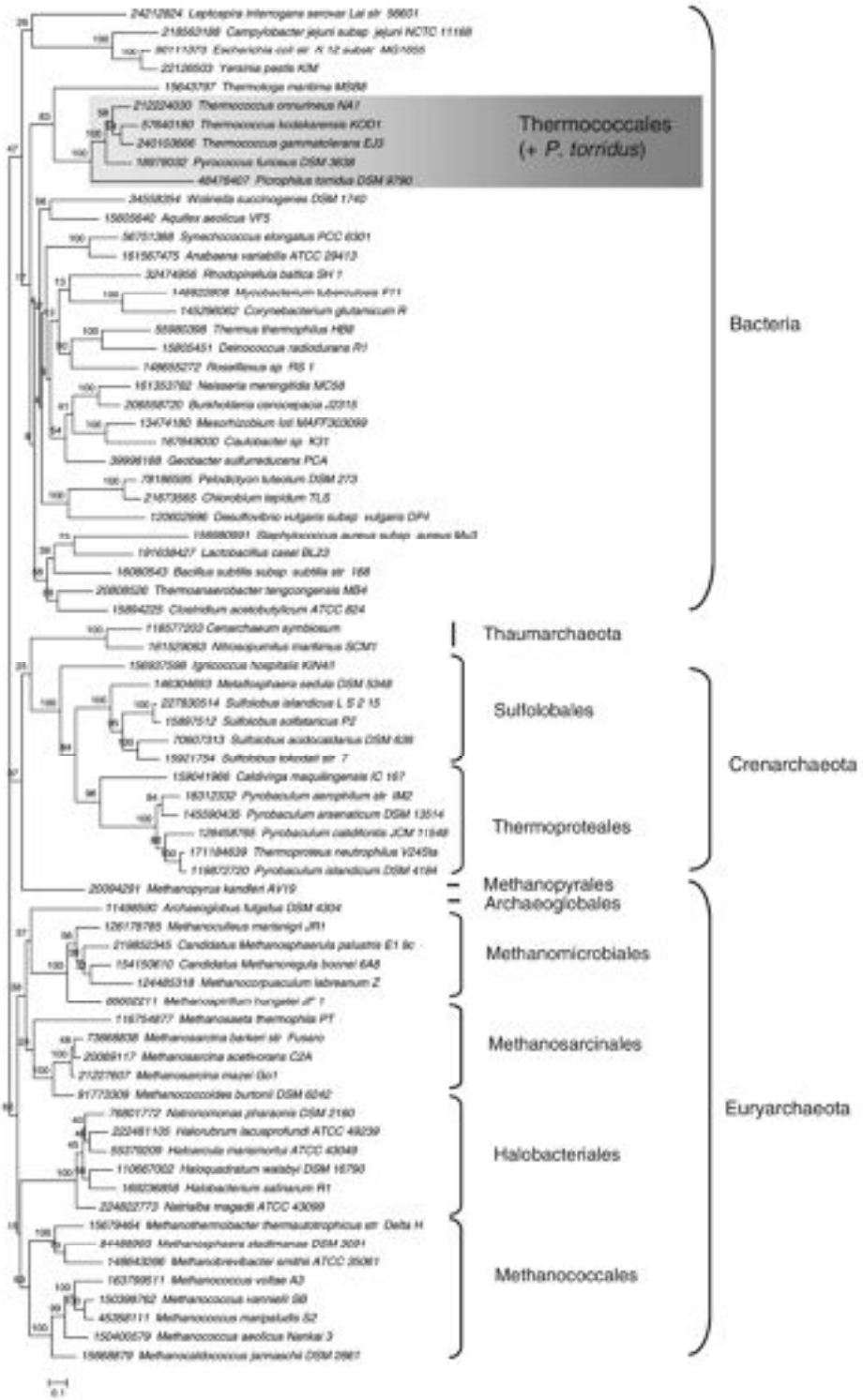


Fig. 4 Phylogenetic tree constructed using a concatenation of five histidine biosynthetic proteins sequences (HisGDBAF) from 45 Archaea and a set of representatives bacterial counterparts

possibility, a phylogenetic analysis was carried out on a data set containing five His proteins (HisGDBAF), which were concatenated into a dataset containing 886 positions. The analysis of the phylogenetic tree obtained (shown in Fig. 4) revealed a clear separation between archaeal and bacterial clades. Furthermore, in the archaeal part of the tree, the monophyly of the major clades (Brochier-Armanet et al. 2008) is respected (Fig. 4), suggesting the absence of HGT events involving the entire set of histidine biosynthetic genes. We cannot a priori exclude the possibility that one (or more) of the stand-alone *his* genes might have undergone HGT events between different Archaea and/or Bacteria. However, the phylogenetic trees constructed using the single His proteins did not help in elucidating the picture, since the bootstrap values were low and the microorganisms belonging to different archaeal clades were intermixed (data not shown). The only archaeal clade that falls outside from the archaeal cluster is represented by the group embedding the HisGDBAF sequences from the Thermococcales and *P. torridus*. In fact, the sequences retrieved from these microorganisms are placed as a well-supported monophyletic cluster in the bacterial part of the tree (Fig. 4). This result clearly indicates the transfer of the entire set of *his* genes (except for *hisN*) from bacteria to these Archaea. A possible scenario is that the *his* operon was acquired once by one of the two groups and then spread through HGT(s) among the Thermococcales group and *P. torridus* (Figs. 4, 5a, b).

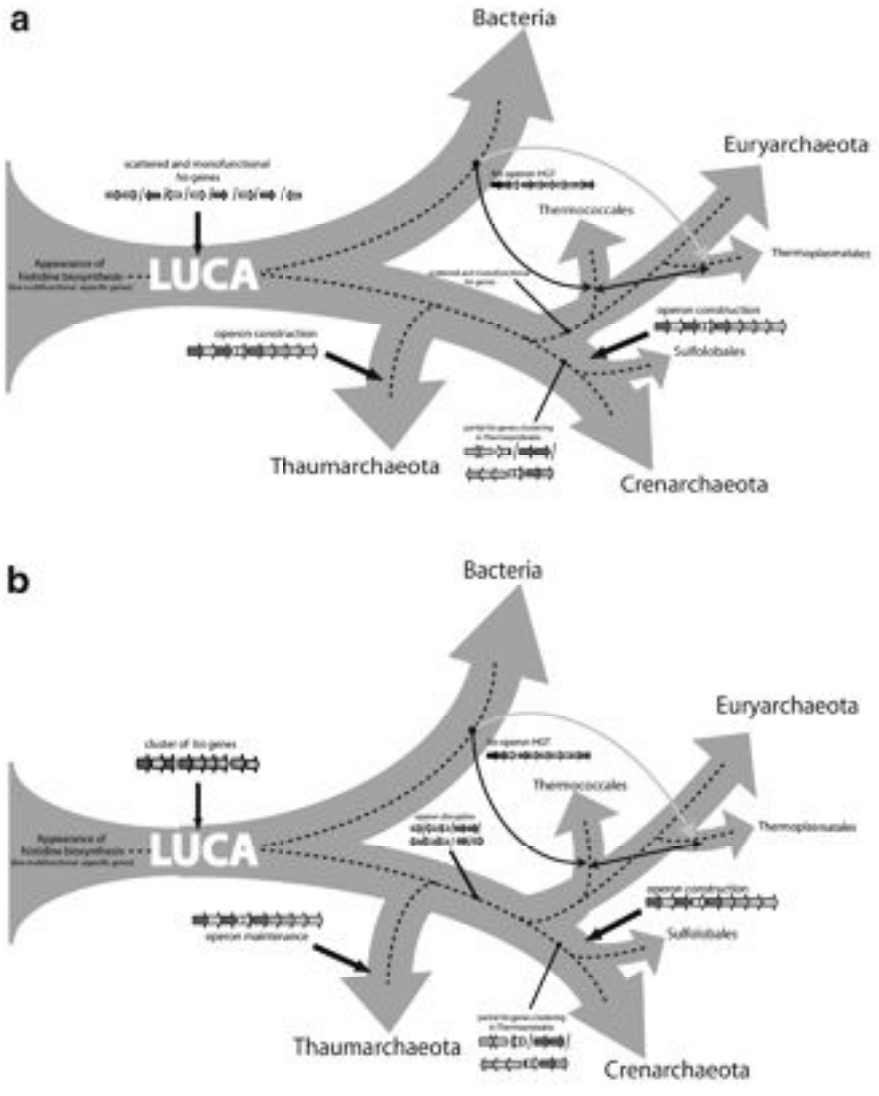


Fig. 5 Two alternative evolutionary models for the origin and evolution of histidine biosynthetic genes in extant Archaea. *Grey arrows* represent an alternative scenario accounting for the presence of a compact *his* operon in both Thermococcales and *P. torridus* genomes

A Model for the Evolution of *his* Genes in Archaea

In this paper, we report the analysis of the structure and organization of histidine biosynthetic genes in Archaea with the aim to understand the mechanism(s) and the forces that have driven the assembly of the histidine biosynthetic pathway, the eventual clustering of genes and operon construction. The analysis of the structure of the *his* genes gave a strong support to the hypothesis that at least three different molecular mechanisms played an important role in shaping the pathway, that is gene elongation, paralogous gene duplication(s), and gene fusion (Alifano et al. 1996; Fani et al. 1995, 2007). The analysis of *his* gene organization in different Archaea revealed that several gene arrays exist within this domain, with genes completely scattered throughout the chromosome, partially scattered/clustered, or strictly compacted (Fig. 2).

The whole body of data presented in this work, combined with previously obtained ones, suggests that histidine biosynthesis is an ancient metabolic route. How the biosynthesis of histidine actually emerged can only be surmised, but the phylogenetic distribution of the genes (Table 1) strongly suggests that the entire pathway existed in the last common ancestor of the three extant domains. Data obtained in this work combined with the analysis of structure and organization of *his* genes in Bacteria (Fani et al. 2007) allow us to propose that eight genes specifically involved in histidine biosynthesis (i.e., *hisGDBHAFIE*) were present in the genome of LUCA, coded for monofunctional enzymes (except HisD), and were scattered throughout the genome. Concerning the two remaining genes, *hisC* and *hisN*, coding for an aminotransferase and a HOL-Pase, respectively, it is quite possible that the two reactions carried out by the extant proteins they code for, might have been performed by enzymes with a broad substrate specificity. It is possible that during evolution a specific *hisC* and/or *hisN* gene might have been recruited from other metabolic processes through duplication and divergence, in agreement with the patchwork hypothesis on the origin and evolution of metabolic pathways (Jensen 1976).

Concerning the organization of histidine biosynthetic genes, it has been proposed (Fani et al. 2005) that, at least in Proteobacteria, the construction of a compact *his* operon would be a recent event in evolution and that it was piece wisely assembled by the joining of short sub-operons comprising two–three genes. On the other hand, Price et al. (2006) showed that a unified *hisGDC(NB)BHAF(I.E)* operon is present in some lineages. This result led the authors to hypothesize that this operon is ancient. The analysis reported in this work allows re-addressing this point (Fig. 5a, b). In principle, if a given phylogenetic lineage includes microorganisms showing a different organization of genes belonging to the same metabolic pathway, i.e., complete scattering, compact operons or partial scattering/partial clustering, at least two hypothetical scenarios can be invoked to explain such a picture:

(i) The LUCA harbored genes (partially) scattered throughout the genome, so that in some of the descendants the construction of clusters and/or operons occurred (Fig. 5a). According to this idea, histidine biosynthetic genes would have then been clustered independently in different archaeal lineages, that is Thaumarchaeota, Sulfolobales and (partially) in Thermoproteales. Conversely, all the other archaeal lineages would have maintained histidine biosynthetic genes (mainly) scattered throughout the chromosome(s). The fact that in Thaumarchaeota the gene order resembles the γ-proteobacterial one may suggest that an event of HGT might have been responsible for the appearance of the *his* genes array in Thaumarchaeota. However, in the phylogenetic tree constructed with a concatenation of a set of representative His protein sequences, the members of this phylum are embedded in a well-supported monophyletic cluster in the archaeal part of the tree, ruling out the hypothesis of an HGT event. The same can be said for the other archaeal lineages displaying a (more or less) compact histidine operon (Sulfolobales and Thermoproteales). Noteworthy, according to this hypothesis, a strong selective pressure in recreating the same relative gene order (starting from scattered genes) must be invoked to account for the emergence of two nearly identical operons in two distantly related prokaryotic clades (Thaumarchaeota and γ-proteobacteria).

(ii) The genome of the LUCA contained genes organized in operons and this organization was completely or partially destroyed during evolution in some of the descendants' lineages (Fig. 5b). Although it is not possible to infer the gene order of this ancient gene cluster, this latter scenario would predict the disruption of the operon in different phylogenetic lineages (all the Euryarchaeal lineages and the Crenarchaeal groups of Thermoproteales and Desulfurococcales) and then the re-assembling of the *his* genes in a different way in respect to the ancestral state (in Sulfolobales). Since Thaumarchaeota likely represent the first emerging lineage in the Archaeal domain (Brochier-Armanet et al. 2008) their histidine genes organization may resemble the ancestral one. However, the fact that, within bacteria, only γ-proteobacteria display a similar organization (*hisGDCNBHAFIE*), whereas in others phyla the histidine genes are differently organized (e.g. *hisDCBHA-impA-FI* in Actinobacteria and *hisZGDBHAFI* in (most of) Firmicutes) would imply that histidine biosynthetic genes would have re-assembled differently in a number of bacterial phyla, while they

would have been kept with the ancestral arrangement in γ-proteobacteria). Moreover, if this scenario is correct, this would mean that the destruction of operon organization should have given rise to scattered genes and/or mini-operons with the creation of new promoters upstream of each of them (Itoh et al. 1999) and the eventual reconstruction of a compact *his* operon with a different gene order in different archaeal lineages.

Concerning the biological significance of the origin and evolution of operons, this issue is still under debate, and at least six different classes of models (listed below) have been proposed to explain the existence of operons:

(i) The co-regulation model predicts that genes are clustered together because regulation is easier under a single promoter, providing both economy of transcription and equal abundance of products, especially when genes belong to the same metabolic pathway.
(ii) The Natal model predicts that operons originated by in situ gene duplication and divergence, whereby the evolution of metabolic pathways took place in a stepwise fashion in an "assembly line of genes" (Horowitz 1945).
(iii) The Fisher Model proposes that the physical proximity of co-adapted alleles in the genome reduces the frequency of the formation of unfavorable combinations of genes by recombination events. This might favor the operon assembly.
(iv) In the molarity model, co-regulation can also guarantee that proteins are synthesized in equimolar amounts reducing stochastic differences in their concentration levels (Swain 2004) and can increase the rate of both formation and folding of multisubunit protein complexes (Dandekar et al. 1998; Pal and Hurst 2004).
(v) The selfish operon theory (Lawrence 1999; Lawrence and Roth 1996) posits that operons—except for some highly conserved operons, thought to be ancient and to have been formed by other mechanisms—form because such compact organization facilitates HGT (and so survival) of non-essential gene clusters, whose function is only occasionally useful and so prone to random deletion of genes by mutation pressure and genetic drift. HGT would save the cluster from extinction and might confer selective advantage(s) to the recipient organism in some environmental conditions.
(vi) Glansdorff (1999) suggested that early adaptation to thermophily played a key role in the emergence of operons. This is supported by the transcription–translation coupling, which is seen as a mechanism capable of protecting the messenger RNA from the degradation caused by high temperatures.

In order to identify the existence of a possible correlation exist between the archaeal lifestyle and the *his* genes organization, as well as the most probable operon formation driving forces, we mapped on the phylogenetic tree the thermophily/hyperthermophily of the corresponding taxa (Fig. 2). Our results revealed that in most cases when the *his* genes are completely clustered in operons or sub-operons, they belong to thermophilic/hyperthermophilic Archaea and that most, but not all, mesophilic Archaea harbor only scattered *his* genes. Thus, apparently, one of the forces driving the assembly of *his* genes into operons might have been the adaptation to high temperature. However, some exceptions to this rule exist in Archaea (see for example Thaumarchaeota). Furthermore, the finding that the entire histidine operon has been horizontally transferred between Bacteria and Archaea might be in agreement with the proposal by Lawrence and Roth (1996), i.e., the Selfish Operon Model. Therefore, different forces might have driven the assembly of more or less compact his operons in Archaea.

Concerning the mechanisms of operon construction, a new idea, referred to as the "piece-wise" model, was recently proposed to explain the origin and evolution of some operons (Fani et al. 2005). According to this model, long and complex operons can be assembled through the progressive clustering of pre-existing sub-operons embedding part of the genes of the final, completely assembled operon. Even though the model was originally suggested to explain the origin and evolution of the proteobacterial histidine operon (Fani et al. 2005), it might be applied to the origin and evolution of any complex operon. The assembly of scattered genes into sub-operons might proceed through different mechanisms. According to Horowitz (1965), in-tandem duplication of ancestral genes may lead to bi- or multicistronic operons; other genes could be recruited via the patchwork mechanism (Jensen 1976) and put close to other genes via recombination or transposition. These sub-operons might be evolutionary fixed by different forces: the necessity of equimolarity and/or co-regulation or the formation of metabolon-like structures. In our opinion, such a mechanism might have also acted in the assembly of the Sulfolobales *his* operon. Indeed the presence in Thermoproteales of sub-operons whose gene order (*hisCG*, *hisFDE*) resembles that of Sulfolobales one (CGABFDEHI) is in agreement with the piece-wise model.

Conclusions

Metabolic pathways of the earliest heterotrophic organisms presumably arose in conjunction with the exhaustion of the prebiotic compounds present in the primordial soup. In the course of molecular and cellular evolution different mechanisms and different forces might have concurred in the rise

of novel metabolic abilities and in the shaping of metabolic routes. The dissemination of metabolic routes between microorganisms might have occurred by horizontal transfer (xenology) or cell fusion (synology) events which could have been facilitated by the absence of a cell wall in primordial cells. The horizontal transfer of entire metabolic pathways or part thereof might have had a special role in shaping genome architectures and in fostering genetic adaptation and evolution during the early stages of cellular evolution when, according to Woese (1998), the "genetic temperature" was high. Gaining of new metabolic traits might have been facilitated by the operon organization of early genes that would have permitted the transfer of entire metabolic routes. However, the organization of genes belonging to the same metabolic pathway followed different routes in different microorganisms. In some of them these genes were scattered throughout the genome, forming regulons, in other cases they were organized in more or less compact operons. Concerning the timing of operon formation, it is possible that some of them are very ancient (i.e., the ribosomal superoperon) (Fondi et al. 2009), whereas others might have been assembled more recently once or multiple times during evolution and then vertically and/or horizontally transferred between organisms belonging to the same or different species/genus. Different forces may have driven operon construction, which might have occurred through in-tandem duplication of ancestral genes or through the recruitment of genes located in other chromosomal loci. It is also possible that the longest operons might have been constructed piecewise by the progressive assembly of shorter sub-operons.

The heterogeneous distribution and organization of *his* genes in Archaea reported in this work suggests the absence of a complete histidine biosynthetic operon in the Archaeal ancestor and probably also in LUCA. Present data, despite not allowing saying whether histidine biosynthetic genes were embedded in a compact operon in the LUCA, revealed that they underwent several recombination events during evolution and this led to the different schemes of *his* genes organization that we observe in modern Archaea (and Bacteria). The organization of *his* genes in some extant archaeal lineages speaks toward a piece-wise construction of *his* sub-operons along with gene fusion events and HTG from bacterial donor. Lastly, data suggest also that different molecular mechanisms may drive operon formation and metabolic pathway origin and evolution.

References

Alifano P, Fani R, Liò P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev 60:44–69

Bovee M, Champagne K, Demeler B, Francklyn C (2002) The quaternary structure of the HisZ-HisG N-1-(5′-phosphoribosyl)-ATP transferase from *Lactococcus lactis*. Biochemistry 41: 11838–11846

Brenner M, Ames B (1971) The histidine operon and its regulation. In: Vogel H (ed) Metabolic pathways. Academic Press, New York, pp 349–387

Brilli M, Fani R (2004) Molecular evolution of *hisB* genes. J Mol Evol 58:225–237

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol 6:245–252

Carlomagno M, Chiarotti L, Alifano P, Nappo A, Bruni C (1988) Structure of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons. J Mol Biol 203:585–606

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23:324–328

Donahue T, Farabaugh P, Fink G (1982) The nucleotide sequence of the His4 region of yeast. Gene 18:47–59

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

Fani R (2004) Gene duplication and gene loading. In: Miller RV, Day MJ (eds) Microbial evolution: gene establishment, survival, and exchange. ASM Press, Washington, pp 67–81

Fani R, Fondi M (2009) Origin and evolution of metabolic pathways. Phys Life Rev 6:23–52

Fani R, Liò P, Chiarelli I, Bazzicalupo M (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. J Mol Evol 38:489–495

Fani R, Liò P, Lazcano A (1995) Molecular evolution of the histidine biosynthetic pathway. J Mol Evol 41:760–774

Fani R, Mori E, Tamburini E, Lazcano A (1998) Evolution of the structure and chromosomal distribution of histidine biosynthetic genes. Orig Life Evol Biosph 28:555–570

Fani R, Brilli M, Liò P (2005) The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. J Mol Evol 60:378–390

Fani R, Brilli M, Fondi M, Liò P (2007) The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. BMC Evol Biol 7(Suppl 2):S4

Fondi M, Emiliani G, Fani R (2009) Origin and evolution of operons and metabolic pathways. Res Microbiol. doi: 10.1016/j.resmic.2009.05.001

Glansdorff N (1999) On the origin of operons and their possible role in evolution toward thermophily. J Mol Evol 49:432–438

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Horowitz NH (1945) On the evolution of biochemical syntheses. Proc Natl Acad Sci USA 31:153–157

Horowitz NH (1965) The evolution of biochemical syntheses—retrospect and prospect. Academic Press, New York

Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol 16:332–346

Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409–425

Kleeman J, Parsons S (1977) Inhibition of histidyl-tRNA-adenosine triphosphate phosphoribosyltransferase complex formation by histidine and by guanosine tetraphosphate. Proc Natl Acad Sci USA 74:1535–1537

Klemm T, Davisson V (1993) Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. Biochemistry 32:5177–5186

Kuenzler M, Balmelli T, Egli C, Paravicini G, Braus G (1993) Cloning, primary structure, and regulation of the *HIS7* gene encoding a bifunctional glutamine amidotransferase: cyclase from *Saccharomyces cerevisiae*. J Bacteriol 175:5548–5558

Lawrence J (1999) Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol 2:519–523

Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–1860

Lazcano A, Fox G, Oró J (1992) Life before DNA: the origin and evolution of early Archean cells. In: Mortlock R, Gallo M (eds) The evolution of metabolic functions. CRC Press, Boca Raton, pp 1–13

Lohkamp B, McDermott G, Campbell S, Coggins J, Lapthorn A (2004) The structure of *Escherichia coli* ATP-phosphoribosyltransferase: identification of substrate binding sites and mode of AMP inhibition. J Mol Biol 336:131–144

Maurel M, Ninio J (1987) Catalysis by a prebiotic nucleotide analog of histidine. Biochimie 69:551–553

Pal C, Hurst LD (2004) Evidence against the selfish operon theory. Trends Genet 20:232–234

Price M, Alm L, Arkin A (2006) The histidine operon is ancient. J Mol Evol 62:807–808

Shen C, Lazcano A, Oró J (1990a) The enhancement activities of histidyl-histidine in some prebiotic reactions. J Mol Biol 31:445–452

Shen C, Mills T, Oró J (1990b) Prebiotic synthesis of histidyl-histidine. J Mol Biol 31:175–179

Shen C, Yang L, Miller S, Oró J (1990c) Prebiotic synthesis of histidine. J Mol Biol 31:167–174

Sissler M, Delorme C, Bond J, Ehrlich S, Renault P, Francklyn C (1999) An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis. Proc Natl Acad Sci USA 96:8985–8990

Swain PS (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. J Mol Biol 344:965–976

Vega M, Zou P, Fernandez F, Murphy G, Sterner R, Popov A, Wilmanns M (2005) Regulation of the hetero-octameric ATP phosphoribosyl transferase complex from *Thermotoga maritima* by a tRNA synthetase-like subunit. Mol Microbiol 55:675–686

Weber A, Miller S (1981) Reasons for the occurrence of the twenty coded protein amino acids. J Mol Evol 17:273–284

White H (1976) Coenzymes as fossils of an earlier metabolic state. J Mol Evol 7:101–117

White D, Erickson J (1980) Catalysis of peptide bond formation by histidyl-histidine in a fluctuating clay environment. J Mol Biol 16:279–290

Winkler M (1987) Biosynthesis of histidine. In: Neidhardt F, Ingraham J, Low K, Magasanik B, Schaechter M, Humbarger H (eds) *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. American Society for Microbiology, Washington, pp 395–411

Woese C (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859

Ycas M (1974) On earlier states of the biochemical system. J Theor Biol 44:145–160

3.3 Structural, evolutionary and genetic analysis of the histidine biosynthetic core in the genus *Burkholderia*

In this work a detailed analysis of the structure, the expression and the organization of *his* genes belonging to the core of histidine biosynthesis (*hisBHAF*) in 40 newly determined and 13 available sequences of *Burkholderia* strains was carried out. Data obtained revealed a strong conservation of the structure and organization of these genes through the entire genus. The phylogenetic analysis showed the monophyletic origin of this gene cluster and indicated that it did not undergo horizontal gene transfer events. The analysis of the intergenic regions, based on the substitution rate, entropy plot and bendability suggested the existence of a putative transcription promoter upstream of *hisB*, that was supported by the genetic analysis that showed that this cluster was able to complement Escherichia coli *hisA*, *hisB*, and *hisF* mutations. Moreover, a preliminary transcriptional analysis and the analysis of microarray data revealed that the expression of the his core was constitutive. These findings are in agreement with the fact that the entire *Burkholderia his* operon is heterogeneous, in that it contains alien genes apparently not involved in histidine biosynthesis. Besides, they also support the idea that the proteobacterial his operon was piecewisely assembled, i.e. through accretion of smaller units containing only some of the genes (eventually together with their own promoters) involved in this biosynthetic route. The correlation existing between the structure, organization and regulation of his core genes and the function(s) they perform in cellular metabolism is discussed.

Gene 448 (2009) 16–28

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Structural, evolutionary and genetic analysis of the histidine biosynthetic "core" in the genus *Burkholderia*

Maria Cristiana Papaleo [a], Edda Russo [a], Marco Fondi [a], Giovanni Emiliani [b], Antonio Frandi [a], Matteo Brilli [c], Roberta Pastorelli [d], Renato Fani [a,*]

[a] Department of Evolutionary Biology, Via Romana 17-19, University of Florence, 50125 Florence, Italy
[b] Department of Environmental and Forestry Sciences, via S. Bonaventura 13, 50145 University of Florence, Italy
[c] UMR CNRS 5558 – LBBE "Biométrie et Biologie évolutive", UCB Lyon 1 - Bât. Grégor Mendel, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex
[d] Research Centre of Agrobiology and Pedology, Piazza M. D'Azeglio 30, Agricultural Research Council (CRA) Florence, Italy

ARTICLE INFO

Article history:
Received 27 April 2009
Received in revised form 25 July 2009
Accepted 5 August 2009
Available online 13 August 2009

Received by R. Britton

Keywords:
Operon evolution
Alien genes
Histidine genes

ABSTRACT

In this work a detailed analysis of the structure, the expression and the organization of his genes belonging to the core of histidine biosynthesis (hisBHAF) in 40 newly determined and 13 available sequences of *Burkholderia* strains was carried out. Data obtained revealed a strong conservation of the structure and organization of these genes through the entire genus. The phylogenetic analysis showed the monophyletic origin of this gene cluster and indicated that it did not undergo horizontal gene transfer events. The analysis of the intergenic regions, based on the substitution rate, entropy plot and bendability suggested the existence of a putative transcription promoter upstream of *hisB*, that was supported by the genetic analysis that showed that this cluster was able to complement *Escherichia coli hisA*, *hisB*, and *hisF* mutations. Moreover, a preliminary transcriptional analysis and the analysis of microarray data revealed that the expression of the *his* core was constitutive. These findings are in agreement with the fact that the entire *Burkholderia his* operon is heterogeneous, in that it contains "alien" genes apparently not involved in histidine biosynthesis. Besides, they also support the idea that the proteobacterial his operon was piece-wisely assembled, i.e. through accretion of smaller units containing only some of the genes (eventually together with their own promoters) involved in this biosynthetic route. The correlation existing between the structure, organization and regulation of *his* "core" genes and the function(s) they perform in cellular metabolism is discussed.

1. Introduction

Histidine biosynthesis is one of the most studied anabolic pathways. It has been studied for over 40 years in *Escherichia coli* and its close relative *Salmonella enterica* (formerly *Salmonella typhimurium*), leading to the accumulation of a very large body of biochemical, genetic, molecular and physiological data (Alifano et al., 1996). Histidine biosynthesis consists of nine intermediates and of eight distinct proteins that in the two enterobacterial species are encoded by eight genes organized in a very compact operon and arranged in the order *hisGDC(NB)HAF(IE)* (Alifano et al., 1996; Fani et al., 1997, 2006). Four of the *his* genes (*hisBHAF*) are particularly interesting from an evolutionary viewpoint and form the so-called "core" of the pathway (Fig. 1), which plays an important role in cellular metabolism. Indeed, it is a metabolic cross-point interconnecting histidine biosynthesis to both nitrogen metabolism and *de novo* synthesis of purines. The available information also showed that after the assembly of the entire pathway, the structure and/or organization of *his* genes underwent major rearrangements in the three domains, which generated a wide variety of structural and/or clustering strategies of *his* genes (Fani et al., 1998, 2005). Thus, the analysis of the structure and organization of *his* genes might help in shedding some light on the origin and evolution of operons (Fani et al., 2005; Price et al., 2006). Recently, we proposed that the proteobacterial *his* operon might be a recent invention of evolution and was piece-wisely constructed (Fani et al., 2005, 2006). According to the model proposed, the *his* genes, scattered on the genome of proteobacterial ancestor, underwent a progressive clustering that culminated in some γ-proteobacteria where the

Abbreviations: hisG, ATP phosphoribosyl transferase (EC 2.4.2.17); *hisD*, histidinol dehydrogenase (EC 1.1.1.23); *hisC*, histidinol-phosphate aminotransferase (EC 2.6.1.9); *hisN*, histidinol-phosphate phosphatase (EC 3.1.3.15); *hisB*, imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19); *hisH*, G-type glutamine amidotransferase; *hisA*, [*N*-(5′-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide isomerise (EC 5.3.1.16); *hisF*, imidazole glycerol phosphate synthase subunit HisF (EC 4.1.3.-); *hisI*, phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19); *hisE*, phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31); *murA*, UDP-N-acetylglucosamine 1-carboxyvinyl-transferase; *murC*, integral membrane protein; Bcc, *Burkholderia cepacia* complex.

* Corresponding author. Tel.: +39 55 2288244; fax: +39 55 2288250.
E-mail address: renato.fani@unifi.it (R. Fani).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.08.002

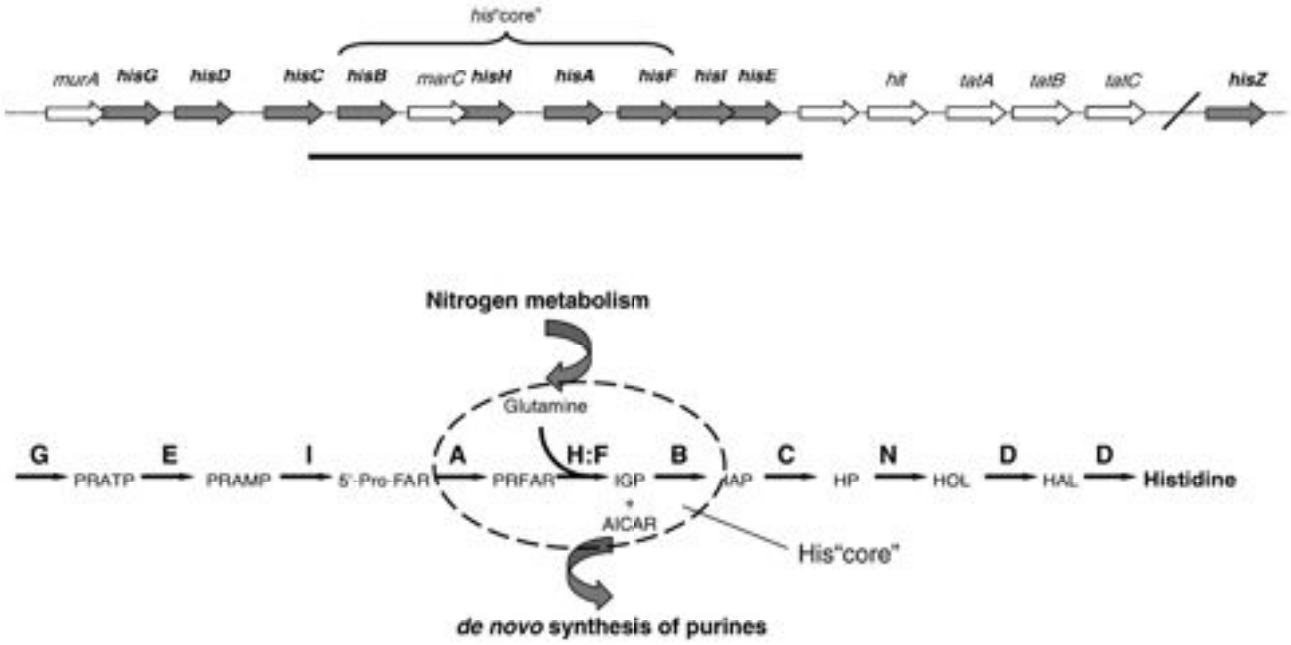


Fig. 1. Structure and organization of the histidine biosynthetic genes in the genus *Burkholderia*. The bold line below the *his* operon represents the region amplified via PCR from 40 Bcc strains.

operons are very compact and include fused and/or overlapping genes. The first step of the operon construction was the formation of the *his* "core" (Alifano et al., 1996; Fani et al., 1995), an event likely occurring in the ancestor of α/β/γ proteobacteria (or even predating its appearance). Then, the ancestral *hisBHAF* operon joined two other independently formed sub-operons (*hisGDC* and *hisIE*) giving an almost complete operon. A corollary of this model is the following: if the evolution of long *his* compact operons actually happened this way, one can imagine intermediate stages in which the *his* genes were organized in less compact operons, eventually containing one or more functionally unrelated ("alien") genes (*heterogeneous* operons), which during evolution were excised and/or transposed to other chromosomal locations. In principle, since the *heterogeneous* operons contain genes not involved in histidine biosynthesis, the pathway should not be tuned as finely as in *E. coli* and *S. enterica* and might be constitutively expressed. Lastly, if the piece-wise model (Fani et al., 2005) is correct, in these operons the internal promoters identified in *E. coli* and *S. enterica* and located upstream of *hisB* and/or *hisI* might still be present and functional. Even though this idea might be intuitive, in our knowledge it has not been demonstrated (at least) for histidine biosynthesis. The only data available are those concerning the α-proteobacterium *Azospirillum brasilense*, possessing a *his* heterogeneous cluster *hisBHorf1AFEhit* where a transcription promoter located just upstream of *hisB* allows the constitutive expression of the downstream genes (Fani et al., 1989, 1993).

Additionally, it is not clear yet whether the gene structure, organization and/or regulation are linked to the taxonomical position of a given strain/species/genus and/or to the metabolism of the organism(s). In some cases (Fondi et al., 2007) the appearance of a given gene structure (i.e. gene fusion) and/or organization can be mapped in a taxonomical branch, whereas in others (Fani et al., 2005) it appears much less related to the phylogeny. Thus, the analysis of genes belonging to a heterogeneous *his* operon at multiple taxonomic scales, ranging from strain to genus level, which has not been carried out until now, can shed some light on these issues. In this context the *his* "core" represents an interesting case study. Thus, the aim of this work was to study the *his* "core" from different "non-model" strains harboring a heterogeneous *his* operon by an approach combining both bioinformatic and experimental methods. Such an analysis may provide useful hints on: i) the degree of conservation of structure and organization of genes belonging to the *his* core at the genus, species and strain level; ii) the degree of conservation/divergence of intergenic sequences; iii) the presence of transcription regulatory signals within them and the degree of their conservation at either taxonomic (between strains belonging to the same or to different species/genus) or ecological level (between strains occupying different environmental niches); iv) the regulation of *his* "core" transcription in microorganisms different from enterobacteria and whether expression is constitutive or either regulated by histidine concentrations; v) the presence and the nature of alien genes; and vi) the validity of the piece-wise model of the origin and evolution of proteobacterial *his* operons. To this purpose, we focused our attention on β-proteobacteria, a taxon including bacteria harboring homogeneous as well as heterogeneous *his* operons and representing key organisms in the construction of compact proteobacterial *his* operons.Within β-proteobacteria, the *Burkholderia cepacia* complex (Bcc) is relevant in ecology and human health. Indeed, these bacteria are important opportunistic pathogens causing lung infections in immuno-compromised patients, especially those with cystic fibrosis (CF). In contrast to these pathogenic properties, Bcc organisms are also found in natural habitats and have a considerable ecological and commercial importance. In fact, some Bcc strains are able to catabolize many toxic compounds and to produce a number of substances that antagonize soil borne plant pathogens. Therefore, the Bcc represents an interesting and heterogeneous taxonomical entity, comprising strains from different environments and with different metabolic abilities. Last, the different ecological niches occupied by Bcc strains and/or species may influence the regulation of *his* genes and provide some insights into the conservation/divergence of transcriptional regulatory signals. A very preliminary analysis of the organization of histidine biosynthetic genes carried out in a very limited number of *Burkholderia* strains (Fani et al., 2005) revealed that the *his* genes are organized in a *heterogeneous* operon comprising 16 genes, seven apparently not belonging to the known histidine biosynthesis chain of reactions (Fig. 1).

Thus, in this work the structure of the histidine biosynthetic core (*hisBHAF*) was studied within the genus *Burkholderia* at multiple

taxonomic levels by analyzing it in a panel of 40 Bcc strains representative of nine species from both clinical and environmental sources as well as in the 13 sequences available in databases.

2. Materials and methods

2.1. Bacterial strains and plasmids

The *E. coli* strains used were DH5α™ F,$^-$ ϕ80 *dlacZ*ΔM15Δ (*lacZYA-argF*) U169, *deoR*, *recA1*, *endA1*, *hsdR17*(rk$^-$, mk$^+$), phoA, *supE44*, λ$^-$, *thi-1*, *gyrA96*, *relA1* (Life Technologies), FB251 *his855 recA56* (Grisolia et al., 1982), FB182 *hisF892* and FB184 *hisA915* (Goldschmidt et al., 1970). The Bcc strains used in this work are listed in Table 1. The plasmid vector used was pGEM-T EASY VECTOR (Amp^r) (Promega), a 3015 bp molecule *ad hoc* constructed for cloning of PCR products. The recombinant plasmids used in this work are listed in Table 2.

2.2. Media and culture preparation

The minimal medium (MMD) used was described by Davis and Mingioli (1950). SOB, SOC, TFB, FSB and LB media were prepared as previously described (Goldschmidt et al., 1970). *E. coli* cells were grown at 37 °C for 24 to 48 h. Ampicillin, IPTG, X-GAL, and histidine were used at 100 μg/ml, 40 μg/ml, 32 μg/ml and 25 μg/ml respectively.

Table 2
Plasmids used in this work.

Plasmid	Relevant genotype	Source
pGEM-T easy vector	Amp^r; *lacZ*	Promega[a]
pBC1	Amp^r; Δ-*10TAbox* (+)[b]; *hisB marCHisH hisA hisF*	This work
pBC2	Amp^r; Δ-*10TAbox* (−); *hisB marCHisH hisA hisF*	This work
pBC3	Amp^r; Δ-*35GCbox* (+); *hisB marC hisH hisA hisF*	This work
pBC4	Amp^r; Δ-*35GCbox* (−); *hisB marCHisH hisA hisF*	This work
pBC5	Amp^r; *hisB marCHisH hisA hisF*[b] (+)	This work
pBC6	Amp^r; *hisB marCHisH hisA hisF* (−)	This work
pBC7	Amp^r; 3′-*hisB marCHisH hisA hisF* (+)	This work
pBC8	Amp^r; 3′-*hisB marCHisH hisA hisF* (−)	This work
pBC9	Amp^r; 3′-*hisH hisA hisF* (+)	This work
pBC10	Amp^r; 3′-*hisH hisA hisF* (−)	This work
pBC11	Amp^r; 3′-*hisA hisF* (+)	This work
pBC12	Amp^r; 3′-*hisA hisF* (−)	This work

[a] Positively or negatively orientated in respect to the *lacZ* promoter.
[b] The cloned fragment *hisB marCHisH hisA hisF* containing the entire biosynthetic core spanning from the 3′ end of *hisC* to the 5′ end of *hisI* belonging to strain LMG16670.

Table 1
Burkholderia cepacia complex *his* sequences used in this work.

Strain	Gvr	Species	Accession no.	Length (bp)	Origin	Reference
FCF2	I	*B. cepacia*	EU057663	4810	CF	a
LMG 2161			EU057647	4811	Environmental	b
FCF7	II	*B. multivorans*	EU057686	4829	CF	a
LMG 18822			EU057652	4825		b
LMG 13010			EU057648	4829		
LMG 17588			EU057651	4829	Environmental	
FCF13	III-A	*B. cenocepacia*	EU057684	4803	CF	a
FCF15			EU057646	4803		
FCF16			EU057679	4803		
FCF17			EU057675	4803		
FCF19	III-B		EU057667	4803		
FCF20			EU057668	4804		
FCF25			EU057664	4803		
FCF22			EU057670	4803		
FCF23			EU057671	4797		
FCF24			EU057662	4804		
FCF27			EU057683	4803		
FCF28			EU057685	4803		
FCF30			EU057672	4803		
FCF31			EU057673	4802		
LMG 19230	III-C		EU057655	4805	Environmental	c
LMG 19240			EU057656	4806		
FCF32	III-D		EU057678	4802	CF	a
FCF34			EU057645	4802		
FCF37			EU057665	4802		
FCF38			EU057674	4787		
FCF39			EU057677	4802		
FCF41	IV	*B. stabilis*	EU057669	4804		
LMG 14294			EU057649	4766		b
TVV75/LMG 10929	V	*B. vietnamiensis*	FJ460229	4611	Environmental	
LMG 18941	VI	*B. dolosa*	EU057653	4833	CF	a
LMG 18942			EU057654	4833		
MCI 7	VII	*B. ambifaria*	EU057659	4807	Environmental	d
LMG 19467			EU057657	4807	CF	e
LMG 16670	VIII	*B. anthina*	EU057650	4967	Environmental	
LMG 20980			EU057658	4967		
FCF43	IX	*B. pyrrocinia*	EU057680	4807	CF	a
FCF44			EU057682	4804		
ATCC15958			EU057644	4808	Environmental	d
MVPC 1/26			EU057666	4818		f

[a] Tabacchioni et al., 2008.
[b] Mahenthiralingam et al., 2000.
[c] Balandreau et al., 2001.
[d] Ciccillo et al., 2002.
[e] Coenye et al., 2003.
[f] Fiore et al., 2001.

2.3. Preparation of template DNA from bacterial cultures

Genomic DNA of each bacterial isolate was prepared using a "Nucleo-Spin Tissue" (Macherey-Nagel) and analyzed by 0.8% (wt/vol) agarose gel electrophoresis.

2.4. PCR amplification of DNA fragments from Bcc strains

The DNA or the cell lysate of Bcc bacterial isolates was a gift from Anna Meyer Children's Hospital (Department of Paediatrics, University of Florence; Division of Paediatrics, Infectious Diseases, Cystic Fibrosis, V.le Pieraccini 24, I-50139 Florence, Italy), ENEA (ENEA-CRE-CASACCIA-Department of Biotechnologies, Agroindustry, Protection of Health-Plant Genetics and Genomics Section, Via Anguillarese 301 S. Maria di Galeria, 00123-Rome, Italy), and G. Manno (Department of Paediatrics-Infectious Diseases Research and Diagnosis Laboratory-Cystic Fibrosis Center, University of Genoa, G. Gaslini Children's Hospital Largo G. Gaslini 5, 16147 Genoa, Italy) Laboratories. Cell lysates were prepared by lysing of 2–3 colonies grown overnight on LB, according to Vandamme et al. (2002). PCR amplification of *his* DNA fragments was performed in a 25 μl reaction mixture containing 2 μl of cell lysate, 1.0 U of PolyTaq DNA polymerase (Polymed), 250 μM of each deoxynucleoside triphosphate, 20 pmol of each primer, 1.5 mM $MgCl_2$, and 1× PCR buffer. A primary denaturation treatment of 2 min at 95 °C was performed and amplification of *his* genes was carried out for 30 cycles consisting of 30 s at 95 °C, 45 s at 56 °C and 60 s at 72 °C, with a final extension of 10 min at 72 °C. Thermal cycling was performed with a gene Amp PCR System 9700 instrument (Applied Biosystems). The primers used in this work are listed in Table 3.

2.5. Transformation

Induction of competence and transformation of *E. coli* cells with plasmid DNAs were carried out as described previously (Hanahan, 1983).

2.6. Plasmid extraction

Plasmid DNA was extracted from *E. coli* cells using the "High pure plasmid isolation" Kit (Roche), according to the manufacturer's instructions.

Table 3
Oligonucleotides used in this work as primers in PCR amplification and/or sequencing reactions.

Primer	Sequence (5′–3′)
for00	CGCTGTCGCGCGTCGTGA
for01	GA(CT)CGCGTGTTCACGAT
for02	CGCCGCGACGTGCTCGC
for03	CGTGAGTAAAATGCATCC
for04	TCGGTGTTCTGCGTGAT
for05	TCGGCGAGCAGATGCTG
for06	TACGTGATCATCGGCAC
for07	CGCGATCGACGCGAAGC
for08	GACGTGCTGATGTTCGC
rev01	GTAGTTGATGATGATCG
rev02	TCGGCCATCTTGCGCGC
rev03	CGCAGATCACGCCTTCG
rev04	TGCGGCACCTTGAAGCG
rev05	GCGATCTGGTCGAGCAT
rev06	ACATCACGAAGCCCGGC
rev07	GCGAGAACAGGTCCATC
HisB-35no	GGATCCCGACGATCAATCCCCATTTACATC
HisB-10no	GGATCCGCCATGCGTGT
HisH_IN_AF	GGATCCTCTGTATCGCAACTTCGTACACT
IN_marCF	GGATCCTCGACCAAGGGCAGCCTCT
HisA_IN_HisF	GGATCCGAGCTCGACGACGCCT
RevHisF_univ	GGATCCTCACAGCCTCACCGGGAT

2.7. DNA sequencing and accession numbers

For sequencing, the PCR products were purified using the "Mini elute gel extraction" purification kit (Qiagen) according to the manufacturer's instructions. The nucleotide sequence of an about 4800 bp DNA fragment was determined on both strands according to Sanger et al. (1977), using an Applied Biosystems Big Dye® Terminator sequencing kit version 3.1, according to the supplier's instructions. Thermal cycling was performed with a gene Amp PCR System 9700 instrument (Applied Biosystems). Each sequence was submitted to GenBank and was assigned the accession number reported in Table 1.

2.8. RNA extraction and RT-PCR

The Bcc strain LMG16670 was grown either in LB or MMD without histidine. Two volumes of RNAprotect Bacteria Reagent (Qiagen) were added to cell suspensions collected from cultures in log, at the end of log, and in stationary phase. Samples were incubated at room temperature for 5 min and centrifuged at 5000×*g* for 10 min. Bacterial pellets were stored at −20 °C until RNA extraction. RNA was extracted with the FastRNA® Kit-Blu (Bio 101) and finally resuspended in Nuclease-free water (Promega). Residual DNA was removed by digestion with RQ1 RNase-free DNase (Promega). The efficacy of DNase digestion was determined by direct PCR (with no reverse transcription step) and DNase-digested samples yielding no PCR amplification products were considered free of DNA contamination. RNA quality was checked by electrophoresis in 1.5% (wt/vol) ethidium-bromide stained agarose gel and RNA templates were quantified spectrophotometrically at 260 nm using BioPhotometer (Eppendorf). For cDNA synthesis 1 μg of RNA, a mixture of random hexadeoxynucleotides (0.5 μg/reaction, Promega) and ImProm-II™ Reverse Transcriptase System (Promega) was used. All mixtures were assembled at 0 °C using Nuclease-free water (Promega) and reverse transcription (RT) reaction was performed as recommended by the manufacturer's protocol. Double stranded *his* cDNA was generated using 1 μl of the RT product in subsequent PCR reactions with the primers listed in Table 3. Amplification reactions were carried out as previously described and RT-PCR products were separated by electrophoresis. Images of 0.8% (wt/vol) agarose gels were captured using the UV illuminator ChemiDoc apparatus (Bio-Rad).

2.9. DNA and protein sequence analysis

BLAST probing of the protein and nucleotide databases was performed with the BLASTp and BLASTn programs (Altschul et al., 1997) using default parameters. Amino acid sequences were aligned using the ClustalW program (Thompson et al., 1994). Coding sequences were aligned with the standalone version of RevTrans software (Wernersson and Pedersen, 2003), which performs a nucleotide alignment using the corresponding alignment of the peptide sequences as a scaffold.

The multialignments were visually corrected and used to build phylogenetic trees using the genetic distance-based neighbour-joining algorithm of MEGA 4.0 (Tamura et al., 2007) and the complete deletion parameter. The Kimura 2-parameter model was used to allow considering differences in transition and transversion rates and to correct for multiple hits; the model also assumes that the four nucleotide frequencies are the same and that rates of substitution do not vary among sites.

The degree of conservation of genes and intergenic regions was calculated with the maximum composite method implemented in Mega version 4.0 (Tamura et al., 2007). The degree of conservation was also estimated using entropy plot calculation implemented in the BioEdit Package (Hall, 1999). Bendability profiles were obtained using

DNase I sensibility patterns estimated by Gabrielian et al. (1997) and a sliding window approach that allowed assigning a score to each overlapping trinucleotide. Bendability profiles for each of the 53 sequences were then smoothed using Lowess local regression with a 20 nucleotides window. Profiles were then averaged and the mean and standard deviations plotted.

2.10. In silico analysis of microarray data

The availability of *Burkholderia cenocepacia* microarray experiments was checked in all public transcriptomic repositories (i.e. GEO, ArraryExpresses, etc.). On January 10, 2009 only one *B. cenocepacia* microarray dataset (E-MEXP-1261, Drevinek et al., 2008) was available in the databases. From such experiment only the hybridizations of control vs. control samples (E-MEXP-1261-raw-data-1556955930 and E-MEXP-1261-raw-data-1556955925) were retrieved and analyzed. The two hybridizations are two dye swapped slides of 2 bulk of reference samples growing in a basal salts medium, containing 14.3 mM glucose and 0.05% casamino acids (Drevinek et al., 2008) and collected in the mid-log growing phase. The available raw microarray data representing the expression of *his* genes in the absence of growing perturbation (i.e. abiotic stresses, nutrients depletion) were loaded into the R statistical package (http://www.r-project.org/) and the LIMMA package (Smyth, 2004) (http://bioinf.wehi.edu.au/limma/) was used to perform background correction and normalization. Standard functions from the LIMMA package were used to produce pre- and post-normalized quality control plots to ensure data quality. Background subtraction was achieved by removing the local median background intensity from the spot foreground median intensity. Background corrected data were then normalized using a print-tip LOWESS normalization. Cross-slide normalization was also applied to ensure consistency of scale. Normalized data were then filtered based on A value $[(\log_2 R + \log_2 G)/2]$ of negative control spots to remove spots with low intensities in both channels (the threshold value for A was set to 7.64, equating to a raw intensity of 200). To explore *his* biosynthesis pathway genes expression levels in *B. cenocepacia* J2315, the average A values of genes belonging to the heterogeneous *his* operon and housekeeping genes subtracted of the (hybridization negative) control spots A level were plotted following operon organization.

Since no annotation of microarray probes is available, locus tags were retrieved from the *B. cenocepacia* J2315 genome: *murA* (UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase, BCAL0310); *hisG* (ATP phosphoribosyltransferase, BCAL0311); *hisD* (Histidinol dehydrogenase, BCAL0312); *hisC* (Histidinol-phosphate aminotransferase, BCAL0313); *hisB* (Imidazoleglycerol-phosphate dehydratase, BCAL0314); *marC* (MarC-family integral membrane protein, BCAL0315); *hisH* (Imidazole glycerol phosphate synthase subunit HisH, BCAL0316); *hisA* (Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase, BCAL0317); *hisF* (Imidazole glycerol phosphate synthase subunit HisF, BCAL0318); *hisI* (Phosphoribosyl-AMP cyclohydrolase, BCAL0319); *hisE* (Phosphoribosyl-ATP pyrophosphatase, BCAL0320); putative membrane protein (BCAL0321); *hit* (Putative uncharacterized protein, BCAL0322); *tatA* (Sec-independent protein translocase protein TatA, BCAL0323); *tatB* (Sec-independent protein translocase protein TatB, BCAL0324); *tatC* (Sec-independent protein translocase protein TatC, BCAL0325); *hisZ* (Putative ATP phosphoribosyltransferase, BCAL1874); *16S* (BCAS r0743e); *gapA* (Glyceraldehyde 3-phosphate dehydrogenase 1, BCAL3388); *gyrA* (DNA Gyrase, subunit A, BCAL2957); *RpoD* (RNA polymerase sigma factor RpoD, BCAM0918); *pgi* (Glucose-6-phosphate isomerase, BCAL1990); *recA* (Putative recombinase A, BCAL0953).

To support the actual expression of *his* operons genes, a quantile analysis on A values was performed comparing the lowest (*hisD*), the highest (*hisI*) and the median (of the full operon) A value of *his* transcripts against the full list of the microarray A values.

3. Results

3.1. Amplification and sequencing of his biosynthetic core from 40 Bcc strains

On December 1, 2007 the complete sequence of the genome from 13 *Burkholderia* strains was available. The nucleotide sequence of each of the entire *his* cluster from these strains was retrieved and aligned with the program ClustalW (Thompson et al., 1994); this allowed to identify highly conserved regions enabling the design of a set of primers (Table 2) to amplify the *his* core from the genome of Bcc strains. The expected size of different *his* amplicons when using each primer set is shown in Table 4.

All the possible primer forward–reverse combinations were firstly tested on the DNA of two strains, i.e. *B. vietnamiensis* TVV75 and *B. cenocepacia* LMG16654 that were chosen because they belong to distantly related Bcc species and, in principle, their sequences should exhibit a degree of sequence divergence higher than that showed by more closely related strains. Thus, it is plausible that if a given primer set works on these two DNAs, it should also work with DNA from more closely related strains. The best results were obtained by using the primer sets For03-Rev04, For03-Rev05, For06-R02, For06-Rev03, and For07-Rev01. Amplification carried out using the other primer sets gave unspecific amplicons or did not yield amplicons (not shown). Since the primer sets For03-Rev04, For05-Rev02 and For7-Rev01 allowed the amplification of the entire *his* "core", they were used to amplify it from all the strains of the experimental panel consisting of 40 strains representative of the Bcc species with either environmental or clinical origin (Table 1). Amplicons of the expected size were obtained from each of the 40 Bcc strains (not shown).

The nucleotide sequence of all the amplicons (about 4800 bp) obtained was determined and analyzed. Each nucleotide sequence and the amino acid sequence of the putative encoded proteins was used as a query in a BLAST (Altschul et al., 1997) probing of nucleotide and protein database. Data obtained (not shown) revealed that each sequence utilized as a query retrieved at the lowest E-values sequences corresponding to the *his* regions from the thirteen *Burkholderia* genomes available in databases. The ClustalW multialignment (Thompson et al., 1994) of the 53 *his* "core" sequences is reported in the Additional file 1. This analysis revealed that the 40 sequences obtained contained the 3′ end of *hisC*, *hisB*, *marC*, *hisH*, *hisA*, *hisF*, *hisI*, *hisE* and the 5′ terminal of a gene coding for a membrane protein.

3.2. Structure of his "core" genes in Burkholderia strains

It has been previously shown that three of the histidine biosynthetic genes (*hisB*, *hisA*, and *hisF*) belonging to the *his* "core", underwent different molecular rearrangements, i.e. elongation, duplication, and/or fusion events in different phylogenetic lineages. For this reason, the structure of these *his* genes from all the 53 *Burkholderia* strains was analyzed.

It has been shown previously (Fani et al., 1994, 1997; Fani and Fondi, 2009) that *hisA* and *hisF* genes have a common ancestry and

Table 4
Size of the different *his* amplicons (in base pairs) obtained from the genomic DNA of Bcc strains.

Primer	For01	For02	For03	For04	For05	For06	For07	For08	
Rev01					3193	2337	1370	889	Length of amplicon
Rev02					1909	1053	86		
Rev03					1192	336			
Rev04	3613	2730	1778	833	99				
Rev05	2094	1211	269						
Rev06	1170	287							
Rev07	197								

are the outcome of a cascade of at least two duplication events, involving an ancestral gene half the size of the present-day ones. This gene underwent a first elongation event giving raise to the ancestor of *hisA*. This *hisA* ancestor gene in turn duplicated generating *hisF*. The analysis of the amino acid sequence (Additional files 2 and 3) of all the 53 *Burkholderia hisA* and *hisF* sequences supported the model proposed for their origin and evolution (Fani et al., 1994).

The sixth and the eighth step of histidine biosynthesis are catalyzed by Histidinol-phosphate phosphatase (EC 3.1.3.15) (HOLPase) and IGP dehydratase (EC 4.2.1.19) (IGPase), respectively (Alifano et al., 1996). In *E. coli* and in other γ-proteobacteria the two activities are coded for by a bifunctional gene (*hisNB*), formed by two domains, i.e. a proximal one (*hisN*) encoding the phosphatase moiety, and a distal one (*hisB*) encoding the dehydratase activity (Brilli and Fani, 2004a). The ClustalW multialignment (Additional file 4) of the 53 *Burkholderia* sequences with a set of bifunctional and monofunctional counterparts revealed that all of the 53 *hisB* genes analyzed corresponded to the 3′ domain of the *E. coli* bifunctional gene, coding for IGPase.

3.3. Organization of his "core" genes in Burkholderia

The organization of histidine biosynthetic "core" genes from the 53 *Burkholderia* strains analyzed is schematically reported in Fig. 1 and Table 5. The analysis revealed a high degree of conservation of gene organization and the presence of an "alien" gene between *hisB* and *hisH* apparently not related to histidine biosynthesis.

The degree of length conservation of each of the five putative genes (*hisB*, *marC*, *hisH*, *hisA*, and *hisF*) and relative intergenic region

Table 5
Length of *hisB*, *marC*, *hisH*, *hisA*, *hisF* and relative intergenic sequences from 53 *Burkholderia* strains.

Species	Strain	Gv	Length of sequence (bp)									
			hisC-hisB	hisB	hisB-marC	marC	marC-hisH	hisH	hisH-hisA	hisA	hisA-hisF	hisF
B. mallei	ATCC 23344		63	588	54	621	−4	641	73	756	83	774
	NCTC 10229											
	NCTC 10247											
	SAVP1											
B. pseudomallei	1710b											
	K96243											
	1106a											
	668											
B. thailandensis	E264		62									
B. cepacia	AMMD		65		61				97		106	
B. cenocepacia	AU 1054				60							
	HI2424											
B. vietnamensis	G4		62		62						110	
B. cepacia	FCF2	I	65		60				106		106	
	LMG2161				65				108			
B. multivorans	FCF7	II			60				128		128	
	LMG18822				57							
	LMG13010				61							
	LMG17588											
B. cenocepacia	FCF13	IIIA			60				106		106	
	FCF15											
	FCF16											
	FCF17								107			
	FCF19	IIIB							106			
	FCF20										107	
	FCF22										106	
	FCF23											768
	FCF24											774
	FCF25											
	FCF27											
	FCF28											
	FCF30											
	FCF31											
	LMG19230	IIIC			61				105		105	
	LMG19240				62							
	FCF32	IIID			60				106		106	
	FCF34											
	FCF37											
	FCF38											
	FCF39											
B. stabilis	FCF41	IV			61							
	LMG14294											
B. vietnamensis	TVV75	V	62		62						110	
B. dolosa	LMG18941	VI	65		61				109			
	LMG18942											
B. ambifaria	MC17	VII							106		106	
	LMG19467											
B. anthina	LMG16670	VIII			59				121		121	
	LMG20980								120		122	
B. pyrrocinia	FCF43	IX			60				107		105	
	FCF44											
	ATCC15958											
	MVPC1/26											

Abbreviations: Gv, genomovar; bp, base pair.

was calculated. Data obtained, reported in Table 5, revealed that the length of each of the five genes is extremely conserved through the genus *Burkholderia*, the only exception being the FCF 23 *hisF* gene, which is 6-bp shorter than the other 52 orthologous genes. Apart from *marC* and *hisH* (showing a 4 bp sequence overlapping), all the other genes are separated by four intergenic regions, two of which (those between *hisC* and *hisB*, and *hisB* and *marC*) exhibiting a degree of length conservation higher than that found in the others. To check the possibility whether the four intergenic regions had a role in regulating *his* genes expression or not, we analyzed them in all the 53 nucleotide sequences by a combination of different bioinformatic tools. Firstly, the 53 sequences were multialigned by ClustalW; data obtained revealed that only the *hisC-hisB* region was highly conserved, especially within the Bcc strains (Fig. 2), where 47 out of 66 sites are conserved. Besides, the 40 nt segment located at the 3′ end of this region is AT rich (GC content of 39%), very rare in Bcc genomes showing a high (70%) GC content.

Moreover, the finding that the *hisC-hisB* intergenic region exhibited a number of base substitutions per site much lower than that found in the other intergenic regions and similar or lower to that of coding regions (Table 6), suggested that a functional constraint may act on it. A further analysis revealed that the entropy value of the *hisC-hisB* region was much lower than that exhibited by the other intergenic regions (Fig. 3). Last, to assess the structural properties of the intergenic regions, we checked the occurrence of curved regions, a property that is usually found associated with prokaryotic promoters. Curvature can be induced by the binding of a protein, as often happens when a transcription factor or the RNA polymerase binds to the promoter (Bultrini and Pizzi, 2006; Kanhere and Bansal, 2005). The *bendability* is a way of quantifying the ease with which a DNA molecule can curve in any direction. Since there are data indicating the wrapping of the promoter DNA around the polymerase at the beginning of transcription (Wong et al., 2008), structural properties of DNA have been used for promoter prediction (Kanhere and Bansal,

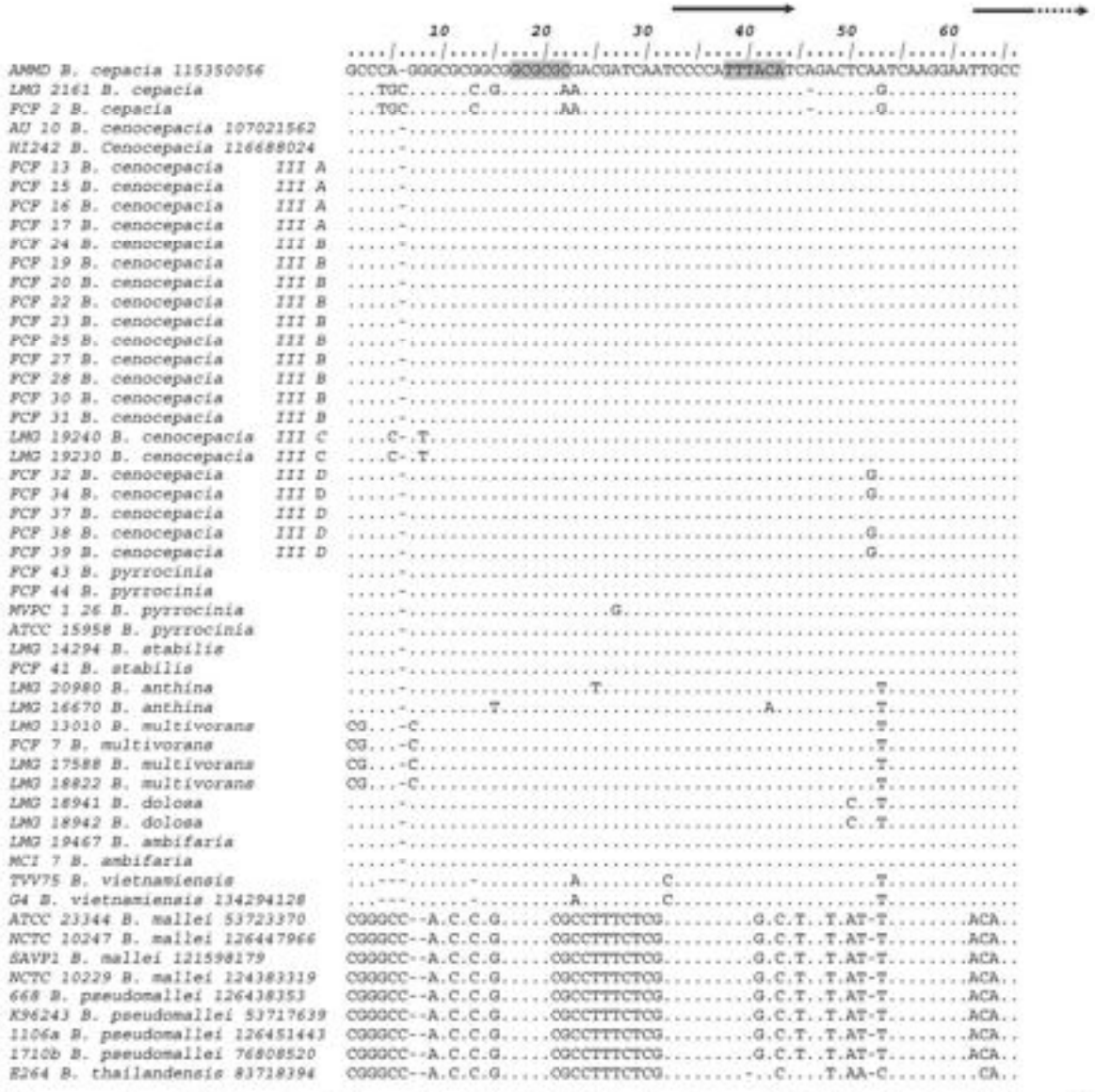


Fig. 2. Multiple sequence alignment of the intergenic region between *hisC* and *hisB* in 53 *Burkholderia* strains. The two regions identified as the −35 and −10 box are highlighted in grey; the two arrows represent the forward primers used for the amplification of a DNA fragment lacking either the putative −35 region, or both the −35 and −10 regions of the putative *hisBp*.

Table 6
Number of base substitutions per site in both coding and non-coding sequences of the Bcc *his* "core".

Region		Base substitutions per site	
		All strains	Bcc strains
Intergenic	*hisC-hisB*	0.2300	**0.0324**
Coding	*hisB*	0.0519	0.0393
Intergenic	*hisB-marC*	0.2370	**0.1300**
Coding	*marC*	0.0700	0.0386
Coding	*hisH*	0.0646	0.0390
Intergenic	*hisH-hisA*	0.1790	**0.1240**
Coding	*hisA*	0.0570	0.0570
Intergenic	*hisA-hisF*	0.1930	**0.0635**
Coding	*hisF*	0.0549	0.0373

Analyses were performed using the Maximum Composite Likelihood method in MEGA4 (Tamura et al., 2007). All positions containing gaps and missing data were eliminated from the dataset (complete deletion option).

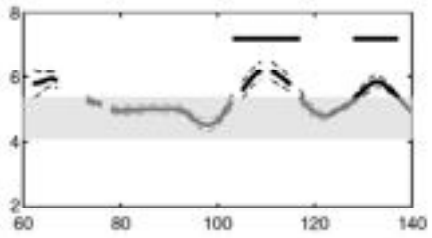


Fig. 4. Bendability profile of the *hisC-hisB* intergenic region. The two peaks marked by the black bars correspond to highly bendable regions. Discontinuities in the thick line correspond to gapped regions in the alignment. The dashed line is the standard deviation of the bendability profile over the 53 sequences. The grey rectangle corresponds to average bendability over the entire operon plus/minus 2 standard deviations. The scale on the y-axis is in arbitrary units, the scale on the x-axis is in nucleotides, corresponding to the location of the region within the *his* operon (see Additional file 1). For the procedure used the curve moves about 20 nucleotides downstream, so that the peaks approximately correspond to the region around nucleotides 100 and 120, respectively.

2005). To determine regions with high bendability we take advantage of the average and standard deviation of the bendability for 100 bootstrap replicates of our dataset (the grey area corresponds to average plus/minus 2 standard deviations); peaks in the plot corresponding to highly bendable DNA regions can be identified. Two bendable stretches closely upstream of *hisB* were observed (Fig. 4), whereas no such structures were found in the other three intergenic sequences (*hisB-marC*, *hisH-hisA*, and *hisA-hisF*).

3.4. Complementation analysis of *E. coli his* mutations and identification of a transcription promoter region upstream of *hisB*

The whole body of data reported in the previous section is in favour of a possible regulatory role (i.e. as a transcriptional promoter) of the *hisC-hisB* intergenic sequence. This might be functionally tested by complementing mutations falling in the *his* genes located downstream from the promoter. However, no Bcc *hisB* mutant is available. Therefore, by assuming that the putative transcription promoter might work in a heterologous host, we used *E. coli* to assay the functionality of the putative promoter. The strategy adopted is based on the idea that if this promoter exists, once that the entire *his* core is cloned in both the orientations in a plasmid vector downstream from the *lac* promoter, it should restore the His^+ phenotype in *E. coli* His^- mutants, independently from its orientation in respect to the *lac* promoter. To check this hypothesis, the PCR amplicons containing the entire histidine biosynthetic "core" and spanning from the 3′ end of *hisC* to the 5′ terminal of *hisF* obtained from the *B. anthina* strain LMG16670 were cloned in the pGEM-T easy vector downstream from the *lac* promoter (*plac*). Since the analysis of the AT-rich region revealed the presence of two sequences that might represent putative −35 and −10 regions (Fig. 4), amplified fragments carrying a progressive deletion of the intergenic region were obtained and cloned into the pGEM-T vector. The ligation mixtures were used to transform competent cells of *E. coli* DH5α. Six recombinant plasmids (pBC1-pBC6) containing the insert in one of the two possible orientations were obtained and introduced by transformation into competent cells of the three *E. coli* His^- mutants FB182, FB184, and FB251. Twenty transformants of each strain were then checked for their ability to grow in MMD either in the presence or without histidine. Data obtained revealed that the His^+ phenotype was restored in all the *E. coli* strains harboring plasmids pBC5 or pBC6 (Fig. 5); all the other strains were unable to grow in MMD without histidine.

This indicated that the transcription starting from *hisBp* proceeds through *marC*, *hisH*, *hisA*, and *hisF*. Similar experiments carried out with DNA fragments including one or more of the three intergenic regions *hisB-marC*, *hisH-hisA*, and *hisA-hisF* (see Fig. 6) showed that none of the recombinant plasmids used were able to restore the His^+ phenotype of strains FB182 and/or FB184, suggesting that these regions do not contain a transcription promoter recognized by the

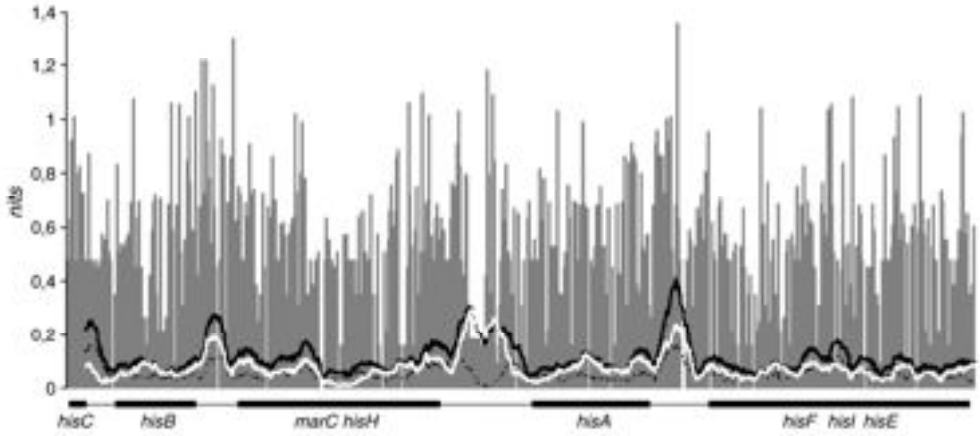


Fig. 3. Entropy plot (thin grey bars) and moving average (black curved line) based on the multiple alignment of the histidine operon fragments of 53 different *Burkholderia* strains. The other curved lines represent: i) the moving average when the same analysis was performed only on the Bcc strains (white line) and ii) the moving average when the Bcc strains were excluded from the analysis (dashed line). Y-axis refers to entropy values, whereas X-axis reports the position along the histidine operon "core".

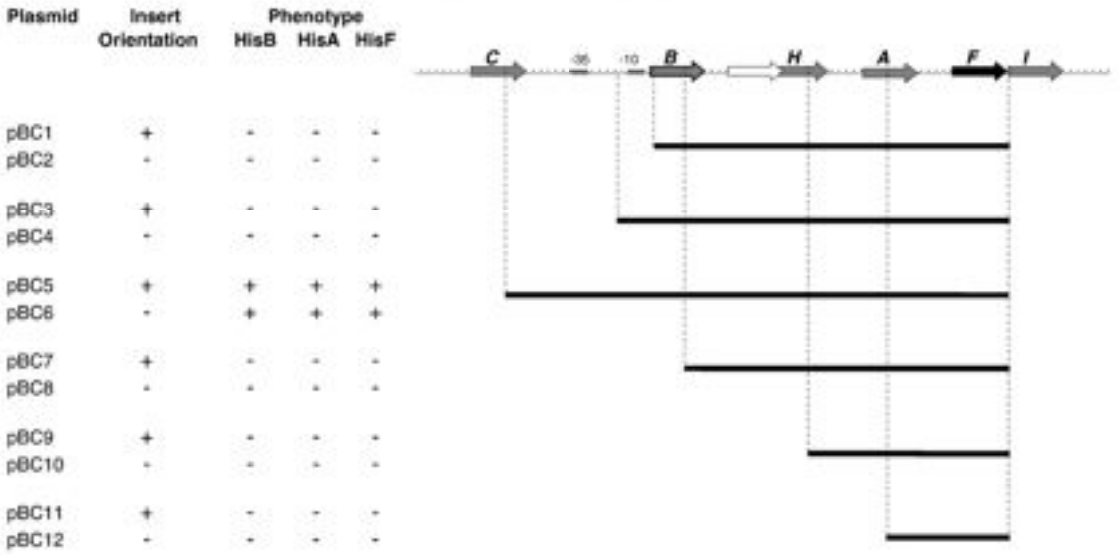


Fig. 5. Complementation analysis of *E. coli hisA*, *hisB*, and *hisF* mutations harboring recombinant plasmids carrying DNA fragment of different length amplified from *B. anthina* LMG16670 genome and cloned in pGEM-T easy vector in the two possible orientations in respect to the *lac* promoter. The black solid bars represent the PCR amplified fragments. The fragments cloned in i) pBC1 and pBC2, ii) pBC3 and pBC4, iii) pBC5 and pBC6, iv) pBC7, pBC8, v) pBC9, pBC10, and vi) pBC11, pBC12 were PCR amplified using as forward primers i) hisB-35no, ii) hisB-10no, iii) for00, iv) IN_marC_F, v) HisH_in_AF, and vi) HisA_in_HisF, respectively, and as reverse primer the oligonucleotide RevHisF_univ.

E. coli transcription apparatus. However, we cannot *a priori* exclude the possibility of the existence of a transcription signal not recognized by the *E. coli* apparatus, but working in *Burkholderia* cells. The same results were obtained with the DNA from *B. multivorans* 17588 strain exhibiting some differences in the intergenic region(s) (Fig. 4).

3.5. Transcriptional analysis

In order to shed some light on the regulation of *his* operon transcription in bacteria belonging to Bcc, a preliminary analysis of *his* core transcripts was carried out. To this purpose, total RNA was extracted from the *B. anthina* strain LMG16670 grown either in LB or minimal medium without histidine to check the influence of presence/absence of histidine on *his* operon transcription. Total RNA was extracted from cultures in log or in stationary phase and two primer sets (Fw04-Rev04 and Fw05-Rev04), allowing the semi-nested RT-PCR amplification of the region downstream from *hisB* promoter, were used as described in Materials and methods. Data obtained revealed (Fig. 6) the presence of amplicons of the expected size in all samples examined, grown either on LB or MMD, and collected from cultures either in log or in stationary phase.

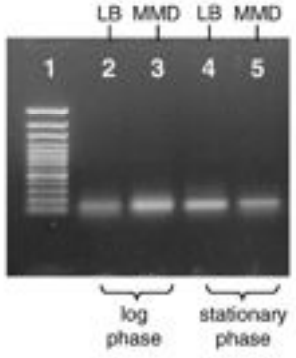


Fig. 6. Agarose gel electrophoresis of amplicons obtained by nested RT-PCR using primers For05-Rev04 performed on the mRNA extracted from LMG 16670 cultures grown either in LB or MMD without histidine.

3.6. Microarray data analysis

Expression level of *his* operon genes was investigated by analyzing the publicly available microarray data. Only one experiment (Drevinek et al., 2008) was available for Bcc species at the time of the present analysis. In Fig. 7 the averaged A (see Materials and methods for details) values of *his* operon and housekeeping genes for untreated bulks of control samples (for which no differential expression is expected, since the samples do not differ for any growing condition) of *B. cenocepacia* J2315 growing in a complete medium are reported. The analysis showed that all the genes belonging to the heterogeneous operon as well as *hisZ* (a gene involved in the regulation of HisG activity, see discussion) are expressed under normal growing conditions (absence of growing perturbations such as abiotic stresses or nutrients depletion). The expression levels of *murA* and *hisG* and for the genes going from *hisH* to *tatB* are comparable or higher than those showed by housekeeping genes like *gyrA*, *recA* and *pgi*. Other six genes, that is *hisD*, *hisC*, *hisB*, *marC*, *tarC* and *hisZ* (the latter located outside the operon), showed expression values ranging from 2.5 (for *hisD*) to 7.5 times (for *hisB*) the (log2 untransformed) intensity threshold value. The quantile analysis showed that the median *his* operons genes A value belong to the 84th percentile (62th and 94th for *hisD* and *hisI*, respectively).

3.7. Phylogenetic analysis

A phylogenetic analysis using either the nucleotide sequence of each of the *his* core genes or the amino acid sequence of the product they code for was performed and the trees obtained are shown in Fig. 8 and Additional file 5. Concerning the trees obtained with the multialignments of the amino acid sequences, a strong bootstrap support for the monophyly of all the sequences embedded in the *his* core and retrieved from the organisms belonging to the Bcc complex was observed. In fact, with the only exception of the HisA phylogenetic tree (where the sequences retrieved from the *B. multivorans* and *B. dolosa* strains are placed outside the main cluster embedding the other Bcc sequences) in all the other trees the monophyly of the Bcc His sequences with a bootstrap support ranging between 91% and 100% was observed.

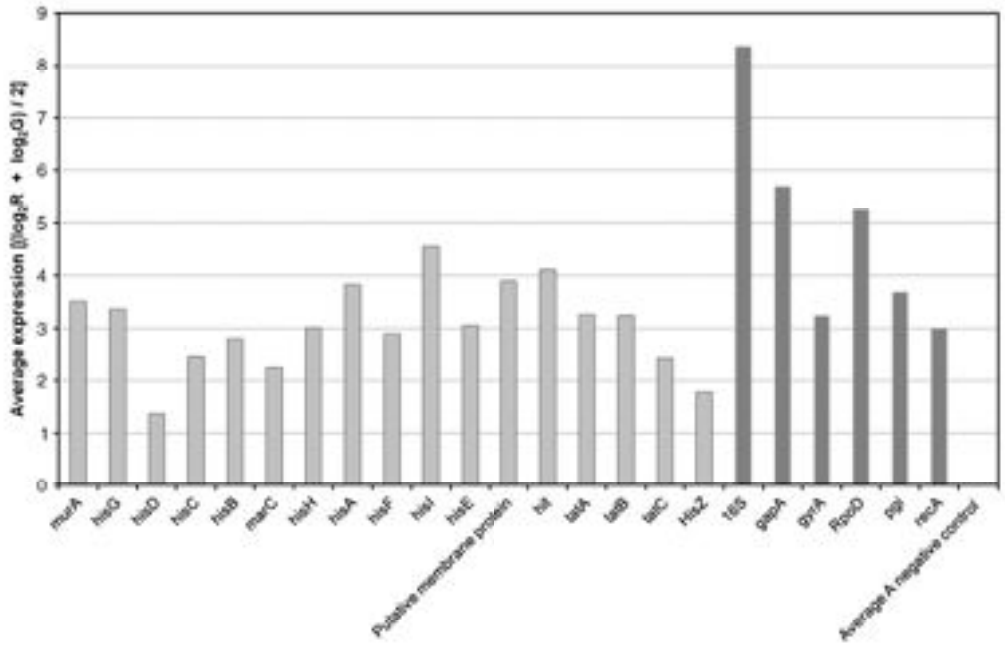


Fig. 7. In silico microarray analysis. Microarray (from experiment E-MEXP-1261 [Drevinek et al., 2008]). Average expression [$A = (\log_2 R + \log_2 G) / 2$] values of *his* operons (light grey) and housekeeping genes (dark grey) minus the hybridization negative spot A value. Genes are annotated following the *B. cenocepacia* J2315 genome locus tagging (see Materials and methods).

These findings were also confirmed by the analysis of the phylogenetic trees constructed with the corresponding nucleotide sequences. In this case, we obtained 100% bootstrap support for the monophyly of the Bcc sequences in all trees, in each of which the Bcc sequences are separated from the other *Burkholderia* ones. Moreover, concerning the Bcc sequences, almost in each tree all the strains belonging to the same species are grouped together in monophyletic clusters, even though the branching order was not always the same and showed low bootstrap values, as it might be expected for very closely related strains. This is probably due to the high degree of similarity of the *his* sequences in these closely related (micro) organisms.

4. Discussion

It is known that bacteria belonging to the genus *Burkholderia* harbor two to three different chromosomes and some of them are among the largest genome-sized and most versatile bacteria known. Besides, these genomes harbor a relevant number of genes coding for transposases, integrases, and resolvases, suggesting that they might frequently undergo DNA rearrangements that, in turn, might alter their gene structure and/or organization. In spite of this possibility, data reported in this work revealed a high degree of conservation of structure and organization of the *his* core genes in all the 53 *Burkholderia* strains analyzed. In our opinion, this is related to the function performed by the *his* core. Since it represents a cross-point interconnecting histidine biosynthesis to nitrogen metabolism and the *de novo* synthesis of purines, it is plausible that the balanced expression of the *hisBHAF* might be crucial for cell metabolism. Accordingly, it has been demonstrated that the unbalance of their expression (i.e. hyperexpression of *hisH*, *hisF*, and/or *hisA*) leads to a drastic alteration of cell form (i.e. filamentation) and other metabolic alterations. Indeed it has been suggested that elevated levels of *hisH* and *hisF* gene products induce filamentation by interfering somehow with cell wall synthesis (see Alifano et al., 1996 for a review). These two genes code for a glutamine aminotransferase and a cyclase, respectively, that must interact to give a heterodimeric active IGP synthase, which catalyzes the central step of histidine biosynthesis (Fig. 1). Elevated levels of IGP synthase cause inhibition of cell division by themselves. Filamentation, as well as the other pleiotropic effects associated with *his* overexpression, was shown to occur in *E. coli* and *S. typhimurium* strains because of interrupting the carbon flow through the histidine and purine pathways in *his pur* double mutants (Frandsen and D'Ari, 1993). It has been also suggested (Fani et al., 2005) that proteins encoded by the four genes *hisBHAF* might interact to form a metabolon (Brilli and Fani, 2004b; Srere, 1987), that is a complex constituted by (transiently) interacting proteins that can facilitate the catalysis of reactions by producing intermediates in the proximity of enzymes that act upon them. This idea is in agreement with the notion that genes coding for proteins that have to interact to form an active complex very often are clustered in conserved operons (Tamames, 2001). If this is true, either a change in the concentration of one (or more) of the interacting proteins or their non-correct co-regulation and co-expression might interfere with the correct assembly of the complex. Since it is well accepted that one of the forces driving the operon assembly is the possibility to co-regulate and co-express genes involved in the same metabolic route (reviewed in Fani et al., 2005), this might explain the conservation of the *his* core. Besides, the high degree of sequence conservation exhibited by the *his* "core" genes is in agreement with the idea that interacting proteins have more functional constraints than stand-alone ones. Hence, both structural and functional constraints appear to be responsible for the high degree of conservation of *his* core. This hypothesis is supported by the phylogenetic analysis that showed the monophyly of *his* core genes and suggested that they very likely did not undergo horizontal gene transfer events, at least in recent times, thus preventing possible DNA rearrangements that may alter the *his* core organization.

Concerning the expression of *his* core genes, the *in silico* prediction of a transcription promoter located closely upstream *hisB* was confirmed by the complementation analysis of the *E. coli* mutants, which, in turn, revealed that the transcription starting from this promoter allows the expression of the downstream genes. It is interesting that, in spite of the phylogenetic and ecological distance

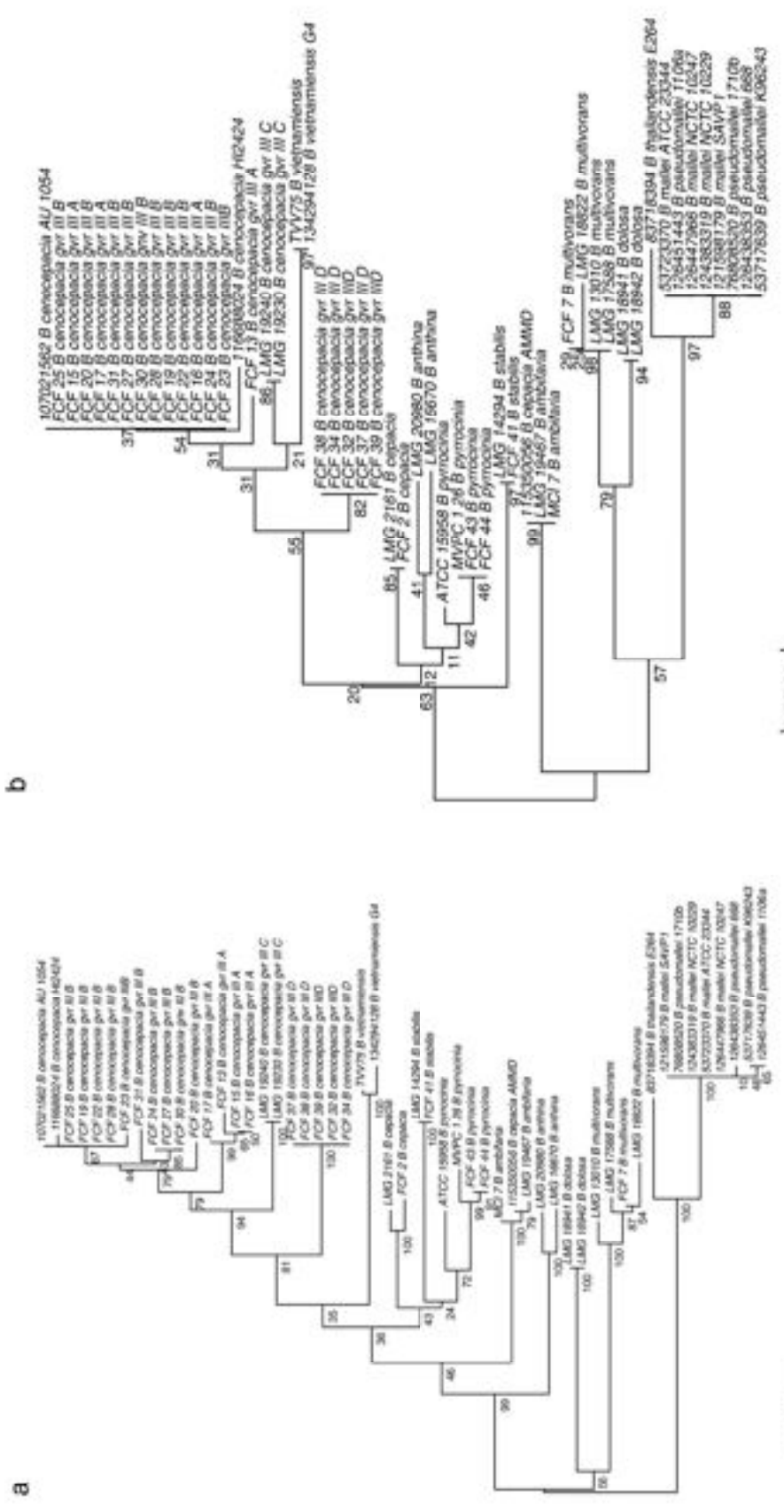


Fig. 5. Phylogenetic trees constructed using either the nucleotide sequence of *hisA* gene (a) or the amino acid sequence of the product it codes for (b). The bootstrap values (1000 replicates) are shown next to each branch (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling, 1965) and are in the units of the number of amino acid substitutions per site.

existing between *E. coli* and *Burkholderia*, the *hisBp* is recognized by the transcriptional apparatus of the former, demonstrating that such a regulatory signal has been maintained during evolution at different taxonomical levels (strain, species, genus). This is in agreement with the prediction of the piece-wise model for the operon construction (Fani et al., 2005, 2006), which is supported by the absence of promoters immediately upstream of *hisA* and *hisF*. Another (preliminary) indication supporting the piece-wise model comes from the analysis of RT-PCR and microarray data that suggested that transcription of the *his* "core" genes might be constitutive, even though more quantitative methods to measure *his* transcripts level under different growth conditions to induce severe histidine limitations will be needed to fully demonstrate the constitutive expression of *his* operon. If it is actually so, this raises the following question: why should a metabolic pathway with a considerable energy-cost (41 ATP molecules for each histidine molecule, Alifano et al., 1996) constitutively transcribed? And hence, how is the expression of this heterogeneous operon controlled?

The answer to the first question might rely on the presence within the operon of "alien" genes not involved in histidine biosynthesis and whose expression cannot be under the control of histidine. Regarding these genes, if some of them can be the relics of the ancestral events that led to the assembly of heterogeneous operons, others can be the outcome of an introgression event that occurred in an already assembled operon, an event facilitated by the presence of non-coding regions located between genes of the operon (Price et al., 2006). Intergenic sequences are present within the Bcc heterogeneous *his* operon and have very likely permitted the introgression of an "alien" gene (*marC*) between *hisB* and *hisH* in the genome of the common ancestor of *Ralstonia* and *Burkholderia* (Fondi et al. manuscript in preparation). This event should have been followed very quickly by its overlapping with *hisH*. As discussed previously (Fukuda et al., 2003) some of the genes overlapping play functional role(s), such as translational coupling, that provides a mechanism to ensure coordinate and equimolar synthesis of proteins coded for by polycistronic messenger RNA (Das and Yanofsky, 1989; Inokuchi et al., 2000; Oppenheim and Yanofsky, 1980). Gene overlapping requires the adjustment of the nucleotide sequence of both genes to avoid any alteration in the conformation of the proteins they code for and then their functionality. This is particularly true for the product of *hisH* (see above). Hence, it is plausible that the placement of *marC* within the *his* operon and its overlapping with *hisH* might have given an (evolutionary) advantage on (at least) *Ralstonia* and *Burkholderia* cells. It is not still clear whether this might be related to the lifestyle of these microorganisms. The involvement of *marC* in histidine biosynthesis is rather unlikely since in all histidine-synthesizing organisms, the biosynthesis of the amino acid proceeds through the same enzymatic steps. The biological significance of the presence of *marC* within the *his core* is still unclear.

Last, on the basis of the high energy cost of histidine biosynthesis, the possible constitutive transcription of the heterogeneous *his* operon suggests the existence of a system controlling the expression of *his* genes in *Burkholderia* at a post-transcriptional level. Regulation of *his* operon expression has been particularly studied in *E. coli* and *S. typhimurium* where the general mechanisms and the molecular details of the process are well established. In these two model-organisms the biosynthetic pathway is under control of distinct regulatory mechanisms that operate at different levels and finely tune the expression of the *his* genes. Feedback inhibition by histidine of the activity of the first enzyme of the pathway, *N*-1-(5′-phosphoribosyl)-ATP transferase (ATP-PRT), which is coded for by *hisG*, almost instantaneously adjust the flow of intermediates along the pathway to the availability of exogenous histidine. However, most of the molecular mechanisms controlling the expression of the pathway act at the transcriptional level (Alifano et al., 1996). Apart from *E. coli* and *S. typhimurium*, few studies concerning the regulation of histidine biosynthesis have been performed in other prokaryotes. However, an interesting and sophisticated molecular mechanism responsible for the feedback inhibition of histidine biosynthesis has been disclosed in *Lactococcus lactis* (Alifano et al., 1996; Sissler et al., 1999). In this bacterium the activity of ATP-PRT is controlled by the product of another gene, called *hisZ*, which is required for histidine biosynthesis and is an essential component of the ATP-PRT holoenzyme. Sissler et al. (1999) showed that the HisZ feedback inhibition is mediated by histidine. This function is achieved by forming heteromeric complexes, in which the number of monomers is dynamically regulated by the binding of inhibitors to HisZ and ligands to HisG, respectively, causing a shift to the inactive and the fully active complexes (Bovee et al., 2002). The *hisZ* gene is absent in *E. coli* and *S. enterica*; on the other hand, the scanning of *Burkholderia* genomes using the *L. lactis* HisZ as seed revealed the presence of an orthologous gene in each genome, located outside the heterogeneous operon (Fondi et al., unpublished data). The analysis of microarray data (Fig. 7) revealed that *hisZ* is expressed at similar level in respect to the other *his* "core" genes. Hence, even though up to now no experimental data concerning this issue is available, it is quite possible that the *his* genes in *Burkholderia* cells might be controlled by a mechanism very similar, if not identical, to that disclosed in *L. lactis*.

Last, even though this issue is beyond the scope of this manuscript, it can be underlined that the phylogenetic trees constructed using either *hisB* or *hisA* sequences (see additional file 5), in spite of the partially different branching order they show, strains belonging to the same species clustered together, separating them from strains of different species or genomovars. This finding might have a clinical relevance for identification purposes, in that one or both of them might be used as molecular marker(s) for Bcc strains identification.

Acknowledgments

This work was supported by Italian Cystic Fibrosis Foundation (project 9#2003) and by Ente Cassa di Risparmio di Firenze (project 2003/1034).

The authors are grateful to the anonymous reviewers for their helpful comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.08.002.

References

Alifano, P., et al., 1996. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol. Rev. 60, 44–69.

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Balandreau, J., Viallard, V., Cournoyer, B., Coenye, T., Laevens, S., Vandamme, P., 2001. *Burkholderia cepacia* genomovar III is a common plant-associated bacterium. Appl. Environ. Microbiol. 67, 982–985.

Bovee, M.L., Champagne, K.S., Demeler, B., Francklyn, C.S., 2002. The quaternary structure of the HisZ-HisG *N*-1-(5′-phosphoribosyl)-ATP transferase from *Lactococcus lactis*. Biochemistry 41, 11838–11846.

Brilli, M., Fani, R., 2004a. Molecular evolution of *his8* genes. J. Mol. Evol. 58, 225–237.

Brilli, M., Fani, R., 2004b. The origin and evolution of eucaryal HIS7 genes: from metabolon to bifunctional protein? Gene 339, 149–160.

Bultrini, E., Pizzi, E., 2006. A new parameter to study compositional properties of non-coding regions in eukaryotic genomes. Gene 385, 75–82.

Ciccillo, F., Fiore, A., Bevivino, A., Dalmastri, C., Tabacchioni, S., Chiarini, L., 2002. Effects of two different application methods of *Burkholderia ambifaria* MCI 7 on plant growth and rhizospheric bacterial diversity. Environ. Microbiol. 4, 238–245.

Coenye, T., Vandamme, P., LiPuma, J.J., Govan, J.R., Mahenthiralingam, E., 2003. Updated version of the *Burkholderia cepacia* complex experimental strain panel. J. Clin. Microbiol. 41, 2797–2798.

Das, A., Yanofsky, C., 1989. Restoration of a translational stop-start overlap reinstates translational coupling in a mutant *trpB-trpA* gene pair of the *Escherichia coli* tryptophan operon. Nucleic Acids Res. 17, 9333–9340.

Davis, B.D., Mingioli, E.S., 1950. Mutants of *Escherichia coli* requiring methionine or vitamin B12. J. Bacteriol. 60, 17–28.

Drevinek, P., et al., 2008. Gene expression changes linked to antimicrobial resistance, oxidative stress, iron depletion and retained motility are observed when *Burkholderia cenocepacia* grows in cystic fibrosis sputum. BMC Infect. Dis. 8, 121.
Fani, R., et al., 1989. Cloning of histidine genes of *Azospirillum brasilense*: organization of the ABFH gene cluster and nucleotide sequence of the *hisB* gene. Mol. Gen. Genet. 216, 224–229.
Fani, R., et al., 1993. The histidine operon of *Azospirillum brasilense*: organization, nucleotide sequence and functional analysis. Res. Microbiol. 144, 187–200.
Fani, R., Fondi, M., 2009. Origin and evolution of metabolic pathways. Physics of Life Reviews 6, 23–52.
Fani, R., Liò, P., Chiarelli, I., Bazzicalupo, M., 1994. The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. J. Mol. Evol. 38, 489–495.
Fani, R., Liò, P., Lazcano, A., 1995. Molecular evolution of the histidine biosynthetic pathway. J. Mol. Evol. 41, 760–774.
Fani, R., et al., 1997. Paralogous histidine biosynthetic genes: evolutionary analysis of the *Saccharomyces cerevisiae* HIS6 and HIS7 genes. Gene 197, 9–17.
Fani, R., Mori, E., Tamburini, E., Lazcano, A., 1998. Evolution of the structure and chromosomal distribution of histidine biosynthetic genes. Orig. Life Evol. Biosph. 28, 555–570.
Fani, R., Brilli, M., Liò, P., 2005. The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. J. Mol. Evol. 60, 378–390.
Fani, R., Brilli, M., Liò, P., 2006. Inference from proteobacterial operons shows piecewise organization: a reply to Price et al. J. Mol. Evol. 63, 577–580.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.
Fiore, A., Laevens, S., Bevivino, A., Dalmastri, C., Tabacchioni, S., Vandamme, P., Chiarini, L., 2001. *Burkholderia cepacia* complex: distribution of genomovars among isolates from the maize rhizosphere in Italy. Environ. Microbiol. 3, 137–143.
Fondi, M., Brilli, M., Fani, R., 2007. On the origin and evolution of biosynthetic pathways: integrating microarray data with structure and organization of the Common Pathway genes. BMC Bioinformatics 8 (Suppl. 1), S12.
Frandsen, N., D'Ari, R., 1993. Excess histidine enzymes cause AICAR-independent filamentation in *Escherichia coli*. Mol. Gen. Genet. 240, 348–354.
Fukuda, Y., Nakayama, Y., Tomita, M., 2003. On dynamics of overlapping genes in bacterial genomes. Gene 323, 181–187.
Gabrielian, A., Vlahovicek, K., Pongor, S., 1997. Distribution of sequence-dependent curvature in genomic DNA sequences. FEBS Lett. 406, 69–74.
Goldschmidt, E.P., Cater, M.S., Matney, T.S., Butler, M.A., Greene, A., 1970. Genetic analysis of the histidine operon in *Escherichia coli* K12. Genetics 66, 219–229.
Grisolia, V., Carlomagno, M.S., Bruni, C.B., 1982. Cloning and expression of the distal portion of the histidine operon of *Escherichia coli* K-12. J. Bacteriol. 151, 692–700.
Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids Symp. Ser. 41, 95–98.
Hanahan, D., 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. 166, 557–580.
Inokuchi, Y., Hirashima, A., Sekine, Y., Janosi, L., Kaji, A., 2000. Role of ribosome recycling factor (RRF) in translational coupling. EMBO J. 19, 3788–3798.
Kanhere, A., Bansal, M., 2005. A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinformatics 6, 1.
Mahenthiralingam, E., Coenye, T., Chung, J.W., Speert, D.P., Govan, J.R., Taylor, P., Vandamme, P., 2000. Diagnostically and experimentally useful panel of strains from the *Burkholderia cepacia* complex. J. Clin. Microbiol. 38, 910–913.
Oppenheim, D.S., Yanofsky, C., 1980. Translational coupling during expression of the tryptophan operon of *Escherichia coli*. Genetics 95, 785–795.
Price, M.N., Arkin, A.P., Alm, E.J., 2006. The life-cycle of operons. PLoS Genet. 2, e96.
Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467.
Sissler, M., Delorme, C., Bond, J., Ehrlich, S.D., Renault, P., Francklyn, C., 1999. An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis. Proc. Natl. Acad. Sci. U. S. A. 96, 8985–8990.
Smyth, G.K., 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3, Article 3.
Srere, P.A., 1987. Complexes of sequential metabolic enzymes. Annu. Rev. Biochem. 56, 89–124.
Tabacchioni, S., Ferri, L., Manno, G., Mentasti, M., Cocchi, P., Campana, S., Ravenni, N., Taccetti, G., Dalmastri, C., Chiarini, L., Bevivino, A., Fani, R., 2008. Use of the *gyrB* gene to discriminate among species of the *Burkholderia cepacia* complex. FEMS Microbiol. Lett. 281, 175–82.
Tamames, J., 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. 2, RESEARCH0020.
Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.
Vandamme, P., et al., 2002. *Burkholderia anthina* sp. nov. and *Burkholderia pyrrocinia*, two additional *Burkholderia cepacia* complex bacteria, may confound results of new molecular diagnostic tools. FEMS Immunol. Med. Microbiol. 33, 143–149.
Wernersson, R., Pedersen, A.G., 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31, 3537–3539.
Wong, O.K., Guthold, M., Erie, D.A., Gelles, J., 2008. Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. PLoS Biol. e232, 6.
Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H.J. (Eds.), Evolving Genes and Proteins. Academic Press, New York, pp. 97–166.

3.4 Conclusions

In this chapter, we have performed a "multi-level" analysis of histidine biosynthetic route, one of the best characterized anabolic pathways. Results obtained have provided hints that might reveal useful in differen fields such as the study of the origin of life, the study of metabolic networks (including regulatory ones), the rapid identification of pathogenic strains. Firstly, we have analyzed the fusions involving histidine biosynthetic genes. At least eight out of ten *his* genes, i.e., *hisA*, *B*, *D*, *E*, *F*, *H*, *I*, and *N* underwent different fusion events strongly supporting a major role of this mechanism in both the assembly and evolution of histidine biosynthesis. Each of the five his fusions detected so far, i.e. *hisA*/*hisF*, *hisIE*, *hisHF* (HIS7), *hisNB*, and *hisIED* (HIS4) has been analyzed for: i) gene structure, ii) phylogenetic distribution, iii) timing of appearance, iv) horizontal gene transfer, v) correlation with gene organization, and vi) biological significance. The whole body of data reported above suggests that the fusion(s) of histidine biosynthetic genes has been driven by different selective pressures. In the case of the elongation events leading to the extant *hisA* and *hisF*, a structural/functional significance can be invoked. Indeed, the elongation events were very likely positively selected in order to optimize the structure and the function of the ancestral TIM-barrel. The fusion of HOL-P phosphatase and IGP dehydratase might have been selected to ensure a fixed ratio of gene products that function in the same biochemical pathway. Concerning the *hisHF* (HIS7) fusion, its biological significance is clear; whilst in prokaryotes the two proteins encoded by *hisH* and *hisF* must interact in a 1:1 ratio to give the active form of IGP synthase, in the eukaryotic bifunctional protein, the two entities are fused allowing their immediate interaction and the substrate tunneling. A similar "substrate channeling" and/or "fixed ratio of gene products" might be invoked for the fusion involving the prokaryotic *hisIE* genes, which code for enzymes performing consecutive steps of histidine biosynthesis. Independently from their case-by-case biological significance, such associations (i.e. gene fusions) might be responsible for a more specific commitment of intermediates in a given pathway by means of the spatial co-localization of enzymes. Operons might allow Bacteria to reach the same target: the translation of polycistronic mRNAs favors protein-protein interactions or the spatial segregation of a pathway. Indeed, genes coding for interacting proteins are often organized in operons; in this context, it has been suggested that the bacterial IGP synthase might be part of a complex metabolon whose entities are encoded by the four genes *hisBHAF*, constituting the so-called *core* of histidine biosynthesis. Data presented here might suggest that the polypeptides coded for by *hisI*, *hisE*, and *hisD* are part of another metabolon.

The heterogeneous distribution and organization of *his* genes in Archaea reported in this chapter, despite not allowing saying whether histidine biosynthetic genes were embedded in a compact operon in the LUCA, revealed that they underwent several recombination events during evolution and this led to the different schemes of *his* genes organization that we observe in modern Archaea (and Bacteria). The organization of *his* genes in some extant archaeal lineages speaks toward a piece-wise construction of *his* sub-operons along with gene fusion events and HTG from bacterial donor. Lastly, data suggest also that different molecular mechanisms may drive operon formation during metabolic pathway

origin and evolution.
Lastly, the analysis of the structure, the organization and the regulation of the *his* biosynthetic *core* in the genus *Burkholderia*) revealed that, at least in this this microorganisms, the entire operon is heterogeneous, in that it contains *alien* genes apparently not involved in histidine biosynthesis. Besides, they also support the idea that the proteobacterial *his* operon was piecewisely assembled, i.e. through accretion of smaller units containing only some of the genes (eventually together with their own promoters) involved in this biosynthetic route. Interestingly, it should be underlined that the phylogenetic trees constructed using either *hisB* or *hisA* sequences, in spite of the partially different branching order they show, strains belonging to the same species clustered together, separating them from strains of different species or genomovars. This finding might have a clinical relevance for identification purposes, in that one or both of the might be used as molecular marker(s) for Bcc strains identification.

Chapter 4

Lysine biosynthesis evolution

The analysis of the structure, organization, phylogeny, and distribution of lysine biosynthetic genes revealed that (together within histidine) this route might represent an interesting case study in the context of metabolic pathways origin and evolution . In particular, the analysis of lysine biosynthesis evolution revealed (at least) two important evolutionary features.

1. Two well-distinct routes have been characterized for the anabolism of lysine, that is the α-aminoadipate (AAA) pathway and the diaminopimelate (DAP) one (Figure 4.1). The first one starts from 2-oxoglutarate and leads to lysine, through nine steps, one of which (catalyzed by LysN) is responsible for the formation of α-aminoadipate. Up to now, genes belonging to this pathway have been found in a limited number of(micro)organisms, such as the Bacteria *Thermus thermophilus* and *Deinococcus radiodurans* and the Archaea *Pyrococcus*, *Thermoproteus*, and (probably) *Sulfolobus*. A distinct variant of the AAA pathway has been disclosed in higher Fungi and in euglenoids. The alternative route leading to lysine, referred to as the DAP pathway, involves nine enzymatic reactions and produces lysine starting from L-aspartate. The DAP pathway also plays a central role in cell-wall biosynthesis of gram-negative bacteria, since meso-diaminopimelate is an essential precursor in the biosynthesis of peptidoglycan. Genes involved in the DAP pathway are widespread in both Prokaryotes and Eucaryotes. Interestingly, AAA and DAP pathways are evolutionary linked to leucine and arginine biosynthesis. However, in spite of the available data, no evolutionary model explaining the extant scenario has been proposed. To this purpose, a comparative analysis of the extant leucine, arginine, and lysine metabolic pathways from (micro)organisms whose genome has been completely sequenced was carried out with the aim to trace the evolutionary history of the three metabolic pathways and to shed some light on the ancestral route(s) and interrelationships existing between them and (eventually) with other metabolic routes.

2. Furthermore, lysine (DAP) biosynthesis shares its three initial enzymatic [referred to as the Common Pathway (CP)], with two other biosynthetic pathways, namely threonine, and methionine (Figure 4.2). . In *Escherichia coli* three different aspar-

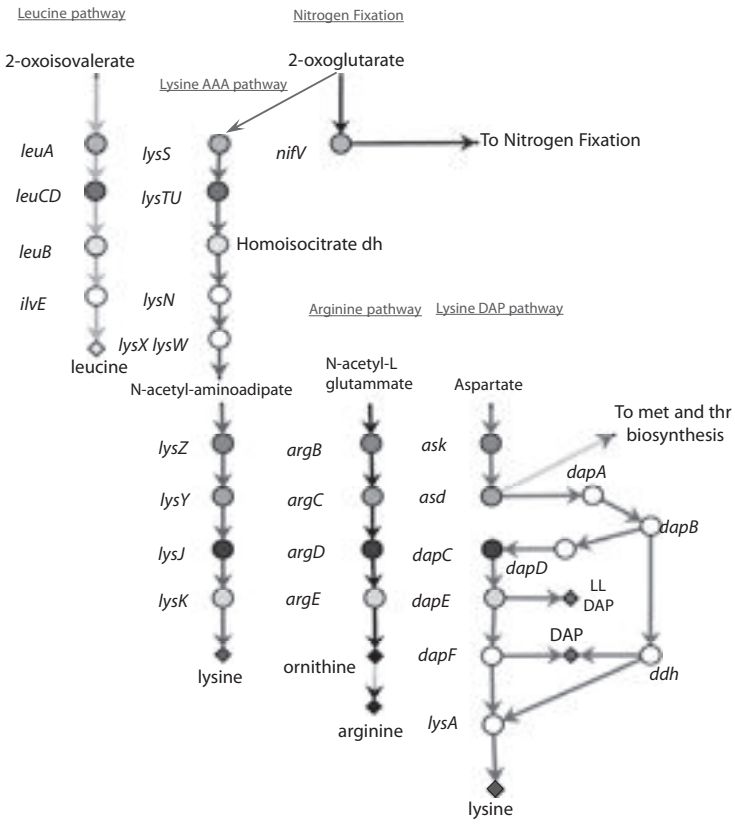


Figure 4.1: The extant lysine, leucine, and arginine biosynthetic routes. Evolutionary relationship between lysine, leucine, and arginine biosynthetic genes. Genes sharing the same colour and the same level are homologs. Genes coloured in white have no homolog in the above mentioned metabolic routes.

tokinases (AKI, AKII, AKIII, the products of *thrA*, *metL* and *lysC*, respectively) can perform the first step of the CP. Moreover, two of them (AKI and AKII) are bifunctional, carrying also homoserine dehydrogenasic activity (*hom* product). The second step of the CP is catalyzed by a single aspartate semialdehyde dehydrogenase (ASDH, the product of asd). Thus, in the CP of *E.coli* while a single copy of ASDH performs the same reaction for three different metabolic routes, three different AKs perfom a unique step. Why and how such a situation did emerge and maintain? How is it correlated to the different regulatory mechanisms acting on these genes? The aim of the work presented in work was to trace the evolutionary pathway leading to this scenario in the extant proteobacteria.

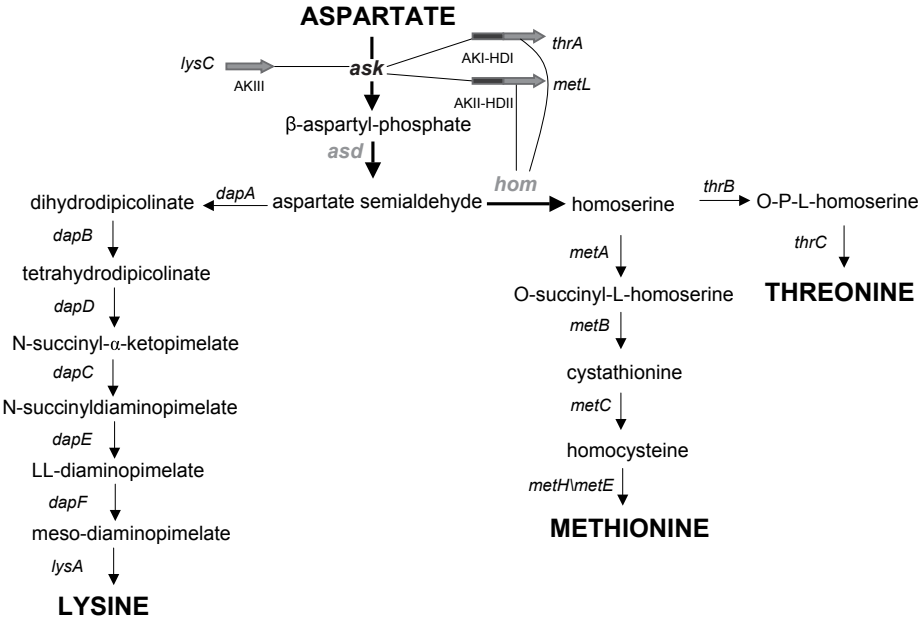


Figure 4.2: The lysine biosynthetic pathway. Genes marked in red (*ask*, *asd*, and *hom*) constitute the Common Pathway.

4.1 An ancestral interconnection between leucine, arginine, and lysine biosynthesis

In the context of metabolic pathways origin and evolution, the lysine, arginine, and leucine biosynthetic routes represent very interesting study-models. In fact, it is known that some of the *lys*, *arg* and *leu* genes are paralogs; this led to the suggestion that their ancestor genes might interconnect the three pathways. The aim of this work was to trace the evolutionary pathway leading to the appearance of the extant biosynthetic routes and to try to disclose the interrelationships existing between them and other pathways in the early stages of cellular evolution. The comparative analysis of the genes involved in the biosynthesis of lysine, leucine, and arginine, their phylogenetic distribution and analysis revealed that the extant metabolic "grids" and their interrelationships might be the outcome of a cascade of duplication of ancestral genes that, according to the patchwork hypothesis, coded for unspecific enzymes able to react with a wide range of substrates. These genes likely belonged to a single common pathway in which the three biosynthetic routes were highly interconnected between them and also to methionine, threonine, and cell wall biosynthesis. A possible evolutionary model leading to the extant metabolic scenarios was also depicted.

BMC Evolutionary Biology

BioMed Central

Research

Open Access

The primordial metabolism: an ancestral interconnection between leucine, arginine, and lysine biosynthesis

Marco Fondi[1], Matteo Brilli[1], Giovanni Emiliani[2], Donatella Paffetti[2] and Renato Fani*[1]

Address: [1]Dipartimento di Biologia Animale e Genetica, Università di Firenze, Via Romana 17\19, Firenze, Italia and [2]Dipartimento di Scienze e Tecnologie Ambientali Forestali, Università di Firenze, Via S. Bonaventura 13, Firenze, Italia

Email: Marco Fondi - marco.fondi@unifi.it; Matteo Brilli - matteo.brilli@dbag.unifi.it; Giovanni Emiliani - giovanni.emiliani@unifi.it; Donatella Paffetti - donatella.paffetti@unifi.it; Renato Fani* - renato.fani@unifi.it

* Corresponding author

from Second Congress of Italian Evolutionary Biologists (First Congress of the Italian Society for Evolutionary Biology)
Florence, Italy. 4–7 September 2006

Published: 16 August 2007

BMC Evolutionary Biology 2007, **7**(Suppl 2):S3 doi:10.1186/1471-2148-7-S2-S3

This article is available from: http://www.biomedcentral.com/1471-2148/7/S2/S3

Abstract

Background: It is generally assumed that primordial cells had small genomes with simple genes coding for enzymes able to react with a wide range of chemically related substrates, interconnecting different metabolic routes. New genes coding for enzymes with a narrowed substrate specificity arose by paralogous duplication(s) of ancestral ones and evolutionary divergence. In this way new metabolic pathways were built up by primordial cells. Useful hints to disclose the origin and evolution of ancestral metabolic routes and their interconnections can be obtained by comparing sequences of enzymes involved in the same or different metabolic routes. From this viewpoint, the lysine, arginine, and leucine biosynthetic routes represent very interesting study-models. Some of the *lys*, *org* and *leu* genes are paralogs; this led to the suggestion that their ancestor genes might interconnect the three pathways. The aim of this work was to trace the evolutionary pathway leading to the appearance of the extant biosynthetic routes and to try to disclose the interrelationships existing between them and other pathways in the early stages of cellular evolution.

Results: The comparative analysis of the genes involved in the biosynthesis of lysine, leucine, and arginine, their phylogenetic distribution and analysis revealed that the extant metabolic "grids" and their interrelationships might be the outcome of a cascade of duplication of ancestral genes that, according to the patchwork hypothesis, coded for unspecific enzymes able to react with a wide range of substrates. These genes belonged to a single common pathway in which the three biosynthetic routes were highly interconnected between them and also to methionine, threonine, and cell wall biosynthesis. A possible evolutionary model leading to the extant metabolic scenarios was also depicted.

Conclusion: The whole body of data obtained in this work suggests that primordial cells synthesized leucine, lysine, and arginine through a single common metabolic pathway, whose genes underwent a set of duplication events, most of which can have predated the appearance of the last common universal ancestor of the three cell domains (Archaea, Bacteria, and Eucaryotes). The model proposes a relative timing for the appearance of the three routes and also suggests a possible evolutionary pathway for the assembly of bacterial cell-wall.

Background

It is commonly assumed that early organisms inhabited an environment rich in organic compounds spontaneously formed in the prebiotic world, an idea that is often referred to as the Oparin – Haldane theory. Those primordial organisms had no need for developing new and improved metabolic abilities since most of the required nutrients were available. However, their increasing number would have led to the depletion of essential nutrients imposing a progressively stronger selective pressure that, in turn, favoured those (micro)organisms that have become able to synthesize the nutrients whose concentration was decreasing in the primordial soup. Thus, the origin and the evolution of basic biosynthetic pathways represented a crucial step in cellular evolution, since it rendered the primordial cells less dependent on the external source of nutrients. But how did these metabolic pathways emerge and evolve? Several theories have been proposed to explain how the metabolic routes have been assembled (see [1] and reference therein). Even though it is possible that different processes might have been responsible for the build up of metabolic routes, a large body of data concerning both sequence comparison and experimental data on enzymes substrate specificity strongly supports one of them, that is the *patchwork assembly* theory [2]. According to this idea, metabolic pathways might have been originated through the recruitment of primitive enzymes that could react with a wide range of chemically related substrates. Such relatively slow, unspecific enzymes may have enabled primitive cells containing small genomes to overcome their limited coding capabilities. Paralogous gene duplication event(s) followed by evolutionary divergence might have permitted the appearance of enzymes with an increased and narrowed specificity and/or the functional diversification. In this way, an ancestral enzyme belonging to a given metabolic route was "recruited" to serve a single or other (novel) pathways. The importance of gene duplication in the course of evolution of genomes and metabolic pathways is well established [3]: the production of two (or more) copies of a DNA sequence leads to an increase of genome size, and it also allows the (rapid) diversification of enzymes, providing material for the invention of new enzymatic properties and complex regulatory and developmental patterns. Therefore, gene duplication should have had a deep impact on primordial metabolism. Indeed, it is quite possible that the biochemical flexibility of the ancestral enzymes might result in an extreme interconnection among different metabolic routes. Hence, the duplication event(s) followed by evolutionary divergence, allowing gaining of novel metabolic capabilities, permitted the new metabolic pathways to be less branched and interconnected to each other. How can this issue be studied? Useful hints to disclose the origin and evolution of metabolic pathways and the possible ancestral interrelationships between different routes may be obtained by comparing the sequence and the structure of genes (and/or the products they code for) of the same and different routes from (micro)organisms belonging to the three cells domains (Archaea, Bacteria and Eucarya). In this context the lysine, arginine, and leucine biosynthetic pathways represent interesting study-models. One of the main reasons is the existence of two well-distinct routes that have been characterized for the anabolism of lysine, that is the α-aminoadipate (AAA) pathway and the diaminopimelate (DAP) one [4-6] (Figure 1). The first one starts from 2-oxoglutarate and leads to lysine, through nine steps, one of which (catalyzed by LysN) is responsible for the formation of α-aminoadipate. Up to now, genes belonging to this pathway have been found in a limited number of (micro)organisms, such as the Bacteria *Thermus thermophilus* and *Deinococcus radiodurans* [7-9] and the Archaea *Pyrococcus* [8], *Thermoproteus* [10], and (probably) *Sulfolobus* [10,11]. A distinct variant of the AAA pathway has been disclosed in higher Fungi [6,12-14] and in euglenoids [15,16]. The alternative route leading to lysine, referred to as the DAP pathway, involves nine enzymatic reactions and produces lysine starting from L-aspartate. The DAP pathway also plays a central role in cell-wall biosynthesis of gram-negative bacteria, since meso-diaminopimelate is an essential precursor in the biosynthesis of peptidoglycan [17,18]. Genes involved in the DAP pathway are widespread in both Prokaryotes and Eucaryotes [19,20]. Interestingly, AAA and DAP pathways are evolutionary linked to leucine and arginine biosynthesis. The relationship existing between genes belonging to these biosynthetic routes has been previously analyzed ([19] and reference therein) and is schematically represented in Figure 1. The products of the four genes involved in the DAP pathway (*ask*, *asd*, *dapC*, and *dapE*) are evolutionary related to arginine biosynthetic enzymes encoded by *argB*, *argC*, *argD* and *argE*, respectively [19]. Some of the enzymes involved in the AAA pathway share a high degree of sequence similarity with enzymes belonging both to leucine and arginine biosynthetic routes [8,19] since the first four enzymes of the *Thermus*-like AAA biosynthetic pathway (LysS, LysT, LysU, and Homoisocitrate dehydrogenase) are homologous to the corresponding enzymes of leucine biosynthesis (LeuA, LeuC, LeuD, LeuB) [8,19,21]. Moreover LysZ, LysY, LysJ and LysK are homologous to ArgB, ArgC, ArgD, and ArgE, respectively [19,22]. The high degree of sequence similarity shared by these enzymes led to the suggestion that: i) the assembly of both the DAP and AAA routes might be explained as the outcome of a series of gene duplication events followed by specialization [19], ii) the DAP route should represent the ancestral pathway leading to lysine and, iii) the AAA pathway should be a more recent invention of evolution[19].

However, in spite of the available data, no evolutionary model explaining the extant scenario has been proposed. The aim of this work was to try to trace the evolutionary history of the three metabolic pathways and to shed some light on the ancestral route(s) and interrelationships existing between them and (eventually) with other metabolic routes. To this purpose, a comparative analysis of the extant leucine, arginine, and lysine metabolic pathways from (micro)organisms whose genome has been completely sequenced was carried out.

Results and discussion

Distribution of leucine, arginine and lysine biosynthetic genes

The amino acid sequence of each of the *E. coli* lysine (DAP), leucine, and *T. thermophilus* lysine (AAA) biosynthetic enzymes, was used as a query to probe the completely sequenced genomes database of KEGG (Kyoto Encyclopaedia of Genes and Genomes) consisting of 29 Archaea, 423 Bacteria and 35 Eucaryotes. The bidirectional best-hit (BBH) criterion (see Methods) was used to retrieve orthologous sequences. Data obtained, schematically reported in Figure 2 and representative of a dataset of 68 bacterial and 15 archaeal genomes, revealed that:

i) Lysine (AAA) biosynthetic genes are very rarely represented in Bacteria, in that just two organisms harbour the complete set of AAA biosynthetic genes, i.e. *T. thermophilus* and *D. radiodurans*. No other bacterium possesses the complete set of enzymes required to synthesize lysine via the AAA route. Six Archaea display a complete set of lysine (AAA) biosynthetic genes, namely *Pyrococcus* and *Sulfolobus* strains, and *Thermococcus kodakaraensis*.

ii) Lysine (DAP) biosynthetic genes are widespread among Bacteria. Among them, 18 (micro)organisms possess a complete set of genes (9) for lysine biosynthesis through the DAP route. The small number of lysine (DAP) biosynthetic genes in some bacterial strains is very likely due to the absence of the corresponding metabolic route, which, in turn, is related to the parasitic lifestyle of these organisms. Such a lifestyle may allow these bacteria to acquire essential compounds directly from the metabolic activities of their host and the adaptation to this environmental condition might have caused the loss of entire metabolic routes or parts thereof.

Although some Archaea are known to synthesize lysine through the DAP pathway [23-28], it is not still completely clear which of the possible variants they use. Nonetheless, no archaeon possesses a number of genes compatible with the number required by the succinylasic variant. When the *ddh* sequence of *Corynebacterium glutamicum*, whose product is involved in the dehydrogenasic variant of lysine (DAP) pathway, was used in a BLAST probing, the only *ddh* archaeal homolog sequence was found in *Archaeoglobus fulgidus* genome (data not shown).

iii) The complete set of leucine biosynthetic genes is present in most of Bacteria and Archaea that possess either all the five leucine biosynthetic genes or four of them (lacking the last gene of leucine biosynthetic route, *ilvE*). This different distribution of *ilvE*, had already been observed by Velasco *et al.* [19].

iv) Arginine biosynthetic genes are widespread among Archaea and Bacteria. Even though a complete set is found only in a restricted number of organisms, most of them possess more than half (from five up to eight) of the enzymes required for the biosynthesis of ornithine and arginine, confirming the previous observation that they are synthesized through *N*-acetylated intermediates both in Bacteria and in Archaea [29].

Structure of leucine, arginine, and lysine biosynthetic pathways

A comparative analysis of the lysine, leucine, and arginine biosynthetic routes of completely sequenced organisms belonging to the three cellular domains (Eucarya, Bacteria, and Archaea) was carried out. Data obtained for some representative organisms belonging to the three cellular domains are shown in Figures 3, 4, and 5 and can be summarised as reported below.

In most of archaeal, bacterial, and eukaryotic (micro)organisms each of the three amino acid is synthesized through a specific metabolic route and each enzyme catalyzes a single step of the metabolic pathway it belongs to. The only exception is represented by the archaeon Pyrococcus horikoshii (see below); moreover a bifunctional enzyme, coded for by argD, interconnecting arginine and lysine biosyntheses was identified and characterized in Escherichia coli (see below).

In all of the organisms analyzed, leucine is synthesized through the typical biosynthetic pathway, consisting of the enzymatic steps catalyzed by 3-isopropylmalate synthase, 2-isopropylmalate isomerase, 3-isopropylmalate dehydrogenase, and α-ketoisocaproate transaminase, respectively. The four enzymes are coded for by the procaryotic *leuA*, *leuCD*, *leuB*, and *ilvE* genes, or by the eucaryotic *LEU9|4*, *LEU1*, *LEU2*, and *BAT1|2* (Figures 3, 4, 5).

Similarly, arginine biosynthesis proceeds through the same steps in all the organisms taken into account (Figures 3, 4, 5).

The lysine biosynthesis is much more intriguing; the amino acid can be synthesized through either the AAA or

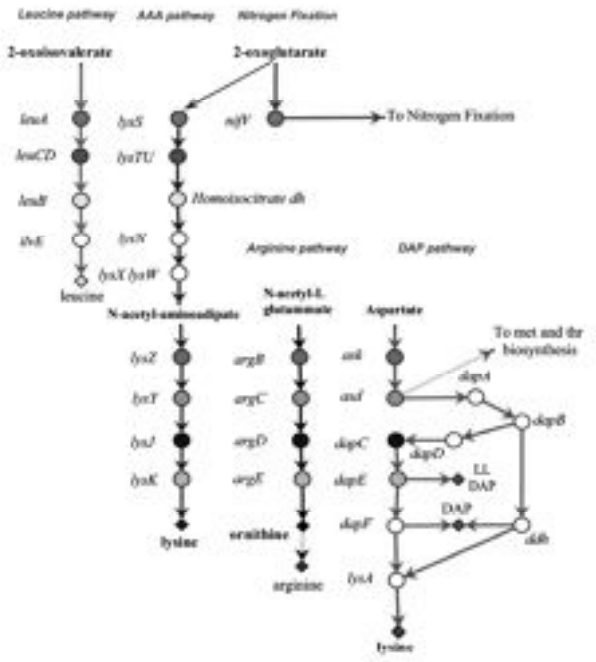


Figure 1
The extant lysine, leucine, and arginine biosynthetic routes. Evolutionary relationship between lysine, leucine, and arginine biosynthetic genes. Genes sharing the same colour and the same level are homologs. Genes coloured in white have no homolog in the metabolic routes taken into account in this work (modified from Velasco et al 2002 [19]).

the DAP pathway (see Background). The AAA pathway was found in *T. thermophilus* and *D. radioduranss*, and a modified version of it was found in *S. cerevisiae* (Figure 4a and Figure 5a). The first four metabolic steps are shared by the two organisms and are evolutionarily related to the corresponding ones of leucine biosynthesis [8,19]. The second moiety of these two AAA versions is different. In *T. thermophilus* lysine biosynthesis is achieved by the products of *lysZ, Y, J*, and *K*, whose sequences are homologous to *argB, C, D*, and *E*, respectively, whereas in *S. cerevisiae* lysine biosynthesis is completed by *LYS2, LYS9, LYS1*, that have no homolog among the genes of the biosynthetic pathways we took into account.

In the other Bacteria and in some Archaea, lysine is synthesized through the DAP pathway. Even in this case, several alternatives are possible. In the γ-proteobacterium *E. coli* (Figure 4c) the product of DapB is converted in LL-meso-diaminopimelate through four sequential enzymatic steps, known as the succinylate variant, and performed by DapD, ArgD, DapE, and DapF. Moreover, the enzyme coded for by *argD* exhibits both N-acetyl-ornithine and N-succinyl-L,L-diaminopimelate aminotransferase activities, with a very similar catalytic efficiency and identical kinetic mechanism, suggesting that this enzyme can play a role in both lysine and arginine biosynthesis [30]. In *C. glutamicum* (Figure 4b) another version of the pathway has been disclosed: tetrahydrodipicolinate can be converted into meso-diaminopimelate by the activity of diaminopimelate dehydrogenase (*ddh*) in a single metabolic step. It is highly likely that this catalytic step is present also in the DAP pathway of the archaeon *Haloarcula hispanica* (Figure 3a) [23]. Interestingly, in *C. glutamicum* two distinct enzymes (ArgD and DapC) can perform the reaction that in *E. coli* is carried out by a single one. Thus, the enzymes encoded by *argD* and *dapC* in *C. glutamicum* can be considered specific for the biosynthetic routes of arginine and lysine, respectively.

Arabidopsis thaliana is known to synthesize lysine through the DAP pathway [15,31]. A DAP variant has been recently identified in this plant that utilizes a novel transaminase (LL-aminoadipate aminotransferase, the product of ORF *At4g33680*) that specifically catalyzes the conversion of tetrahydrodipicolinate to LL-diaminopimelate, a reaction requiring three enzymes in the DAP-pathway variant found in *E. coli* [32].

The archaeon *P. horikoshii* represents the main exception. In fact, in this organism just one pathway for the biosynthesis of all these three amino acids has been identified so far (Figure 3b). On the basis of sequence comparison and phylogenetic analyses, it has been suggested [9] that ORF PH1722, PH1724, PH1726, and PH1727 (see Figure 3b) from *P. horikoshii* might be involved in leucine biosynthesis as well as in the AAA variant of lysine biosynthesis, and that ORF PH1720, PH1718, PH1716, and PH1715 (Figure 3b) might be involved in the biosynthesis of both lysine (through the AAA one) and arginine. Even though it cannot be ruled *a priori* out the possibility that other uncharacterized routes for the biosynthesis of these amino acids may exist [33], it has been suggested that *P. horikoshii* possesses a unique amino acid biosynthetic system by which several amino acids are synthesized by a limited number of enzymes with broad substrate specificity [8].

A hypothetical ancestral pathway for lysine, leucine, and arginine

Data shown in Figures 3, 4 and 5 reveal that some steps of lysine, leucine, and arginine biosynthetic pathways are strongly conserved among the organisms we have taken into account. These steps are the first three of leucine route (performed by LeuA, LeuCD, and LeuB), the central ones of the arginine route (performed by ArgB, ArgC, ArgD, and ArgE), and the first two of lysine (DAP) route (performed by Ask and Asd). Moreover, these enzymes, share a signif-

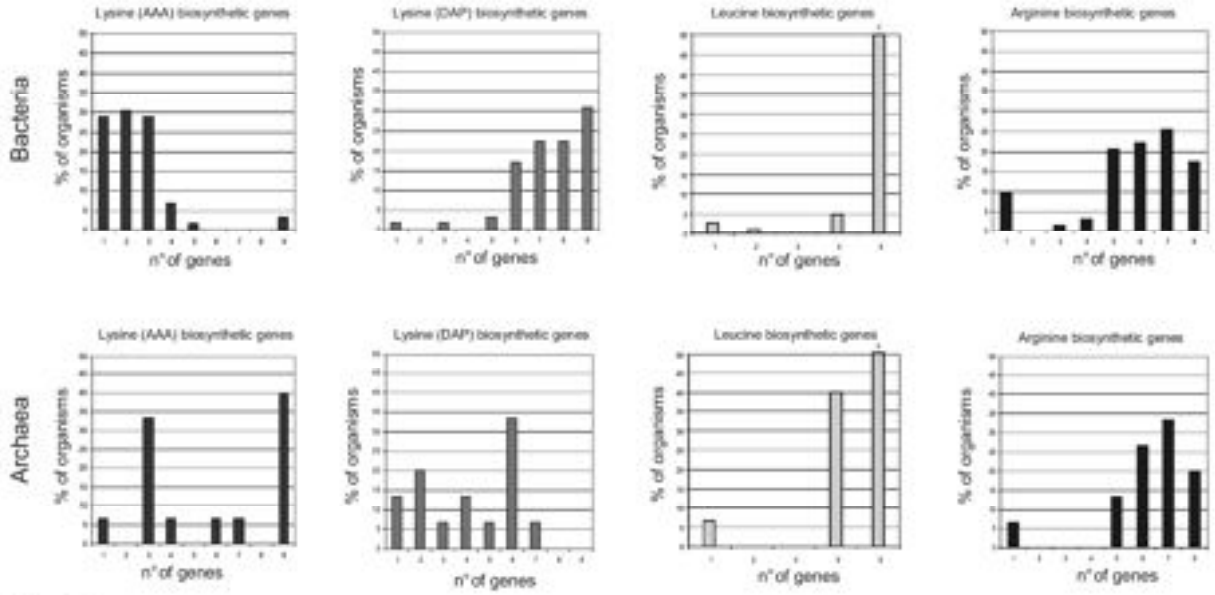


Figure 2
Distribution of lysine, leucine, arginine biosynthetic genes. Histogram showing the number of lysine, leucine, and arginine biosynthetic genes possessed by Bacteria and Archaea.

icant degree of sequence similarity with some of those involved in the AAA lysine biosynthesis [19]. As shown in Figure 1, *leuA*, *leuCD*, and *leuB* appeared to be paralogous to *lysS*, *lysTU*, and homoisocitrate dehydrogenase coding gene, respectively. A similar paralogy exists between *argB*, *ask*, and *lysZ*, as well as among *argC*, *asd*, and *lysY*. Velasco *et al.* [19] suggested that all of them are the outcome of one or more duplication events of an ancestral set of genes. According to the patchwork hypothesis [2], these ancestral genes might have encoded enzymes possessing broad substrate specificity and have been involved in different metabolic pathways.

On the basis of the analysis of phylogenetic distribution, Velasco *et al.* [19] also proposed that the DAP variant of lysine biosynthesis appeared earlier than the AAA one. However, a different scenario can be proposed. The model that we describe here predicts the existence of an ancestral pathway consisting of a set of genes, some of which coding for unspecific enzymes able to react with a wide range of chemically related substrates. This metabolic pathway interconnected lysine, leucine, arginine and also methionine and threonine biosyntheses and nitrogen fixation (Figure 6). According to the model, the first step of the primordial route was catalyzed by the ancestor of the extant isopropylmalate synthase (IPMase, LeuA) and homocitrate synthase (HCase, NifV), that are involved in leucine biosynthesis and nitrogen fixation, respectively. The paralogy existing between their coding genes has not been analyzed in detail up to now, but its description is beyond the scope of the present work. Moreover, the first three steps of this ancestral biosynthetic pathway might have included the reactions that, in the extant organisms, are separately accomplished by the enzymes of leucine and lysine-AAA pathways. Moreover, the last four steps, involving the ancestral copies of the extant ArgB, C, D, and LysZ, Y, J, K, may have been able to recognize different substrates and to catalyze their conversion into ornithine, a fundamental intermediary in arginine biosynthesis.

Even though these ancestral enzymes may have been the backbone of the hypothetical common route, some others might have been necessary to complete the biosynthesis of all the corresponding products. An ancestor of the extant IlvE might have catalyzed the conversion of 2-oxoisocaproate into leucine, the final step of leucine biosynthesis, whereas an unspecific aminotransferase may have accomplished the transamination of α-aminoadipate, the extant role of LysN.

Moreover, an ancestral copy of LysX, which is evolutionarily related to *E. coli* RimK [34], was probably involved in the modification of α-aminoadipate to N-acetyl-α-aminoadipate, releasing the substrate of LysZ/ArgB/Ask ancestor. As shown in Figure 5, the model proposed does not

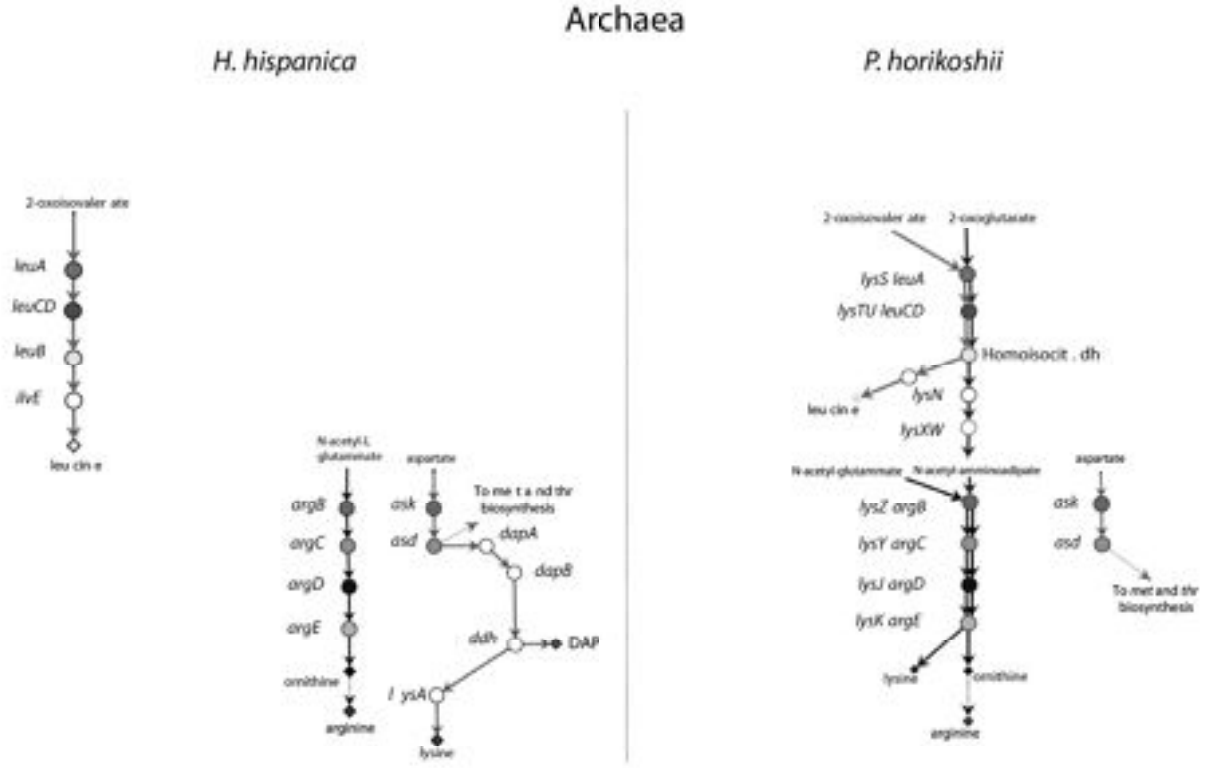


Figure 3
Structure of lysine, leucine, and arginine biosynthetic routes in the Archaea H. hispanica and P. horikoshii. The lysine, leucine, and arginine biosynthetic routes in H. hispanica (a) and in P. horikoshii (b) (see references in the text). Genes sharing the same colour and the same letter are homologs. Genes coloured in white have no homolog in the metabolic routes studied in this work.

contemplate the biosynthesis of LL-diaminopimelate in primordial cells, implying that the AAA pathway predated the appearance of the DAP pathway.

In conclusion, ancestral organisms may have been able to synthesize lysine, leucine and ornithine using just one pathway, consisting of a small number of enzymes endowed with a broad substrate specificity, permitting a clear interconnection of different metabolic routes.

A model for the evolution of lysine, leucine, and arginine metabolic routes

On the basis of the structure and the relationships existing in the extant biosynthetic routes leading to lysine, leucine, or arginine we depict a possible evolutionary model that might explain the (complex) extant scenario.

The model predicts the existence of the above described ancestral pathway (Figures 6 and 7a) that appears to be very similar to that responsible for the biosynthesis of lysine, leucine, and ornithine in P. horikoshii (Figure 3b and 6). It has been proposed [8] that P. horikoshii might have developed unique amino acid biosynthetic systems in which several amino acids are synthesized by a limited number of enzymes with broad substrate specificity and that these enzymes might be the ancestors of the extant biosynthetic ones of lysine, leucine, and arginine.

If the idea of an ancestral common pathway for the biosynthesis of lysine, leucine and arginine is correct, this raises the question of how did the extant routes originate. Besides, which was the timing of their appearance? Here we propose a model that is in agreement with the Nishida's idea and predicts that the extant biosynthetic routes might have emerged after a set of duplication events involving the genes of the ancestral common pathway (Figure 7). According to the model, the first step (or one of the first steps) might have been the duplication of the genes encoding the enzymes catalysing the sixth and seventh steps of the route common to lysine and ornithine.

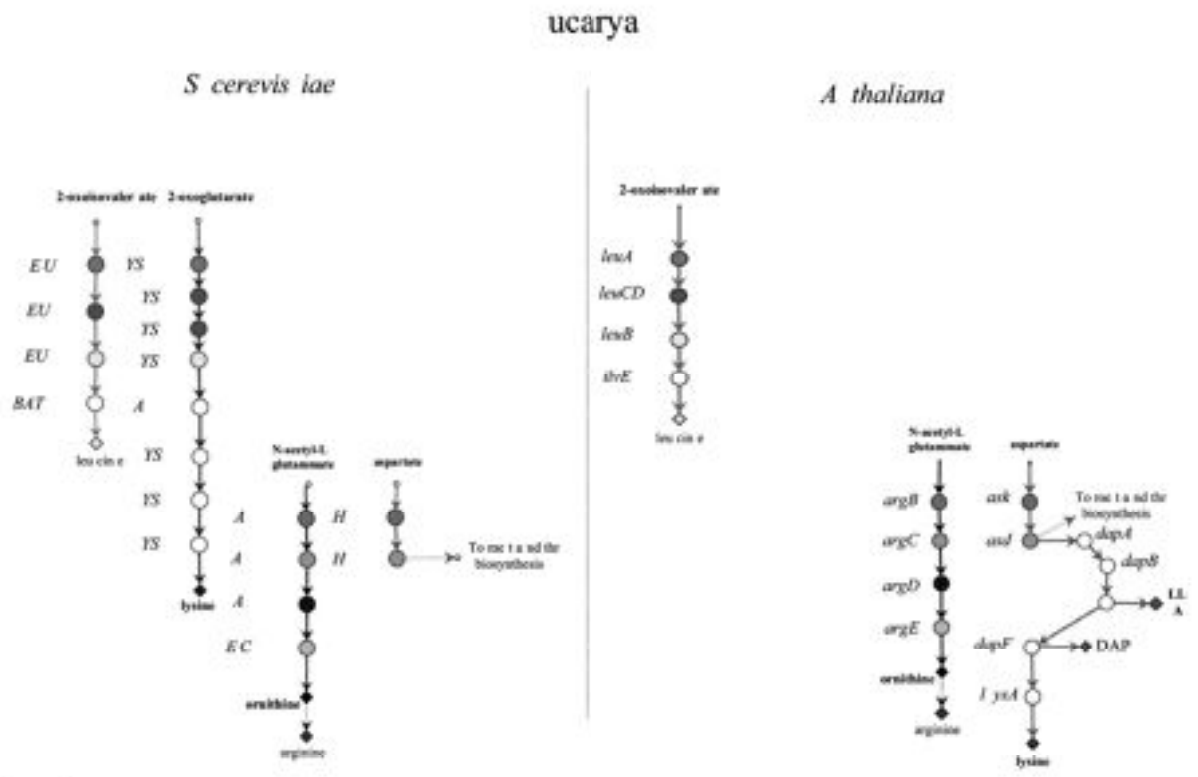


Figure 4

Structure of lysine, leucine, and arginine biosynthetic routes the Eucarya *S. cerevisiae* and *A. thaliana*. The lysine, leucine, and arginine biosynthetic routes in *S. cerevisiae* **(a)** and in *A. thaliana* **(b)** (see references in the text). Genes sharing the same colour and the same level are homologs. Genes coloured in white have no homolog in the metabolic routes studied in this work.

These events gave raise to the ancestors of *ask* and *asd* genes and to the overall metabolic "grid" that has been found in *P. horikoshii* (Figures 3b and 7b). This idea is also supported by the phylogenetic analysis (see next section).

Later on, the duplication of the genes encoding the first three steps of the ancestral pathway and the further evolutionary divergence of the new copies might have originated the ancestor of *leuA*, *leuCD*, and *leuB*, which coded for enzymes with a more narrow substrate specificity, rendering the leucine biosynthesis independent from the ancestral common pathway (Figure 7c). The central steps of arginine biosynthesis, catalyzed by ArgB, ArgC, ArgD and ArgE, might have arisen from the duplication and further evolutionary divergence of the corresponding ancestral aspecific genes (Figure 7d). In this way the four metabolic routes leading to leucine, lysine, arginine, and methionine\threonine, respectively, became one independent from each other. This metabolic scheme corresponds to that found in *T. thermophilus* and *D. radiodurans*.

A further duplication of the bottom genes of the ancestral pathway led to the appearance of DapC and DapE; during this step the branch of the "DAP pathway" leading to lysine was completely assembled (Figure 7e) with the recruitment of *dapC* and *dapE* (that are homolog to the extant *lysI*/*argD* and *lysZ*/*argB*, respectively). *dapF*, and *lysA*. On the basis of the available data, it is not still possible to discern between the possibility that the product of the reaction catalyzed by Asd directly interacts with DapC, or if other enzymes (possibly the ancestor of the extant *dapA*, *dapB*, and *dapD*) were required to complete the DAP route. However, it is possible that once the assembly of the lysine-branch of the DAP pathway was completed the primordial cells possessed two alternative ways to synthesize lysine. If the model we propose here is correct, it is possible that, in this context, the entire AAA pathway became superfluous. Hence, the fourth step (Figure 7e) of the model predicts that, after the completion of the DAP pathway to lysine, genes belonging to the AAA pathway were progressively lost; therefore those organisms lacking

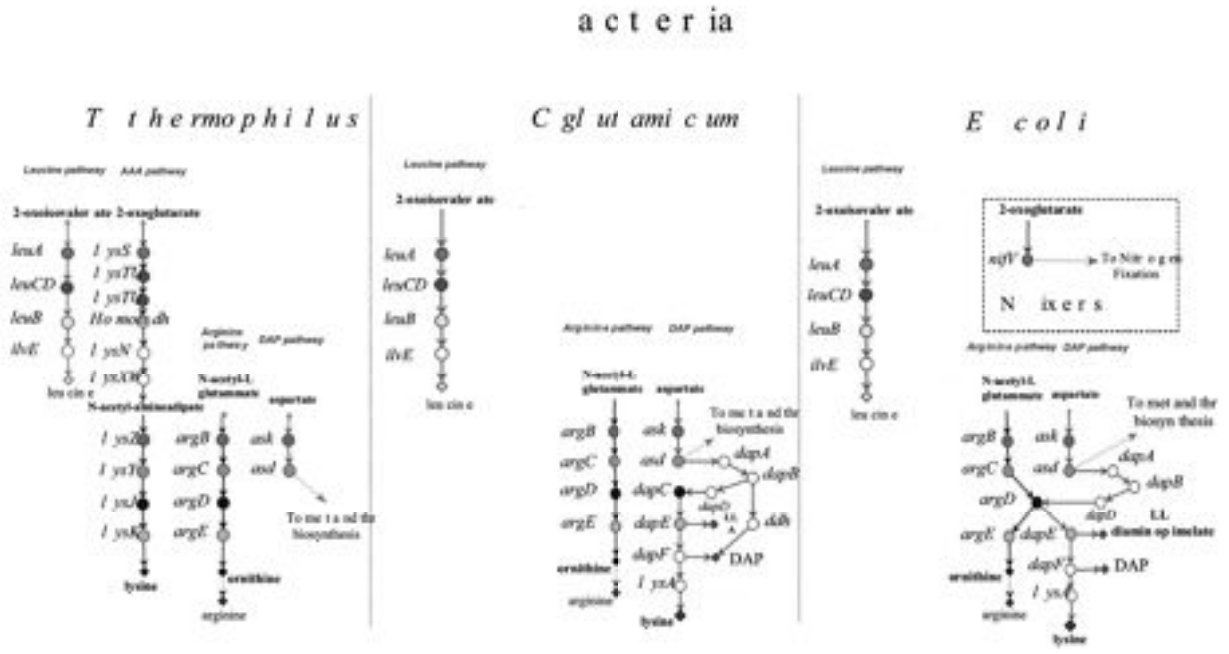


Figure 5

Structure of lysine, leucine, and arginine biosynthetic routes in some Bacteria. The lysine, leucine, and arginine biosynthetic routes in *T. thermophilus* **(a)**, in *C. glutamicum* **(b)**, and in *E. coli* **(c)** (see references in the text). Genes sharing the same colour on the same level are homologs. Genes coloured in white have no homologs in the metabolic routes studied in this work.

the DAP pathway, maintained the ability to synthesize lysine through the AAA route. The final step of the model (Figure 7f) predicts that lysine (DAP) may have been completely assembled, with the appearance of all the extant variants (deacetylasic, desuccynilasic, and dehydrogenasic), leading to the metabolic patterns found in *E. coli* and in many other microorganisms. The development of the DAP pathway might allow bacterial cells to acquire the possibility to insert LL-diaminopimelate and meso-diaminopimelate, that are both intermediaries of the DAP route, into their cell wall.

In our opinion, the duplication events from Figure 7a to Figure 7e would have predated the appearance of the Last Universal Common Ancestor (LUCA), which, according to Woese [35], would have been a complex community of progenotes, highly dynamic from a genomic viewpoint, a mix of heterogeneous primordial cells with different metabolic abilities that could be easily exchanged between the different entities of the community. According to this idea, cells with different metabolic grids such as those represented in Figure 7b and 7d might have co-existed in the "LUCA community". The model also predicts, according to Cavalier-Smith [36], that if the primordial cells were surrounded by a cell wall containing peptidoglycan, ornithine should have been one of the first, if not the very first, component of peptidoglycan (see *Conclusions*).

Phylogenetic analysis

According to the model proposed, the metabolic pathway leading to methionine and threonine diverged from the ancestral pathway before lysine, leucine, and arginine biosynthetic ones. However, (at least) another plausible scenario can be depicted, involving the previous appearance of the leucine pathway and the further assembly of the others biosynthetic routes. In other words, is it possible to establish the relative timing of the separation of the extant three pairs of genes involved in lysine, arginine, threonine, and methionine biosynthesis, that is *lysZ* and *lysY*, *argB* and *argC*, and *ask* and *asd*?

Useful hints concerning this issue can be obtained by a phylogenetic analysis of the two paralogous triads (*LysZ*, *ArgB*, *Ask*, and *LysY*, *ArgC*, and *Asd*). To this purpose a dataset of all the retrieved LysZ, ArgB, Ask and, LysY, ArgC, and Asd amino acid sequences was aligned using the pro-

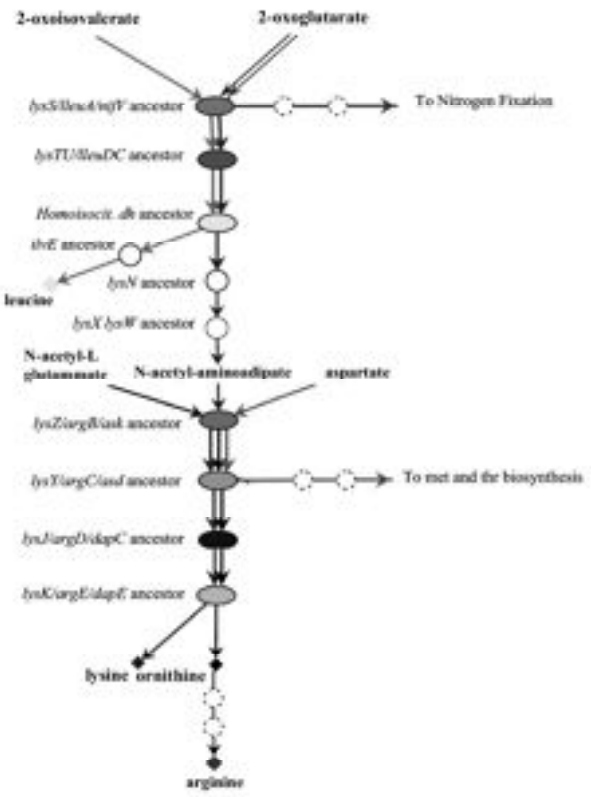


Figure 6
A hypothetical ancestral common route for lysine, leucine, and arginine. The ancestral metabolic pathway, constituted by a restricted, unspecific set of enzymes, able to synthesize leucine, lysine, ornithine, and LL-diaminopimelate. Coloured genes are related to the extant lysine, leucine, and arginine biosynthetic ones. Dashed circles indicate more than one metabolic step. Genes coloured in white have no homologs in the metabolic routes studied in this work.

gram ClustalW [37] and the multialignments obtained used to draw the phylogenetic trees shown in Figure 8. This analysis revealed that:

1. Ask sequences form a unique cluster, which is clearly separated from the other one containing all the ArgB and LysZ sequences involved in arginine and lysine (AAA) biosynthesis, respectively (Figure 8a).

2. Asd sequences form a unique cluster separated from ArgC and LysY sequences, involved in arginine and lysine biosynthesis, respectively (Figure 8b).

Hence, the topology of the tree in Figure 8a suggests that ArgB and LysZ sequences share a degree of sequence similarity higher than that exhibited with Ask, respectively. Similarly, the tree shown in Figure 8b, revealed that ArgC and LysY sequences share a degree of similarity higher than that exhibited with Asd. The overall body of phylogenetic data suggests that the first duplication event(s) involving the ancestral common genes might have originated the ancestor of the extant *ask* and *asd* genes; hence, *argB* and *argC* might be the outcome of a later duplication event of their corresponding ancestral sequences.

Thus, in our opinion, the model proposed in the previous section appeared to be in agreement with both phylogenetic analyses and the metabolic schemes shown in Figures 2, 3, 4.

Conclusion

In this work a likely model for the evolution of genes involved in the biosynthesis of lysine, leucine, and arginine is depicted. The model proposed is based on the analysis of the structure of these biosynthetic routes and the phylogenetic distribution of their genes. The phylogenetic analysis performed allowed us also to determine a possible relative timing of the appearance of genes that are involved in the extant lysine (DAP) and arginine biosynthetic routes. This analysis gave a strong support to the hypothesis that extensive gene duplication events played a key role in shaping the extant biosynthetic routes of lysine, leucine and arginine. According to the model proposed in this work a common metabolic pathway for the biosynthesis of these three amino acids predated the appearance of the last universal common ancestor. This ancestral metabolic route was probably composed of a set of unspecific enzymes able to react with chemically related substrates interconnecting different biosynthetic routes (Figure 6). The occurrence of multiple gene duplication events (Figure 7) would have led to the appearance of specific metabolic pathways responsible for the biosynthesis of each amino acid. The evolutionary history of lysine, leucine, and arginine biosynthetic routes strongly supports the hypothesis on the origin and evolution of metabolic pathways proposed by Jensen [2], strengthening the idea that the gene duplication and the recruitment of genes encoding enzymes with a broad substrate specificity played a key role in the assembly of primitive metabolic routes.

Hence, starting from a metabolic network whose (highly interconnected) nodes represent unspecific enzymes (Figure 6), novel metabolic networks have emerged, consisting of highly specialized (and less interconnected) enzymes (Figure 1). In this way, ancestral enzymes belonging to a given metabolic route, have been "recruited" to serve a single or other (novel) pathways.

One of the main consequences of this evolutionary pathway, i.e. gene duplication followed by evolutionary diver-

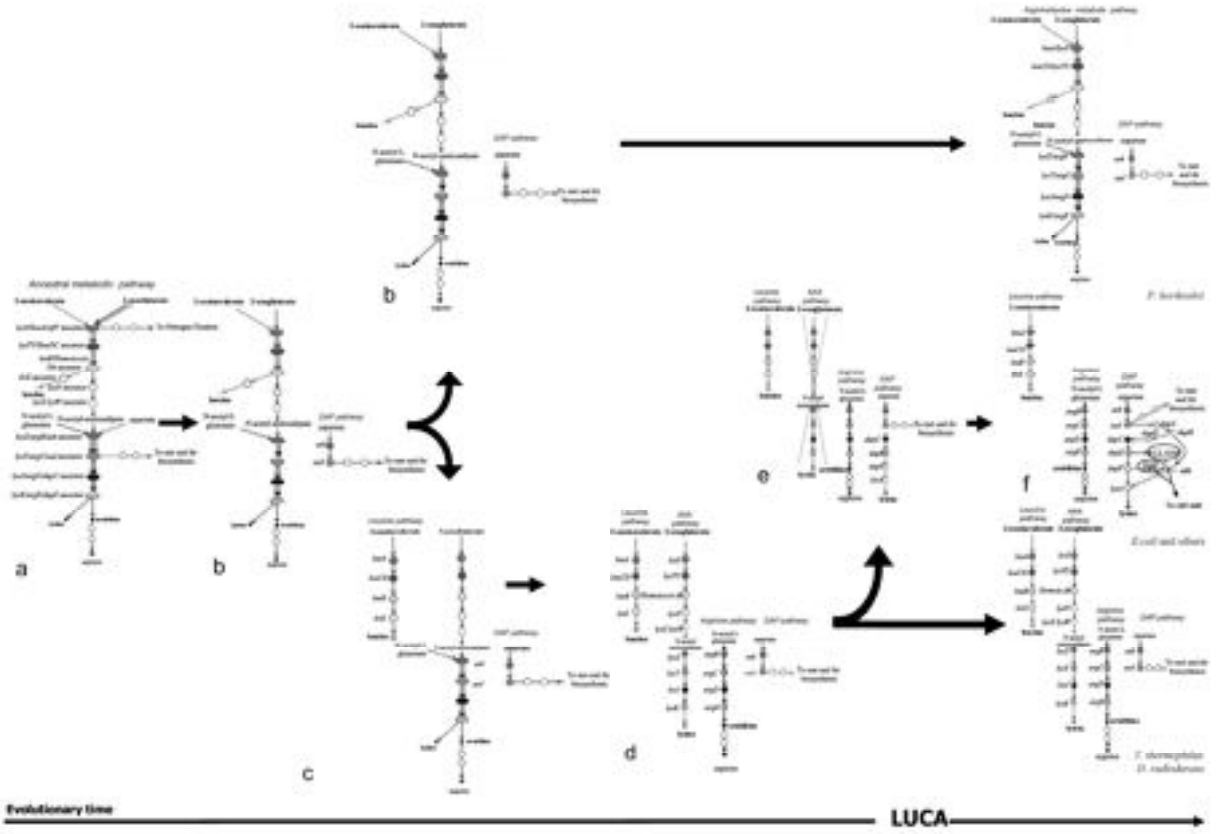


Figure 7
Evolutionary model for the assembly of lysine, leucine, and arginine biosynthetic pathways. Evolutionary model proposed to explain the evolution and the assembly of lysine, leucine, and arginine biosynthetic pathways starting from the hypothetical common route. Dashed circles indicates more than one metabolic step.

gence, is the separation of metabolic routes that were originally fused in a single, common one. The biological significance of this cascade of duplication events might rely on both (i) the appearance and refinement of regulatory mechanisms specific to each novel metabolic pathway and on (ii) the increased robustness acquired by the whole metabolic network, i.e. its ability to respond to environmental changes or to "knock-out" mutations falling within biosynthetic genes (while maintaining unchanged biosynthetic activity). In fact, the removal of one of the multifunctional enzymes belonging to the ancestral common pathway of lysine, leucine and arginine (Figure 6) would lead to auxotrophy to all of the three amino acids. On the contrary, mutations silencing one of the extant, specific biosynthetic genes may generate auxotrophy only for the amino acid whose biosynthetic route has been interrupted, whereas all the others may still be synthesized. In this latter case, the presence of a paralogous copy may still account for the reaction catalyzed by the product of the silenced gene, permitting the biosynthesis of the corresponding amino acid and avoiding the occurrence of any other auxotrophy. In this context the *dapC* and *argD* genes (Figure 14b, 4c) represent a paradigmatic example. In *E. coli* (Figure 4c) the *argD* product exhibits both N-acetyl-ornithine and N-succinyl-L,L-diaminopimelate aminotransferase activities [30] whereas in *C. glutamicum* two distinct enzymes, DapC and ArgDrespectively, perform these reactions [38]. Although the genetic inactivation of *E. coli argD* and the consequences of this disruption are still under investigation, it is likely that it may generate auxotrophy to both diaminopimelate (and lysine) and L-arginine [30]. On the contrary, the simultaneous deletion of the *dapC* gene and *ddh* in *C. glutamicum* does not generate auxotrophy to lysine, suggesting that the product of *argD* may substitute the DapC function [38]. Since the simultaneous deletion of *dapC*, *ddh* and *argD* genes from *C. glutamicum* does not affect its growth, the existence of another aminotransferase able to

substitute ArgD and DapC, has been invoked [38]. This, in turn, is in agreement with the well-established low substrate specificity of aminotransferases [39], that may allow the enzymes carrying this activity to serve several different metabolic pathways.

On the basis of gene distribution analysis, Velasco *et al.* [19] proposed that the DAP pathway appeared earlier than the AAA one. The model we have proposed is in disagreement with this view, in that it suggests that the AAA pathway is the oldest one and that both pathways might have been simultaneously (and transiently) present in the "LUCA community". Since most bacteria, with the only exception of *T. thermophilus* and *D. radiodurans* (TD group), lack the AAA pathway, we suggest that, during the separation of Archaea from Bacteria, the latter may have lost the AAA pathway, maintaining the DAP pathway for the biosynthesis of lysine, LL-diaminopimelate, and meso-diaminopimelate (Figure 7e). The presence of the AAA pathway in the TD group might be the result of an event of horizontal gene transfer between an ancestor of *P. horikoshii* and the ancestor of the (micro)organisms belonging to the TD group as suggested by Nishida [9]. Alternatively, *T. thermophilus* and *D. radiodurans* have maintained the AAA pathway. The model proposed fits with the metabolic schemes shown in Figures 2, 3, 4 and the timing of some duplication events is supported by the phylogenetic analysis of triads of paralogous genes.

The distribution of the AAA and the DAP pathway in Archaea defies a simple explanation since all currently known pathways for lysine biosynthesis in Bacteria and Eucarya exist also in the domain of Archaea [23]. Hence, in the case of Archaea, a lineage specific maintenance of the DAP or the AAA pathway for lysine biosynthesis seems to be the most reliable hypothesis.

Lastly, on the basis of the scenario depicted in this work, the evolutionary history of lysine, leucine, and arginine biosynthetic routes might supply useful hints to disclose the origin and the assembly of bacterial cell wall. In the extant bacteria, the rigidity of the cell wall is due to a huge macromolecule containing acylated amino sugars and three to six different amino acids. This heteropolymer, the peptidoglycan, is built out of glycan strands cross linked through short peptides. In contrast to the uniform structure of the glycan, the peptide moiety reveals considerable variations. The greatest variation occurs at position 3 of the peptide moiety, where usually a diamino acid is found. The most widely distributed diamino acid is meso-diaminopimelate. It is present in gram-negative bacteria and in many other (micro)organisms, such as some species of bacilli, clostridia, lactobacilli, corynebacteria, propionibacteria, actinomycetales, myxobacteriales, rickettsiae, and cyanobacteria. Some studies have shown that the L-asymmetric carbon of meso-diaminopimelate is bound to the peptide subunit (see [40] and references therein). Studies on the amino acid composition and sequence of the peptidoglycans of different gram-negative bacteria have also shown that there is no great variation within this group. The peptidoglycan of these bacteria contains mainly meso-diaminopimelate, although D-glutamate and L-alanine can replace it.

The gram-positive bacteria reveal, contrary to the gram-negative organisms, a great variability in the composition and structural arrangement of their peptidoglycan since it can alternatively contain, at position 3, L-lysine, L-ornithine, D-glutamate, and LL-diaminopimelate.

Lastly, archaeal cells possess a fundamentally different type of peptidoglycan. Its glycan moiety contains L-talosaminuronic acid instead of muramic acid, and its peptide moiety lacks D-amino acids, but it is present L-lysine (see [40] and references therein).

As expected by the lack of DAP pathway, all those bacterial (micro)organisms biosynthesizing lysine through the AAA pathway (*T. thermophilus*, *D. radiodurans*) do not have diaminopimelic acid within their cell wall [41,42]. In these microorganisms diaminopimelic acid is replaced by ornithine. In addition to this, LL-diaminopimelate and meso-diaminopimelate, intermediaries of the DAP pathway (Figure 1), are compounds that can be found in the cell wall of extant Bacteria [40], whereas they are absent in the archaeal cell wall of Archaea [43]. Hence we suggest that the maintenance of the DAP pathway for the biosynthesis of lysine, at least in Bacteria, might be correlated with the appearance of the extant structure of their cell wall. The ability of inserting these molecules in the cell wall, might be an "invention" of Bacteria. In our opinion, the primordial cells had ornithine as a component of peptidoglycan. This is in agreement with the recent suggestion that the presence of ornithine within the cell wall might be an ancestral feature [36] whose biosynthesis, in the primordial organisms, might have occurred through the previously proposed common route (Figure 6).

This evolutionary step corresponds to the last one that is depicted in the evolutionary model that we have previously proposed (Figure 7f). This view might support the idea that the insertion of the meso-diaminopimelate and, probably, the appearance of the fully assembled DAP pathway might be a metabolic invention of Bacteria.

Methods

Sequence retrieval

Amino acid sequences were retrieved from GenBank and KEGG databases and were used to build a local database for BLASTp [44] probing that was performed using default

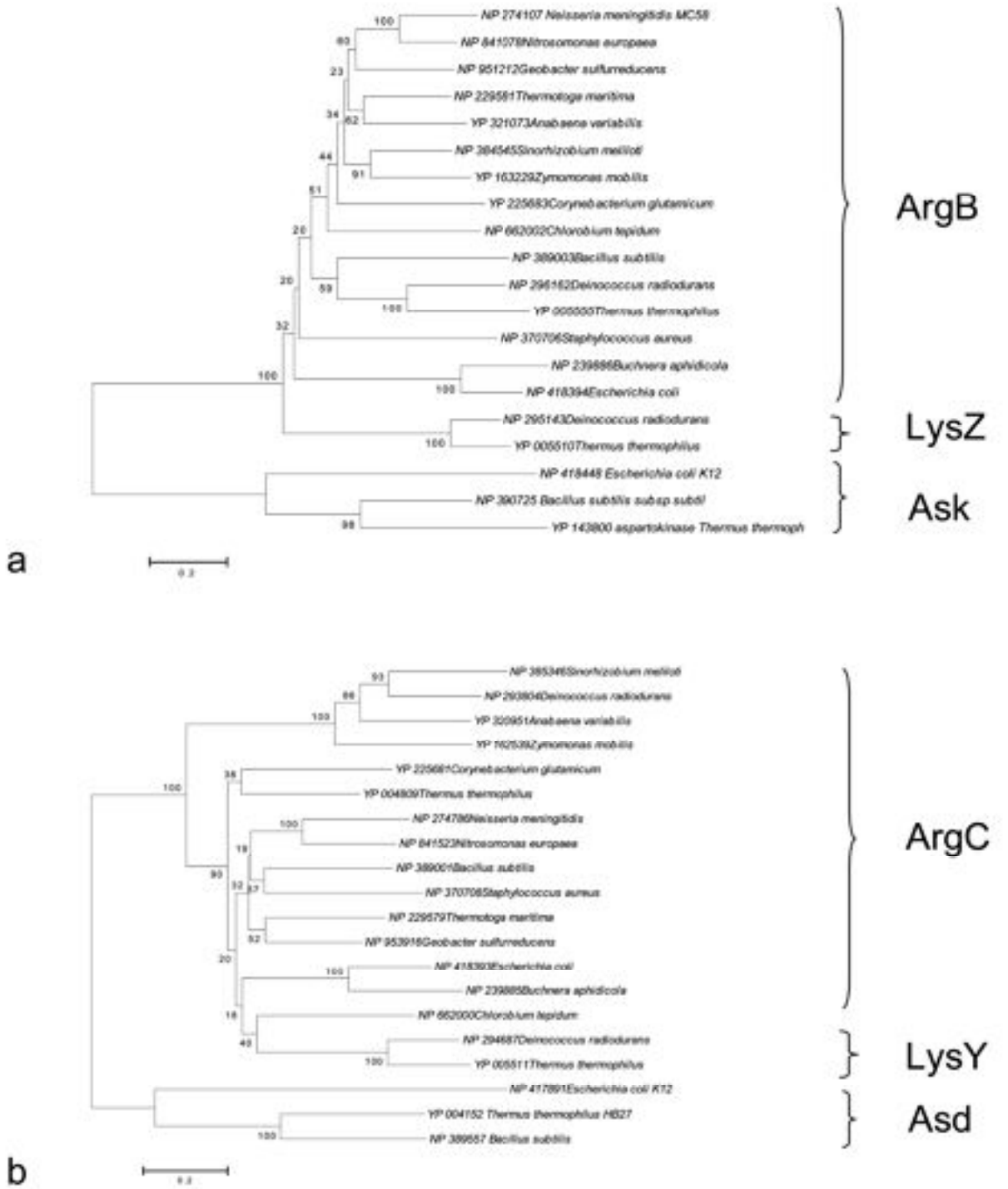


Figure 8
Phylogenetic trees. Phylogenetic trees (Neighbor Joining, 2250 Bootstrap Replicates, Complete Deletion, Poisson Correction) constructed with the sequences of LysZ, ArgB, Ask **(a)** and LysY, ArgC, Asd **(b)**.

parameters. Orthologs identification was achieved according to the bidirectional best-hit (BBH) criterion [45,46]. The relationship between gene *x* in genome A and gene *y* in genome B is called best-best hit when *x* is the best hit of query *y* against all genes in A and *vice versa*, and it is often used as an operational definition of ortholog [45,47].

Sequence alignment

The ClustalW [37] program in the BioEdit [48] package was used to perform pairwise and multiple amino acid sequences alignments.

Phylogenetic analysis

Phylogenetic trees were obtained with Mega 3 software [49] using the Neighbor-Joining (NJ) and the Minimum Evolution (ME) methods.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors equally contributed to the preparation of the final version of the manuscript; MF performed the analyses during its PhD thesis under the supervision of Prof. RF.

Acknowledgements

This article has been published as part of *BMC Evolutionary Biology* Volume 7 Supplement 2, 2007: Second Congress of Italian Evolutionary Biologists (First Congress of the Italian Society for Evolutionary Biology). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2148/7?issue=S2

References

1. Fani R: **Gene duplication and gene loading.** In *Microbial evolution: gene establishment, survival and exchange* Edited by: *Miller RV, Day MJ. Washington DC: ASM Press*; 2004:67-81.
2. Jensen RA: **Enzyme Recruitment in Evolution of New Function.** *Annual Review of Microbiology* 1976, **30**:409-425.
3. Ohno S: *Evolution by Gene Duplication New York: Springer-Verlag*; 1970.
4. Vogel HJ: **Lysine biosynthesis and evolution.** In *Evolving genes and proteins* Edited by: *Bryson V, Vogel HJ. New York Academic Press*; 1965:25-40.
5. Umbarger HE: **Amino acid biosynthesis and its regulation.** *Annu Rev Biochem* 1978, **47**:532-606.
6. Bhattacharjee JK: **Evolution of α-aminoadipate pathway for the synthesis of lysine in fungi.** In *The Evolution of Metabolic Function* Edited by: *Mortlock RP. Boca Raton, Florida CRC Press*; 1992:47-80.
7. Kosuge T, Hoshino T: **Lysine is synthesized through the α-aminoadipate pathway in *Thermus thermophilus*.** *FEMS Microbiol Lett* 1998, **169**:361-367.
8. Nishida H, Nishiyama M, Kobashi N, Kosuge T, Hoshino T, Yamane H: **A Prokaryotic Gene Cluster Involved in Synthesis of Lysine through the Amino Adipate Pathway: A Key to the Evolution of Amino Acid Biosynthesis.** *Genome Res* 1999, **9**:1175-1183.
9. Nishida H: **Distribution of genes for lysine biosynthesis through the aminoadipate pathway among prokaryotic genomes.** *Bioinformatics* 2001, **17**:189-191.
10. Schafer S, Paalme T, Vilu R, Fuchs G: **^{13}C-NMR study of acetate assimilation in *Thermoproteus neutrophilus*.** *Eur J Biochem* 1989, **186**:695-700.
11. Ragan MA: **Biochemical pathways and the phylogeny of the eukariotes.** In *The hierarchy of life* Edited by: *Fernhulm B, Bremer K, Jornvall H. New York: Elsevier*; 1989:145-160.
12. Matsuda M, Ogur M: **Separation and specificity of the yeast glutamate-α-ketoadipate transaminase.** *J Biol Chem* 1969, **244**:3352-3358.
13. Rowley B, Tucci AF: **Homoisocitric dehydrogenase from yeast.** *Arch Biochem Biophys* 1970, **141**:499-510.
14. Weidner G, Steffan B, Brakhage AA: **The Aspergillus nidulans lysF gene encodes homoaconitase, an enzyme involved in the fungus-specific lysine biosynthesis pathway.** *Mol Gen Genet* 1997, **255**:237-147.
15. Vogel HJ: **On biochemical evolution: lysine formation in higher plants.** *Proc Natl Acad Sci USA* 1959, **45**:1717-1721.
16. Léjohn HB: **Enzyme Regulation, Lysine Pathways and Cell Wall Structures as Indicators of Major Lines of Evolution in Fungi.** *Nature* 1971, **231**:164-168.
17. Cirillo JD, Weisbrod TR, Banerjee A, Bloom BR, Jacobs WR Jr: **Genetic determination of the meso-diaminopimelate biosynthetic pathway of mycobacteria.** *J Bacteriol* 1994, **176**:4424-4429.
18. Wehrmann A, Phillipp B, Sahm H, Eggeling L: **Different Modes of Diaminopimelate Synthesis and Their Role in Cell Wall Integrity: a Study with *Corynebacterium glutamicum*.** *J Bacteriol* 1998, **180**:3159-3165.
19. Velasco AM, Leguina JI, Lazcano A: **Molecular Evolution of the Lysine Biosynthetic Pathways.** *Journal of Molecular Evolution* 2002, **55**:445-449.
20. Fondi M, Brilli M, Fani R: **On the origin and evolution of biosynthetic pathways: integrating microarray data with structure and organization of the Common Pathway genes.** *BMC Bioformatics* 2007, **8(Suppl 1)**:S12.
21. Irvin SD, Bhattacharjee JK: **A Unique Fungal Lysine Biosynthesis Enzyme Shares a Common Ancestor with Tricarboxylic Acid Cycle and Leucine Biosynthetic Enzymes Found in Diverse Organisms.** *Journal of Molecular Evolution* 1998, **46**:401-408.
22. Miyazaki J, Kobashi N, Fujii T, Nishiyama M, Yamane H: **Characterization of a *lysK* gene as an *argE* homolog in Thermus thermophilus HB27.** *FEBS Lett* 2002, **512**:269-274.
23. Hochuli M, Patzelt H, Oesterhelt D, Wüthrich K, Szyperski T: **Amino Acid Biosynthesis in the Halophilic Archaeon *Haloarcula hispanica*.** *Journal of Bacteriology* 1999, **181**:3226-3237.
24. Bakhiet N, Forney FW, Stahly DP, Daniels L: **Lysine biosynthesis in Methanobacterium thermoautotrophicum is by the diaminopimelic acid pathway.** *Current Microbiology* 1984, **10**:195-198.
25. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb J-F, Adams KD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Weidman JF, Fuhrmann JL, Nguyen D, Utterback TR, Kelley JM, Peterson JD, Sadow PW, Hanna MC, Cotton MD, Roberts KM, Hurst MA, Kaine BP, Borodovsky M, Klenk H-P, Fraser CM, Smith HO, Woese CR, Venter J: **Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*.** *Science* 1996, **273**:1058-1073.
26. Klenk H-P, Clayton RA, Tomb J-F, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Peterson S, Reich CI, McNeil LK, Badger JH, Glodek A, Zhou L, Overbeek R, Gocayne JD, Weidman JF, McDonald L, Utterback T, Cotton MD, Spriggs T, Artiach P, Kaine BP, Sykes SM, Sadow PW, D'Andrea KP, Bowman C, Fujii C, Garland SA, Mason TM, Olsen GJ, Fraser CM, Smith HO, Woese CR, Venter JC: **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus.** *Nature* 1997, **390**:364-370.
27. Selkov E, Maltsev N, Olsen GJ, Overbeek R, Whitman WB: **A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data.** *Gene* 1997, **197**:GC11-26.
28. Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K: **Complete genome sequence of *Methanobacterium thermoautotrophicum deltaH*: Functional analysis and comparative genomics.** *J Bacteriol* 1997, **179**:7135-7155.
29. Caldovic L, Tuchman M: **N-acetylglutamate and its changing role through evolution.** *Biochem J* 2003, **372(2)**:279-290.

30. Ledwidge R, Blanchard JS: **The dual biosynthetic capability of N-acetylornithine aminotransferase in arginine and lysine biosynthesis.** *Biochemistry* 1999, **38**:3019-3024.
31. Bryan JK: **Synthesis of the aspartate family and branched chain amino acids.** In *The biochemistry of plants Volume 5*. Edited by: *Miflin BJ. New York: Academic Press*; 1980:402-452.
32. Hudson AO, Singh BK, Leustek T, Gilvarg C: **An LL-Diaminopimelate Aminotransferase Defines a Novel Variant of the Lysine Biosynthesis Pathway in Plants.** *Plant Physiology* 2006, **140**:292-301.
33. Miyazaki K: **Bifunctional isocitrate-homoisocitrate dehydrogenase: a missing link in the evolution of beta-decarboxylating dehydrogenase.** *Biochem Biophys Res Commun* 2005, **331**:341-346.
34. Sakai H, Vassylyeva MN, Matsuura T, Sekine S, Gotoh K, Nishiyama M, Terada T, Shirouzu M, Kuramitsu S, Vassylyev DG: **Crystal Structure of a Lysine Biosynthesis Enzyme, LysX, from Thermus thermophilus HB8.** *J Mol Biol* 2003, **332**:729-740.
35. Woese C: **The universal ancestor.** *PNAS* 1998, **95**:6854-6859.
36. Cavalier-Smith T: **Rooting the tree of life by transition analyses.** *Biology Direct* 2006, **1**:19.
37. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
38. Hartmann M, Tauch A, Eggeling L, Bathe B, Mockel B, Pühler A, Kalinowski J: **Identification and characterization of the last two unknown genes, dapC and dapF, in the succinylase branch of the L-lysine biosynthesis of Corynebacterium glutamicum.** *J Biotechnol* 2003, **104**:199-211.
39. Jensen RA, Gu W: **Evolutionary Recruitment of Biochemically Specialized Subdivisions of Family I within the Protein Superfamily of Aminotransferases.** *Journal of Bacteriology* 1996, **178(8)**:2161-2171.
40. Schleifer KH, Kandler O: **Peptidoglycan Types of Bacterial Cell Walls and their Taxonomic Implications.** *Bacteriological Reviews* 1972, **36**:407-477.
41. Quintela JC, Pittenauer E, Allmaier G, Aran V, Pedro MA: **Structure of peptidoglycan from *Thermus thermophilus* HB8.** *Journal of Bacteriology* 1995, **177(17)**:4947-4962.
42. Mkarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, Daly MJ: **Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics.** *Microbiol Mol Biol Rev* 2001, **65**:44-79.
43. Kandler O, Konig H: **Cell wall polymers in Archaea (Archaebacteria).** *Cell Mol Life Sci* 1998, **54**:305-308.
44. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Nucleic Acid Res* 1997, **25**:3389-3402.
45. Bono H, Ogata H, Goto S, Kanehisa M: **Reconstruction of Amino Acid Biosynthesis Pathways from the Complete Genome Sequence.** *Genome Res* 1998, **8**:203-210.
46. Uchiyama I: **Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes.** *Nucleic Acids Research* 2006, **34**:647-658.
47. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet.** *Nucleic Acids Res* 2002, **30**:42-46.
48. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41**:95-98.
49. Kumar S, Tamura K, Nei M: **Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Briefings in Bioinformatics* 2004, **5**:150-163.

4.2 On the origin and evolution of the Common Pathway of lysine, threonine and methionine

In this work, data concerning gene structure, organization, phylogeny, distribution and microarray experiments were integrated, in order to depict a model for the evolution of *ask* and *hom*, the two genes representing the Common Pathway (CP) of lysine, threonine and methionine. In *Escherichia coli* three different aspartokinases (AKI, AKII, AKIII, the products of *thrA*, *metL* and *lysC*, respectively) can perform the first step of the CP. Moreover, two of them (AKI and AKII) are bifunctional, carrying also homoserine dehydrogenasic activity (*hom* product). The second step of the CP is catalyzed by a single aspartate semialdehyde dehydrogenase (ASDH, the product of *asd*). Thus, in the CP of *E.coli* while a single copy of ASDH performs the same reaction for three different metabolic routes, three different AKs perform a unique step. Why and how such a situation did emerge and maintain? How is it correlated to the different regulatory mechanisms acting on these genes? The aim of this work was to trace the evolutionary pathway leading to the extant scenario in proteobacteria. Analyses revealed that the presence of multiple copies of these genes and their fusion events are restricted to the γ-subdivision of proteobacteria. Furthermore, the appearance of fused genes paralleled the assembly of operons of different sizes, suggesting a strong correlation between the structure and organization of these genes. A statistic analysis of microarray data retrieved from experiments carried out on *Escherichia coli* and *Pseudomonas aeruginosa* was also performed.

BMC Bioinformatics

BioMed Central

Research

Open Access

On the origin and evolution of biosynthetic pathways: integrating microarray data with structure and organization of the Common Pathway genes

Marco Fondi, Matteo Brilli and Renato Fani*

Address: Dipartimento di Biologia Animale e Genetica, Università di Firenze, Via Romana 17\19, Firenze, Italy

Email: Marco Fondi - marco.fondi@unifi.it; Matteo Brilli - matteo.brilli@dbag.unifi.it; Renato Fani* - renato.fani@unifi.it

* Corresponding author

from Italian Society of Bioinformatics (BITS): Annual Meeting 2006
Bologna, Italy. 28–29 April, 2006

Published: 8 March 2007

BMC Bioinformatics 2007, **8**(Suppl 1):S12 doi:10.1186/1471-2105-8-S1-S12

This article is available from: http://www.biomedcentral.com/1471-2105/8/S1/S12

Abstract

Background: The lysine, threonine, and methionine biosynthetic pathways share the three initial enzymatic steps, which are referred to as the Common Pathway (CP). In *Escherichia coli* three different aspartokinases (AKI, AKII, AKIII, the products of *thrA*, *metL* and *lysC*, respectively) can perform the first step of the CP. Moreover, two of them (AKI and AKII) are bifunctional, carrying also homoserine dehydrogenasic activity (*hom* product). The second step of the CP is catalyzed by a single aspartate semialdehyde dehydrogenase (ASDH, the product of *asd*). Thus, in the CP of *E. coli* while a single copy of ASDH performs the same reaction for three different metabolic routes, three different AKs perfom a unique step. Why and how such a situation did emerge and maintain? How is it correlated to the different regulatory mechanisms acting on these genes? The aim of this work was to trace the evolutionary pathway leading to the extant scenario in proteobacteria.

Results: The analysis of the structure, organization, phylogeny, and distribution of *ask* and *hom* genes revealed that the presence of multiple copies of these genes and their fusion events are restricted to the γ-subdivision of proteobacteria. This allowed us to depict a model to explain the evolution of *ask* and *hom* according to which the fused genes are the outcome of a cascade of gene duplication and fusion events that can be traced in the ancestor of γ-proteobacteria. *Moreover*, the appearance of fused genes paralleled the assembly of operons of different sizes, suggesting a strong correlation between the structure and organization of these genes. A statistic analysis of microarray data retrieved from experiments carried out on *E. coli* and *Pseudomonas aeruginosa* was also performed.

Conclusion: The integration of data concerning gene structure, organization, phylogeny, distribution, and microarray experiments allowed us to depict a model for the evolution of *ask* and *hom* genes in proteobacteria and to suggest a biological significance for the extant scenario.

Background

The metabolic routes leading to the synthesis of lysine\diaminopimelic acid, methionine and threonine\isoleucine are closely interconnected forming a complex system, three steps of which represent the so-called Common Pathway (CP) [1] (Figure 1). The first of them is the phosphorylation of aspartate, carried out by an aspartokinase (AK, the product of the *ask* gene) leading to β-aspartyl-phosphate, which, in turn, is oxidised by an aspartate semialdehyde dehydrogenase (ASDH, the enzyme encoded by *asd*) to aspartate semialdehyde that, finally, may be transformed either into dihydrodipicolinate, the precursor of diaminopimelic acid and lysine, by dihydrodipicolinate synthase (coded for by *dapA*) or homoserine by homoserine dehydrogenase (HD, encoded by *hom*). Homoserine can be then channeled towards threonine and/or methionine biosyntheses. From an evolutionary point of view, the genes coding for these three enzymes are particularly interesting, since at least two different molecular mechanisms, i.e. paralogous gene duplication and gene fusion, appeared to have played a key role in their origin and evolution. In addition to this, in some bacteria each CP step is catalyzed by enzymes coded for by single monofunctional genes, whereas in the enterobacterium *Escherichia coli* it has been shown [2] (Figure 1) that:

i) the first step of the CP can be performed by three different aspartokinases (AKI, AKII and AKIII);

ii) the second step is catalyzed by a monofunctional ASDH encoded by *lysC*; and, lastly,

iii) the third step is carried out by two different homoserine dehydrogenases, referred to as HDI and HDII, which are fused to two of the three AKs: AKI and AKII, respectively. These two bifunctional proteins are coded for by two genes, *thrA* and *metL*, respectively.

The expression of the two *E. coli* bifunctional proteins are differently regulated: threonine and isoleucine regulate the expression of *thrA*, and threonine controls both enzymatic activities by a negative feedback. The transcription of *metL* is repressed by methionine but no feedback inhibition, by methionine itself, has been observed on this enzyme. Finally, the expression of the gene coding for AKIII (*lysC*) and the activity of its product, are regulated in response to lysine concentration [2].

This particular structure pattern has raised the question of how and why it emerged in the course of evolution. On the basis of limited sequence data, Cassan et al. [3] proposed that the present-day bifunctional enzymes may have arisen from a fusion event involving the AK and the HD ancestral coding genes. The duplication of this bifunctional gene may have originated two redundant copies carrying both AK and HD activity. Another gene duplication event may have led to the formation of the three AK copies we observe nowadays. According to this model, the monofunctional AK could have emerged in two different ways: either by a partial gene duplication event involving only the AK activity coding region of the bifunctional genes, or by inactivation, as a result of accumulation of mutations, of the HD coding sequence. Thus, both paralogous gene duplication and gene fusion might have been responsible for shaping the CP. The importance of gene duplication in the course of evolution of genomes and metabolic pathways is well established, (see [4] and references therein): the production of two copies of a DNA sequences leads to an increase of genome size, and it also allows the rapid diversification of enzymatically catalyzed reactions, providing new material for the invention of new enzymatic properties and complex regulatory and developmental patterns. In addition to gene duplication, (see [4] and references therein), one of the major routes of gene evolution is the fusion of independent cistrons leading to bi- or multifunctional proteins [5-9]. Gene fusions provide a mechanism for the physical association of different catalytic domains or of catalytic and regulatory structures [5]. Fusions frequently involve genes coding for proteins functioning in a concerted manner, such as enzyme catalyzing sequential steps within a metabolic pathway [10]. Fusion of such catalytic centres likely promotes the channelling of intermediates that may be unstable and/or in low concentration [5]; this, in turn, requires that enzymes catalysing sequential reactions are colocalized within cell [11] and may (transiently) interact to form complexes that are termed metabolons [12]. The high fitness of gene fusions can also rely on the tight regulation of the expression of the fused domains. This might be the case of *metL* and *thrA*.

Thus, the CP might represent a very interesting model study to shed some light on the mechanisms driving the assembly of metabolic pathways and the refinement of regulatory networks. Nonetheless, in spite of the availability of several completely sequenced genomes and microarray data, neither a detailed analysis of the structure and organization of CP genes has been carried out nor any correlation of these data with expression (microarray) ones has been established until now. The aim of this work was to try to reconstruct the possible evolutionary and timing pathway(s) leading to the extant *ask* and *hom* genes, to analyse their phylogenetic distribution, to shed some light on the molecular mechanisms responsible for the assembly of the CP genes in bacteria and on the role that gene duplication(s), fusion(s) and clustering might have had in this context. To this purpose, the structure, organization and phylogenetic distribution of all the available proteobacterial *ask*, *hom*, and *asd* genes were analysed. Data

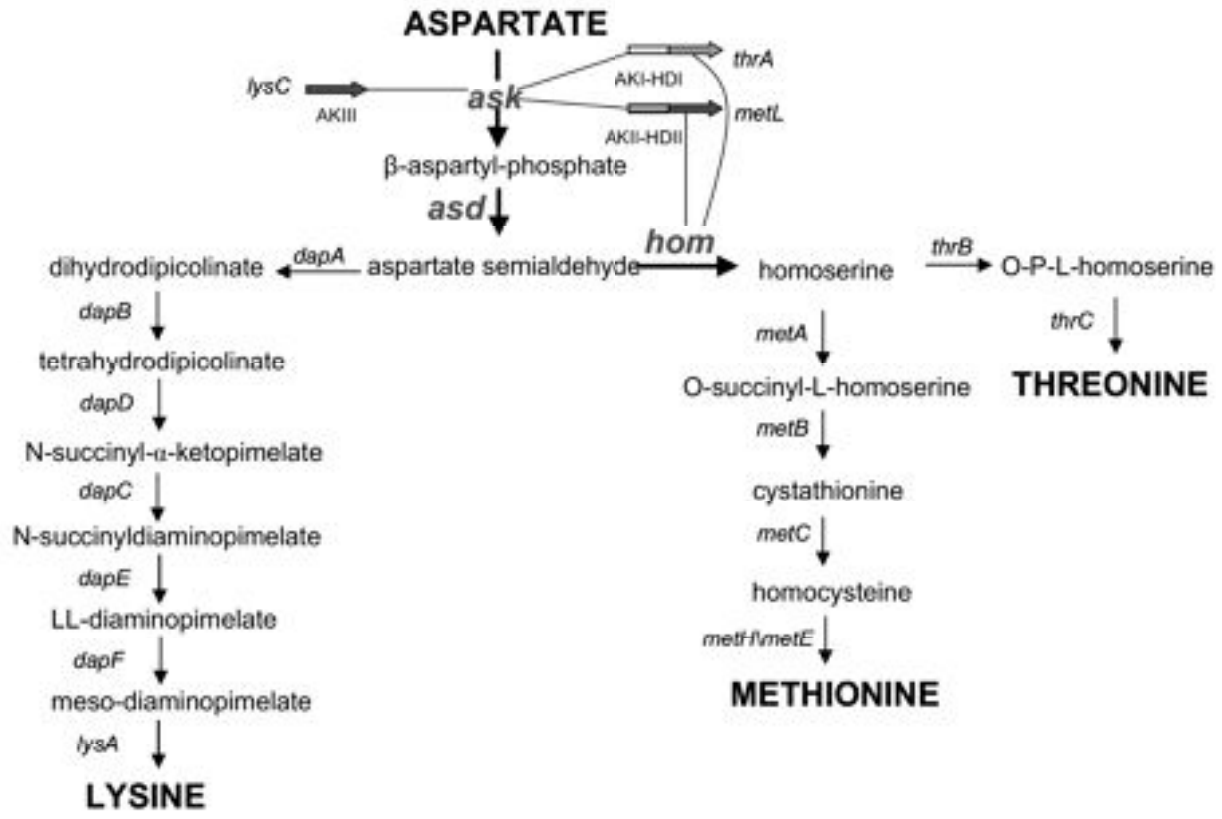


Figure 1
The aspartate pathway. Genes marked in red (*ask*, *asd*, and *hom*) constitute the Common Pathway [1].

obtained were integrated with expression data deriving from microarray analyses. We focused our attention on Proteobacteria for the following reasons: i) previous works [6,7,9] have shown that gene rearrangement events, such as gene duplication, fusion, and/or clustering have strongly influenced their evolution, ii) this phylogenetic branch includes the γ-subdivision, that is thought to be one of the most recent branching point among Bacteria and iii) they represent a good case-study since comprise organisms living in very different habitats (going from the deep-sea hydrothermal environments of the ε-subdivision to the roots of plants in the case of some α-proteobacteria), and with very different lifestyles, including endosymbionts and parasites.

Results and discussion

Structure and phylogenetic distribution of the genes coding for AK, ASDH and HD in Proteobacteria

The aminoacid sequences of the *E. coli* AK, ASDH, and HD sequences were used as a query to probe the protein database of completely sequenced proteobacterial genomes with the BLASTP option of BLAST program [13], in order to retrieve the most similar sequences. To this purpose 58 proteobacterial genomes were selected and, in most cases, only one strain for each species was taken into account. Data obtained are schematically reported in Figure 2, where a phylogenetic tree constructed using the RpoD sequences of the 58 proteobacteria is shown together with the number and the structure of all the retrieved AK, and HD coding genes. The *asd* genes were not included in Figure 2, since just one copy of this gene was retrieved from the 58 proteobacteria. The analysis of data reported in Figure 2 revealed that:

Figure 2
The structure of *ask* and *hom* genes. Phylogenetic tree constructed using the RpoD sequences (Neighbor Joining, 2250 Boostrap Replicates, Complete Deletion, Poisson Correction) of the 58 proteobacteria together with the number and the structure of all the retrieved *ask* and *hom* genes.

a) in all the α-, β- and δ\ε-proteobacterial genomes a single, monofunctional, stand-alone, copy of the gene coding for AK or HD was detected; moreover, neither duplicated copies nor fusion events involving these genes were detected.

b) multiple as well as fused copies of AK and HD were found only in γ-proteobacteria, where the scenario is (apparently) more complex and intriguing. Indeed, a variable structure and copy-number of genes coding for AK (1 to 5) and HD (1 to 2) can be observed. Moreover, there is an apparent increasing complexity concerning these genes that is parallel to the evolutionay branching of γ-proteobacteria, with enterobacteria and vibrionaceae showing the highest number of redundant and fused copies of AK and HD. This phylogenetic distribution strongly suggests that the duplication of AK coding genes and the fusion to HD apparently can be traced within γ-proteobacteria or soon after the divergence of the γ-proteobacterial ancestor from α-, β- and δ\ε-proteobacteria.

A model for the evolution of the AK and HD coding genes

On the basis of the phylogenetic distribution of standalone and bifunctional genes of the CP we propose a possible, plausible evolutionary and timing model explaining the extant scenario. The model, which is schematically reported in Figure 3, predicts that the proteobacterial ancestor possessed a single copy of *hom*, *ask* and *asd* genes. During evolution, this organization was maintained in proteobacteria belonging to the α-, β- and δ\ε-subdivisions. One of the cross-roads for the evolution of these genes is represented by the branching point between β- and γ-proteobacteria. It appears quite possible that, in the ancestor of γ-proteobacteria, a first duplication of the *ask* gene may have taken place, generating two redundant copies that underwent an evolutionary divergence. The finding that no bacterium (with the exception of *Vibrio* strains, see below) shows two copies of monofunctional *ask* genes, strongly suggests that this duplication event and its further fusion to *hom* might have occurred in a relatively short evolutionary time, giving raise to an ancestral bifunctional gene, which might have retained the function of the extant *metL* and *thrA*. This sort of "gene duplication-gene fusion coupling" is quite similar to that described recently for the evolution of γ-proteobacterial *hisN* and *hisB* histidine biosynthetic genes [6,7,9]. Finally, a paralogous duplication event of this bifunctional ancestor gene followed by evolutionary divergence (which very likely concerned with the regulatory mechanism, rather than the catalytic activity) led to the extant *metL* and *thrA* genes. On the basis of the phylogenetic distribution of the bifunctional genes (Figure 3), this "final" step might have occurred just before the separation between the "clusters" 1 and 2 of the γ-proteobacterial subdivision.

The biological significance of this cascade of duplication and fusion events might rely on the "patchwork" hypothesis on the origin and evolution of metabolic pathways [14]. According to this idea, metabolic pathways may have been assembled through the recruitment of primitive enzymes that could react with a wide range of chemically related substrates. Such relatively slow, unspecific enzymes may have been enabled primitive cells containing small genomes to overcome their limited coding capabilities [4]. Paralogous gene duplication event(s) followed by evolutionary divergence might have permitted the appearance of enzymes with an increase and narrow specificity and/or the diversification of function. In this way, an ancestral enzyme belonging to a given metabolic route, is "recruited" to serve a single or other (novel) pathways. Besides, it may permit the *evolution and refinement of regulatory mechanisms*coincident with the development of new pathways and/or the refinement of pre-existing ones.

In our opinion, the evolutionary model proposed here to explain the origin and evolution the extant *metL* and *thrA* genes is in full agreement with the Jensen hypothesis and the cascade of gene duplications and fusions involving *ask* and *hom* genes might actually represent a mechanism for the refinement of the feedback regulation mechanisms controlling the activity of the enzymes they code for.

Phylogenetic analysis

If the evolutionary model proposed here is correct, one should expect that the fused copies of AK (AKI and AKII) and HD (HDI and HDII) share a degree of sequence similarity higher than that exhibited with AKIII and HD, respectively, and cluster together in a phylogenetic tree. In order to check this hypothesis, the AK and HD aminoacid sequences were aligned using the program ClustalW [15] and the multialignments obtained used to draw the phylogenetic trees shown in Figure 4 and 5. The analysis of the AK tree (Figure 4) showed that all the α-, β- and δ\ε-proteobacterial sequences form a unique cluster separated from γ-proteobacterial ones. Besides, the γ-proteobacterial AKI, AKII, and AKIII sequences form three different and separated clusters with AKIII representing the root of the others. A similar situation can be observed in the HD tree (Figure 5): α-, β- and δ\ε-proteobacterial HD sequences form a distinct unique cluster, while HDI and HDII form two close clusters.

The topology of the phylogenetic trees obtained fits well with the evolutionary model proposed and indicates that horizotal gene transfer of these genes rarely occurred and did not strongly influenced the evolution of AK and HD domanis. However, even though the evolutionary model reported in Figure 3 is in agreement with gene structure and phylogenetic analyses, the following exceptions have to be explained:

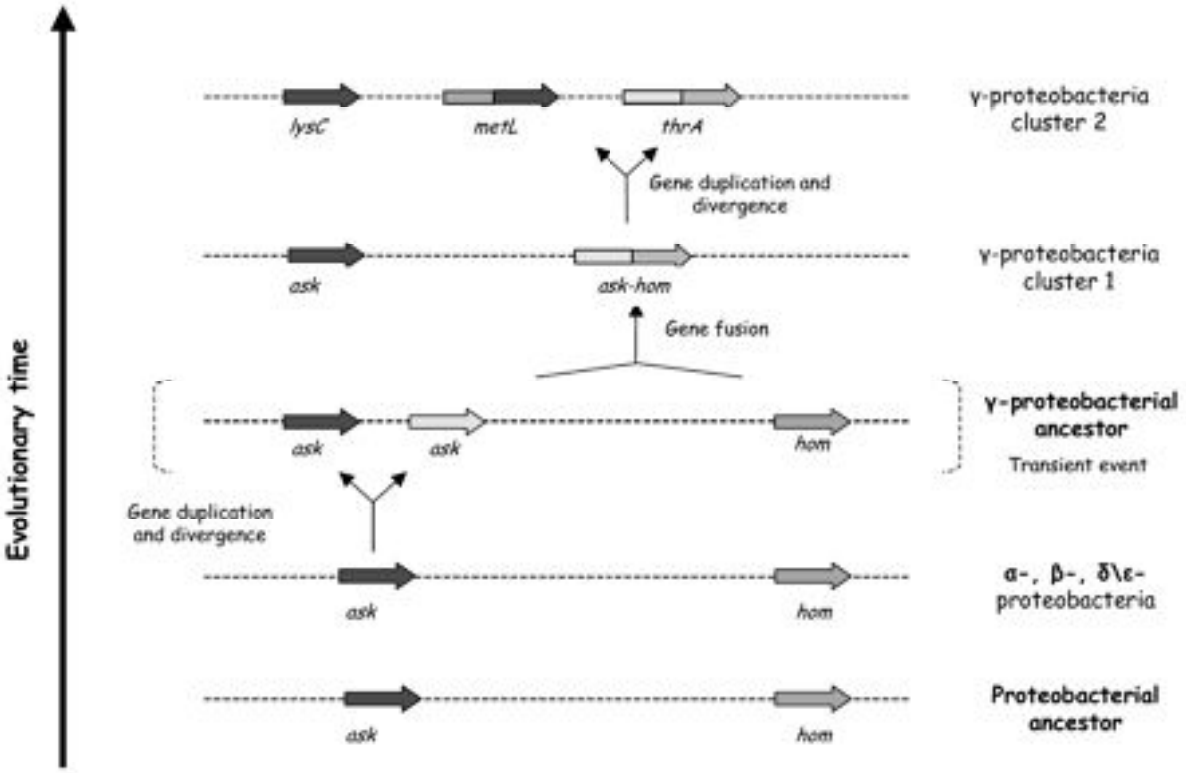


Figure 3
The evolutionary model. Evolutionary model proposed to explain the evolution of *ask* and *hom* genes in proteobacteria.

1) The absence of *lysC* and *metL* in a group of enterobacteria (*Buchnera aphidicola* strains, *Candidatus Blochmannia floridanus*, *Wigglesworthia glossinidia*) and in *Haemophilus influenzae*, the absence of bifunctional genes in *H. ducrey*, and the lack of *hom* in *Coxiella burnetii*, *Richetsia prowazekii*, *Wolbachia endosymbiont of Drosophila melanogaster* and *Bdellovibrio bacteriovorus*. This is very likely due to the absence of the correnspondingmetabolic route(s), which, in turn, is correlated to the parasitic lifestyle of these proteobacteria. Such a lifestyle may allow the bacteria to acquire essential compounds directly from the metabolic activities of their host and the adaptation to this environmental condition might have caused the loss of entire metabolic routes or part thereof.

2) The increase of the AK copies in *Vibrio* strains in respect to other γ-proteobacteria is probably related to the high genomic rearrangement rate typical of these species.

3) The absence of bifunctional *ask-hom* genes in *Pseudomonas* and *Methylococcus capsulatus* that, in spite of their taxonomical position within γ-proteobacteria, exhibit the same structural and organization pattern of bacteria belonging to the *α*-, *β*- and *δ*/*ε*-subdivisions. This is not an isolated example; in fact, the same situation has been recorded for other biosynthetic pathways, such as histidine biosynthesis [6,7]. The reason(s) of such structure and organization is still unclear.

4) The fusion of *ask* to *lysA* in *Xanthomonadaceae*, which represents an exception to this general model. In these bacteria the paralogous duplication of *ask* gene originated two copies, one of which fused to *hom*, whereas the other one underwent another fusion event with *lysA*, a gene coding coding for DAPDC activity). The biological significance of the last fusion might rely in the spatial colocalization of the products of the two modules and a faster feedback inhibition of the first enzyme (AK) by the end product of the pathway (lysine), whose last biosynthetic step is catalyzed by the enzyme coded for by *lysA*.

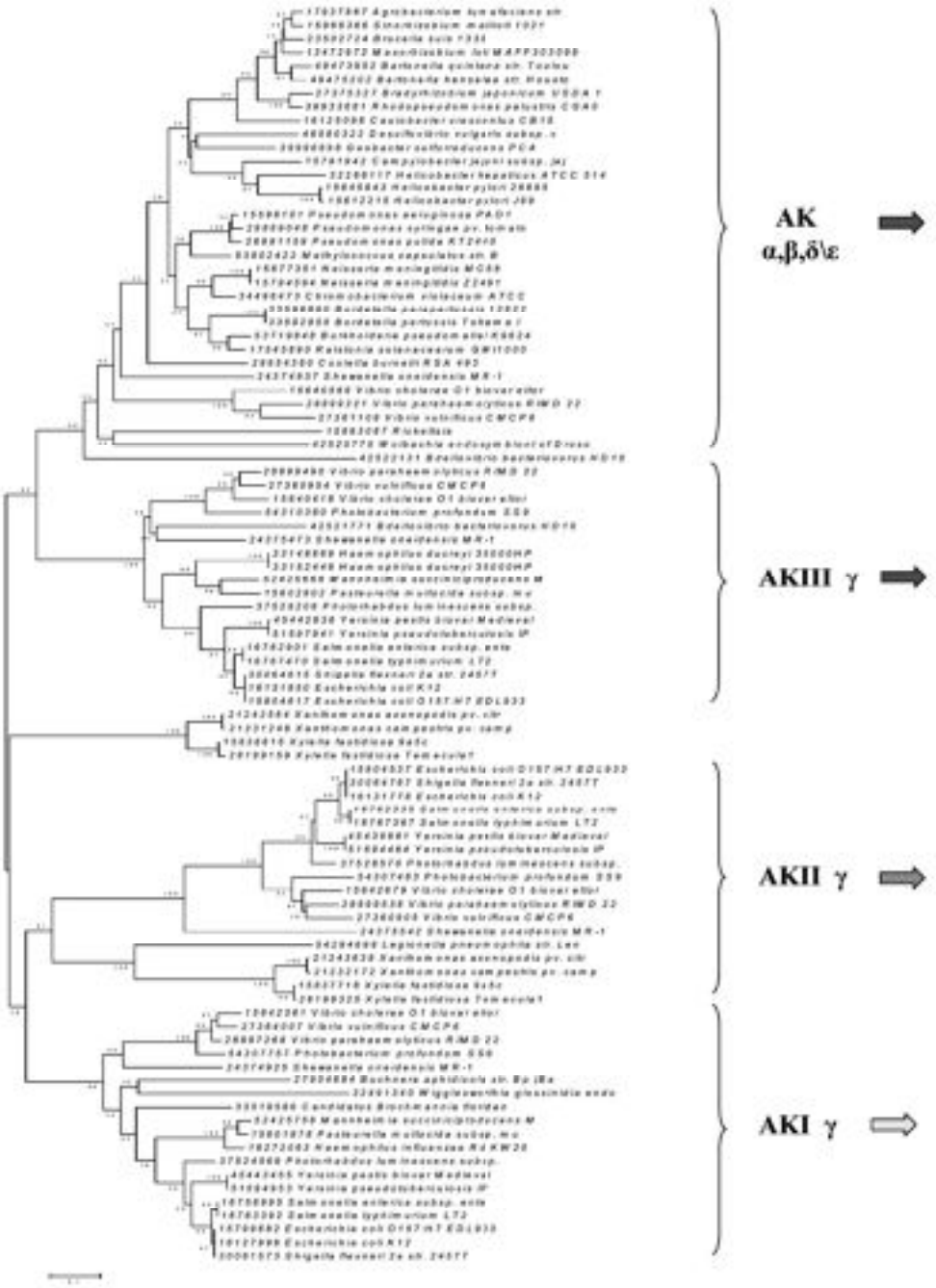


Figure 4
Phylogenetic tree of AK sequences. Phylogenetic trees (Neighbor Joining, 2250 Boostrap Replicates, Complete Deletion, Poisson Correction) constructed with all the retrieved sequences of AK.

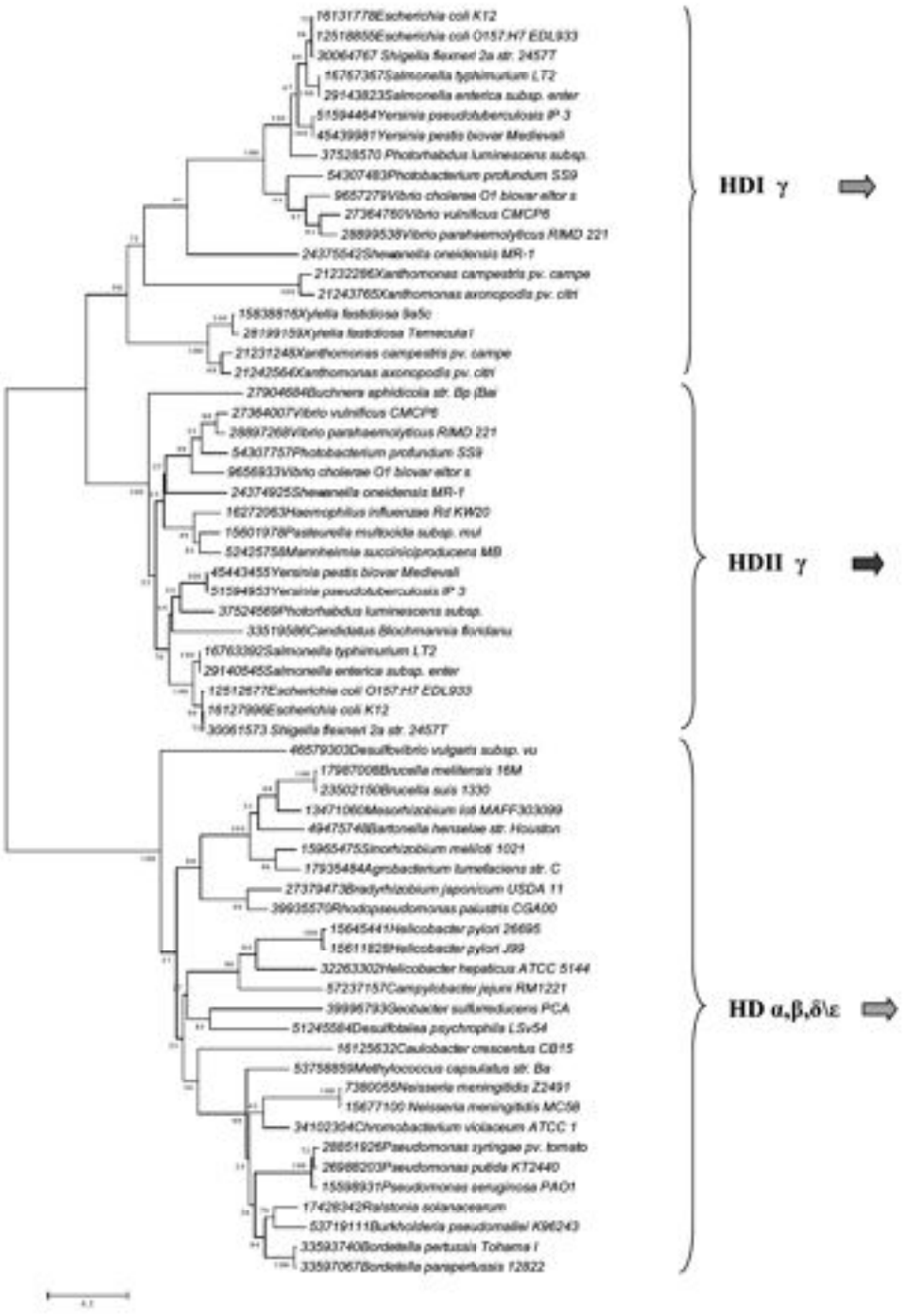


Figure 5
Phylogenetic tree of HD sequences. Phylogenetic trees (Neighbor Joining, 2250 Boostrap Replicates, Complete Deletion, Poisson Correction) constructed with all the retrieved sequences of HD.

Analysis of gene organization

If the model proposed and its biological significance is correct, i.e. that the duplication and fusion events, and the successive evolutionary divergence allowed the three copies of AKs and the two of HDs to narrow their specificity and to become increasingly more sensitive to specific regulatory signals, then it is plausible to assume that the ancestral copy of AK (AKIII) might serve different metabolic pathways and hence might have been under the control of multiple different regulatory signals (i.e. the availability of DAP, lysine, threonine, methionine etc). On the other hand, the expression of the bifunctional genes, *thrA* and *metL*, once they were channelled towards the biosynthesis of threonine and methionine, should have become increasingly more dependent on more specific signals (for example the concentration of the final product of that route). In general, it is plausible that once a "new" gene introgresses and becomes part of a pre-existing metabolic pathway, it will become co-regulated with the other genes belonging to the same metabolic pathway. In some cases, co-regulation of genes of the same biosynthetic route is achieved by organizing genes in operon structures, even though co-regulation may also be obtained by regulon construction. This is particularly true for fused genes; as reported in previous works, based on the analysis of the histidine biosynthetic pathway in γ-proteobacteria, the appearance of fused genes (specific for a single pathway) is often parallel to their presence within operons [6,7,9]. This raises the question whether the structure and distribution of duplicated and fused copies of *ask* and *hom* genes might somehow be correlated to their organization in the proteobacterial genome. Therefore, we analysed the organization of all the genes of the *lys*, *met* and *thr* biosynthesis in all the 58 proteobacteria. Data obtained revealed that:

1. Genes involved in the DAP\lysine biosynthesis are scattered throughout the chromosome(s) of all the 58 proteobacteria taken into account (data not shown).

2. In addition to *ask*, *asd* and *hom* genes, the other two genes involved in threonine biosynthesis (*thrB* and *thrC*) are scattered on the chromosome of bacteria belonging to α-, β- and δ\ε subdivisions (except *Bordetella* strains that own a *hom-thrC* operon) (Figure 6). The γ-proteobacterial scenario is completely different; according to the hypothesis mentioned above, in all of organisms possessing a bifunctional *thrA* gene, it is endowed within a three-cystronic operon, in the same relative gene order (*thrABC*), also suggesting that its construction should have been occurred once during evolution.

3. The organization of methionine biosynthetic genes in proteobacteria partly reflects that exhibited by *lys* or *thr* genes. In fact, in the α-, β- and δ\ε branches all the *met* biosynthetic genes are scattered on the chromosome(s) (Figure 7). This organization is also shared by γ-proteobacteria; the only exception is represented by the bifunctional *metL*, which is clustered with *metB* to form a bicistronic *metLB* operon.

Thus, no bifunctional gene of the CP is located outside operons. Data obtained strongly suggest that the production of genes coding for enzymes specific of a single metabolic pathway coincides with their presence within a polycistronic transcriptional unit that includes all (or at least some of) the other genes of that route. Concerning the timing of the operons construction, the comparative analysis of Figure 2, 5, and 6 revealed that the "gene duplication-gene fusion coupling" occurring in γ-proteobacteria appears to be coincident with gene clustering and the formation of operons of different length.

Analysis of microarray experiments data

In order to elucidate the correlation existing between the structure and organization of *lys*, *met*, and *thr* genes and their expression within the cell, we analyzed the microarray data from *E. coli* and *P. aeruginosa*, which show two different arrays of structure and organization of CP genes. Microarray data were downloaded as supplemental material to published papers (see Additional File 1: Additional References for the Expression compendium); only normalized and filtered data were used. Values were transformed into base 2 logarithm of the ratio of the wild type (untreated) / mutant (treated) expression levels, if not yet in that form.

For each of the three metabolic pathways we carried out a pairwise comparison of the expression pattern of each gene, by calculating the Pearson's correlation coefficient.

Data obtained are reported in Figure 8, whose analysis revealed:

1. A low co-regulation of the methionine biosynthetic genes (Figure 8a). Most of these genes are scarcely co-expressed, and they appeared to be expressed independently from each other. The fact that both *metL* and *metB* show very high correlation coefficient value in respect to the other *met* genes is in agreement with their operonic organization.

2. The three *E. coli thrABC* genes (Figure 8b) are highly co-expressed, with correlation coefficient > 0.84. This is in agreement with their organization in a compact operon.

3. The trend of the lysine pathway genes in the γ-proteobacterium *E. coli* (Figure 8c) is quite surprising; although the *lys* genes are scattered throughout the *E. coli* chromosome, they show a high degree of co-expression with cor-

Figure 6
Gene organization of threonine genes. Structure and organization of threonine biosynthetic genes of the 58 proteobacteria correlated with their phylogenetic position as established by RpoD analysis.

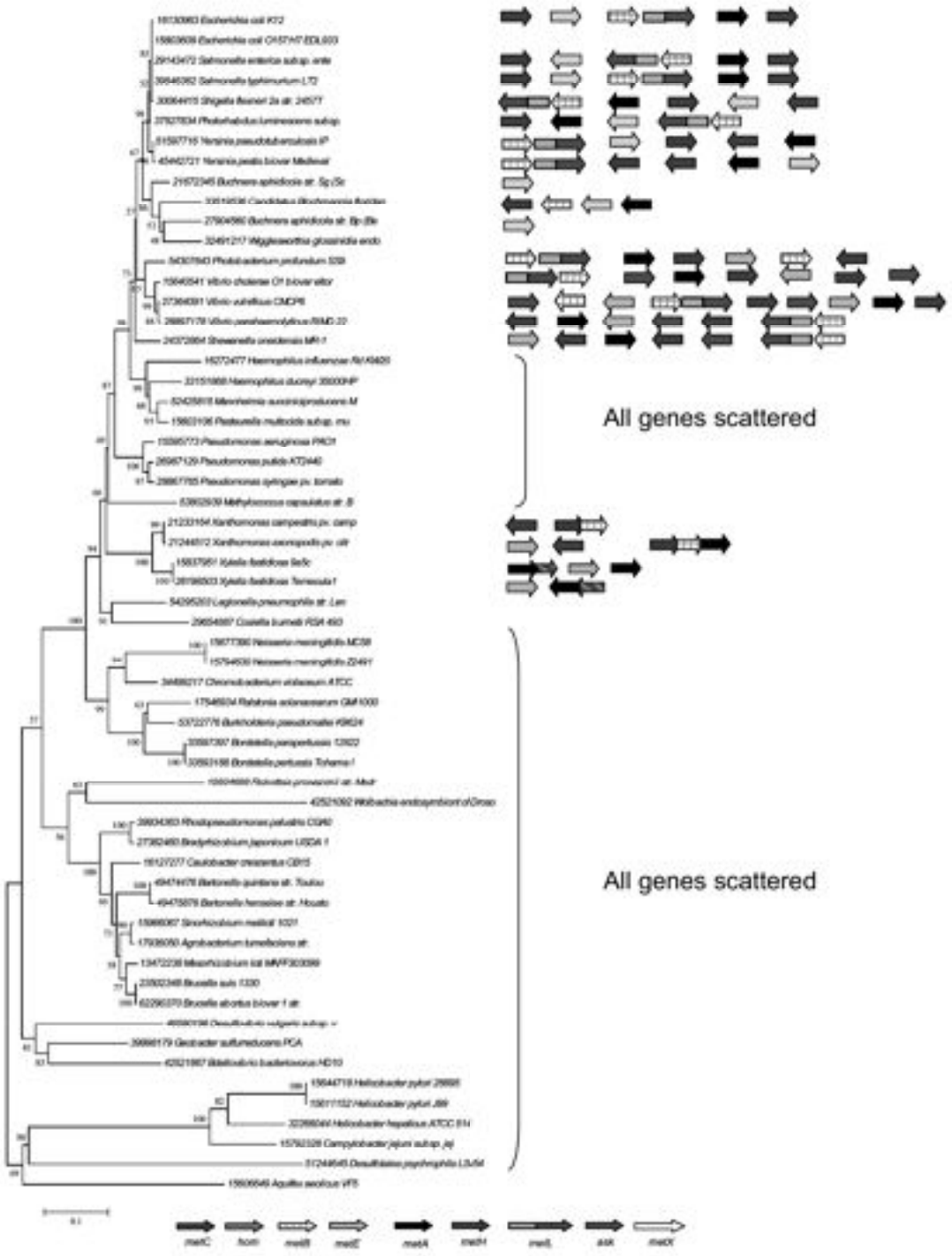


Figure 7
Gene organization of methionine genes. Structure and organization of methionine biosynthetic genes of the 58 proteobacteria correlated with their phylogenetic position as established by RpoD analysis.

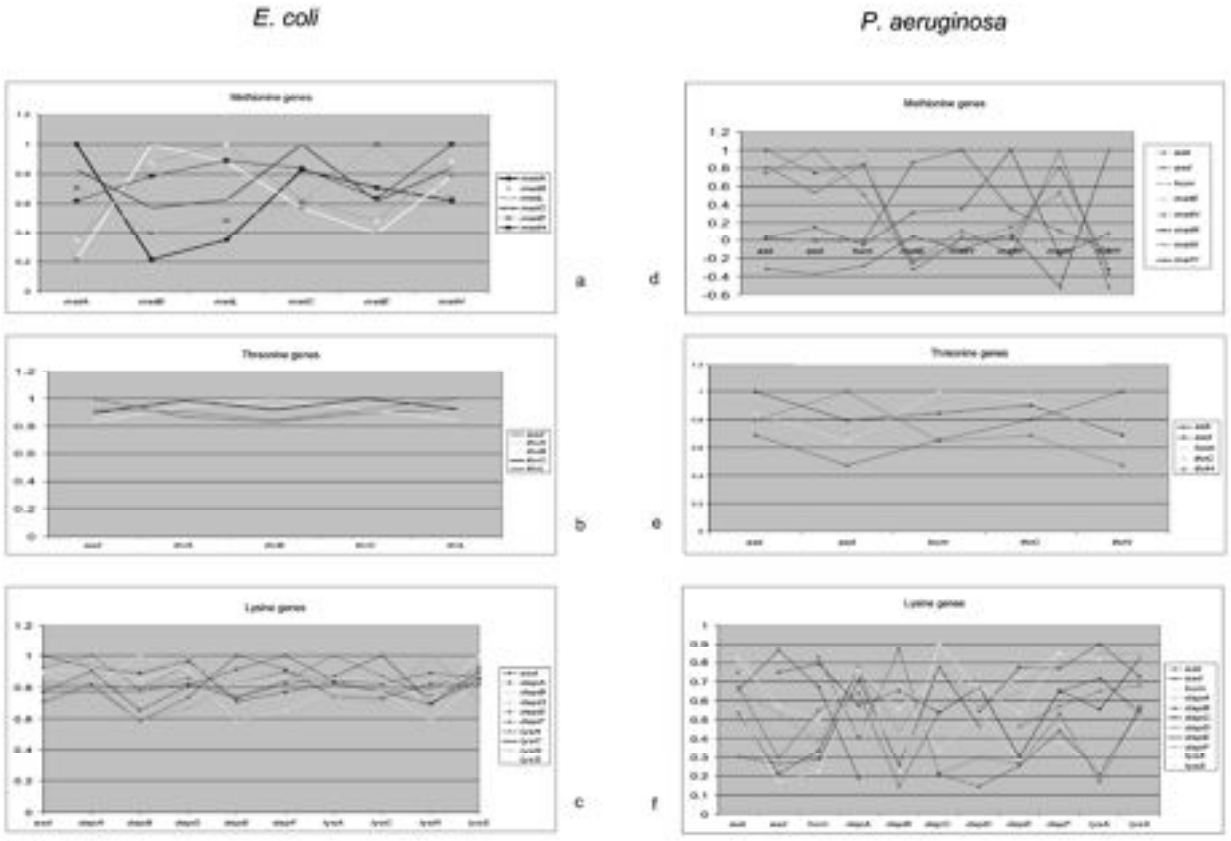


Figure 8
Microarray data analysis. Comparison between the expression pattern of each *met*, *lys*, *thr* gene of *E. coli* (a, b, c) and *P. aeruginosa* (d, e, f).

relation coefficient values often > 0.8. It is not clear how these genes can be highly co-expressed in the absence of an operonic organization. However, it is known [16] that lysine biosynthetic genes are regulated by the so-called LYS *element* (lysine-specific RNA element) located in their regulatory regions and able to repress or to allow their trascription in response to lysine concentration. The high coexpression pattern of lysine bosynthetic genes might be due to this mechanism.

The same analysis was carried out on lysine, methionine and threonine biosynthetic genes of *Pseudomonas aeruginosa*, whose structure and organization pattern is the same of the α-, β-, and δ\ε subdivision of proteobacteria. Data obtained (reported in Figure 8) showed that, overall, there is a low degree of co-expression between genes belonging to the same pathway; this is particularly pronounced for methionine, where in some cases, the correlation coefficient assumes negative values (Figure 8e), and lysine genes, whereas the *thr* biosynthetic genes were more correlated between them. The low degree of co-expression of *P. aeruginosa* genes is in agreement with their scattering on the bacterial genome.

Conclusion

In this work a likely model for the evolution of the genes involved in the common pathway (CP) is depicted, which is based on the comparative analysis of data concerning the structure, phylogenetic distribution, organization, phylogeny and expression of *ask* and *hom* genes in proteobacteria. The analysis of the structure of the CP genes gave a strong support to the hypothesis that at least two different molecular mechanisms played an important role in shaping the pathway, that is paralogous gene duplication(s) and gene fusion [17,4]. The analysis of *thr*, *met* and *lys* gene organization in different proteobacteria revealed

that several gene arrays exist within this phylogenetic lineage, with genes completely scattered throughout the genome, partially scattered/clustered, or strictly compacted. Even though different scenarios can be depicted for this different organization, i.e. the presence of scattered or clustered genes in the ancestor of proteobacteria, data reported in this work supported the first hypothesis. According to the model proposed, the ancestor of proteobacteria possessed monofunctional *hom*, *ask*, and *asd* genes scattered throughout the genome. The extant multiple and fused copies of *ask* and *hom* genes are the outcome of a cascade of paralogous gene duplication and fusion events, which led to the appearance of bifunctional enzymes catalyzing the same metabolic steps, but "sensing" different regulatory signals.

The evolutionary history of the CP genes gives another important support to the Jensen's hypothesis on the origin and evolution of metabolic pathways [14], strengthening the idea that gene duplication, gene fusion and recruitment of genes encoding enzyme with a broad range of substrate specificity played a crucial role in the assembly of biosynthetic pathways and in the appearance of new and/or more sophisticated regulatory networks [4,9]. Indeed, the biological significance of the presence of multiple copies of *ask* and *hom* genes might rely on the refinement of regulatory mechanisms allowing each *ask* copy to be regulated by specific signals, such as the availability of the end-product of the pathway.

The question of why the duplicated copies of *ask* fused to *hom* is rather intriguing. It is evident from their phylogenetic distribution that, once occurred, the fusion has been fixed; thus, it should have been evolutionary advantageous. Even though it cannot be *a priori* excluded, we do not favour the possibility that this fusion might permit the substrate tunnelling. It is possible that this gene fusion (and gene organization) resulted from both regulatory and metabolic constraints, for instance it might permit the spatial colocalization of their products and so a faster feedback inhibition of the first enzyme of the pathway, coded for by *ask*, by the product of *hom*.

The existence of the *thrA and metL* gene fusions in the genome of γ-proteobacteria is not an isolated example; additional gene fusions occurred in these genomes, such as those involving some histidine biosynthetic genes. It is worth of note that most of bifunctional proteins recognized to date are involved in metabolic pathways of the γ-subdivision of proteobacteria [18]. Even though there is no apparent reason to think that these organisms are more prone to gene fusions than any others, it is interesting that these gene fusions appeared to be parallel to the increasing compactness of some operons [9] or to their construction, as in the case of the *thrABC* and *metLB* ones. Actually, the analysis of the organization of these genes revealed that all the *metL* and *thrA* genes are embedded within (compact) operons, whereas their monofunctional counterparts as well as the second CP gene, *asd*, are located outside gene clusters. This is not so surprising if we agree on the existence of unspecific enzymes that might serve different metabolic pathways. Indeed, it is plausible that the expression of a gene, whose product catalyses a chemical reaction leading to a product involved in different metabolic pathways should be constitutively expressed or controlled by multiple mechanisms rather than being controlled by mechanisms specific for a single route.

This is also in agreement with expression data retrieved from the available microarray data; in fact, the greater the scattering of genes belonging to the same pathway, the lower the degree of correlation between them.

If our model is correct, the building up of *thrABC* and *metLB* operons represents a recent invention of evolution (dated in the γ proteobacterial ancestor) and is apparently co-incident with the appearance of bifunctional *ask-hom* genes. The origin and evolution of operons is still under debate, and at least six different classes of models have been proposed to explain the existence of operons (see [9] and references therein); although different forces might have driven the assembly of operons, in our opinion the major ones were those enabling the *fused* genes to be coregulated finely and the protein coded for synthesized in the correct stoichiometric ratio.

Material and methods

Sequence retrieval

Amino acid sequences were retrieved from GenBank database. BLAST [13] probing of database was performed with the BLASTP option of this program using default parameters. Only those sequences retrieved at an E-value below the 0.05 threshold were taken into account.

Sequence alignment

The ClustalW [15] program in the BioEdit package was used to perform pairwise and multiple amino acid sequences alignments.

Phylogenetic trees construction

Phylogenetic trees were obtained with Mega 3 [19] software using the Neighbor-Joining (NJ) and the Minimum Evolution (ME) methods.

List of abbreviations

AKI, AKII, AKIII, Aspartokinase I, II, III; *askI* and *askII* can also be named as *thrA* and *metL*; ASHD, Aspartate semialdehyde dehydrogenase; DAPDC, meso-diaminopimelate decarboxylase; HD, homoserine dehydrogenase.

Authors' contributions

All authors equally contributed to the preparation of the final version of the manuscript; MF performed the analyses during its MS degree work under the supervision of Prof. RF.

Additional material

Additional File 1

Additional References for the Expression compendium. List of the references used to retrieve microarray experiments data.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-S1-S12-S1.pdf]

Acknowledgements

This article has been published as part of *BMC Bioinformatics* Volume 8, Supplement 1, 2007: Italian Society of Bioinformatics (BITS): Annual Meeting 2006. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/8?issue=S1.

References

1. Cohen GN: **The common pathway to lysine, methionine and threonine.** *Amino Acids: Biosynthesis and Genetic Regulation* 1983:141-147.
2. Patte JC: **Biosynthesis of threonine and lysine.** In *Escherichia coli and Salmonella typhimurium* Edited by: *Neidhardt FC. ASM Press, Washington, DC*; 1996:528-541.
3. Cassan M, Parsot C, Cohen GN, Patte JC: **Nucleotide sequence of *lysC* gene encoding the lysine-sensitive aspartokinase III of *Escherichia coli* K12. Evolutionary pathway leading to three isofunctional enzymes.** *J Biol Chem* 1986, **261(3)**:1052-1057.
4. Fani R: **Gene duplication and gene loading.** *Microbial evolution: gene establishment, survival, and exchange* 2004:67-81.
5. Jensen R: **Evolution of metabolic pathways in enteric bacteria.** In *Escherichia coli and Salmonella typhimurium* Edited by: *Neidhardt FC. ASM Press, Washington, DC*; 1996:2649-2662.
6. Brilli M, Fani R: **The origin and evolution of eucaryal *HIS7* genes: from metabolon to bifunctional proteins?** *Gene* 2004, **339**:149-160.
7. Brilli M, Fani R: **Molecular evolution of *hisB* genes.** *J Mol Evol* 2004, **58(2)**:225-237.
8. Xie G, Keyhani NO, Bonner CA, Jensen RA: **Ancient origin of the tryptophan operon and the dynamics of evolutionary change.** *Microbiol Mol Biol Rev* 2003, **67(3)**:303-342.
9. Fani R, Brilli M, Lió P: **The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon.** *J Mol Evol* 2005, **60(3)**:378-390.
10. Yanai I, Wolf YL, Koonin EV: **Evolution of gene fusions: horizontal gene transfer versus independent events.** *Genome Biol* 2002, **3(5)**:research0024-.
11. Mathews CK: **The cell-bag of enzymes or network of channels?** *J Bacteriol* 1993, **175(20)**:6377-6381.
12. Srere PA: **Complexes of sequential metabolic enzymes.** *Ann Rev Biochem* 1987, **56**:89-124.
13. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25**:3389-3402.
14. Jensen RA: **Enzyme recruitment in evolution of new function.** *Annu Rev Microbiol* 1976, **30**:409-425.
15. Thompson JD, Higgins DG, Gibson TJ: **Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucl Acids Res* 1994, **22**:4673-4680.
16. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch?** *Nucl Acids Res* 2003, **31(23)**:6748-6757.
17. Fani R, Lio P, Lazcano A: **Molecular evolution of the histidine biosynthetic pathway.** *J Mol Evol* 1995, **41(6)**:760-774.
18. Ahmad S, Weinburg WG, Jensen RA: **Evolution of aromatic amino acid biosynthesis and application to the fine-tuned phylogenetic positioning of enteric bacteria.** *J Bacteriol* 1990, **172(2)**:1051-1061.
19. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Brief Bioinform* 2004, **5(2)**:150-163.

4.3 Conclusions

The comparative analysis of the genes involved in the biosynthesis of lysine performed in this chapter revealed that the extant metabolic "grids" and their interrelationships might be the outcome of a cascade of duplication of ancestral genes that, according to the patchwork hypothesis, coded for unspecific enzymes able to react with a wide range of substrates. Firstly, a likely model for the evolution of genes involved in the biosynthesis of lysine, leucine, and arginine can be depicted. The model proposed is based on the analysis of the structure of these biosynthetic routes and the phylogenetic distribution of their genes. The phylogenetic analysis performed allowed us also to determine a possible relative timing of the appearance of genes that are involved in the extant lysine (DAP) and arginine biosynthetic routes. This analysis gave a strong support to the hypothesis that extensive gene duplication events played a key role in shaping the extant biosynthetic routes of lysine, leucine and arginine. According to the model proposed in this work a common metabolic pathway for the biosynthesis of these three amino acids predated the appearance of the last universal common ancestor. This ancestral metabolic route was probably composed of a set of unspecific enzymes able to react with chemically related substrates interconnecting different biosynthetic routes. The occurrence of multiple gene duplication events would have led to the appearance of specific metabolic pathways responsible for the biosynthesis of each amino acid. The evolutionary history of lysine, leucine, and arginine biosynthetic routes strongly supports the hypothesis on the origin and evolution of metabolic pathways proposed by Jensen, strengthening the idea that the gene duplication and the recruitment of genes encoding enzymes with a broad substrate specificity played a key role in the assembly of primitive metabolic routes. Further support to this idea was provided also when examining the evolutionary history of the CP genes. According to the model proposed, in fact, the ancestor of proteobacteria possessed monofunctional *hom*, *ask*, and *asd* genes scattered throughout the genome. The extant multiple and fused copies of *ask* and *hom* genes are the outcome of a cascade of paralogous gene duplication and fusion events, which led to the appearance of bifunctional enzymes catalyzing the same metabolic steps, but "sensing" different regulatory signals.

Chapter 5

On the origin and evolution of nitrogen fixation genes

The building up of nitrogen fixation represented a metabolic innovation that is not only crucial for the extant life, but played a key role in the early stages of evolution as the prebiotic supply of all nitrogen sources decreased. The ancestral *nif* pathway might have originated in the early stages evolution and the entire process might have been carried out by a limited number of genes coding for multifunctional, unspecific enzymes that could react with a wide range of chemically related substrates. These primordial enzymes were responsible for the interconnection of nitrogen fixation to other metabolic routes, such as bacterial photosynthesis and biosynthesis of leucine/lysine. Gene and operon duplications, gene recruitment and elongation events and an extensive horizontal transfer of *nif* genes shaped the entire pathway that was likely completely assembled before the appearance of the Last Universal Common Ancestor. Data reported in this chapter were obtained performing a phylogenomic analysis, based on a computational biology approach, of the available sequences of proteins involved in nitrogen fixation and propose a model for the major evolutionary steps of nitrogen fixation process. Lastly the applied strategy allowed to map on the species phylogeny tree the appearance of several genes related to nitrogen fixation in several different bacterial lineages. This, in turn, suggests that, their appearance (and/or recruitment) during microbial evolution, probably allowed the refinement of nitrogen fixation process, initially carried out by a limited number of genes.

5.1 The Nitrogen Cycle

On the planetary scale the biogeochemical N cycle has suffered major anthropogenic alterations in the last decades shifting the priorities from boosting food production to control large scale environmental changes [Galloway *et al.*, 2008]. Half of the fixed nitrogen entering Earth ecosystems is produced via the Haber-Bosch process and cultivation of nitrogen fixing crops. Furthermore, reactive nitrogen is also produced by fossil and bio-fuels combustion. These inputs of reactive nitrogen might alter the terrestrial and marine N cycles [Deutsch *et al.*, 2007; Houlton *et al.*, 2008] as well as interconnected biogeochemical cycles, such as those related to carbon and phosphorus [Gruber & Galloway,

2008]. In the absence of human perturbations, the nitrogen cycle is the result of geological time-scale abiotic processes including NH_4^+ production from N_2 [Wachtershauser, 2007], combustion of N_2 to nitrate [Mancinelli & McKay, 1988a; Navarro-Gonzalez *et al.*, 2001], mineralization [McLain & Martens, 2005] and biologically driven metabolic reactions. The abiotic production of fixed nitrogen, which is mainly due to lightening discharge, is ten-fold lower than microbial production [Falkowski, 1997]. It has been postulated that abiotic fixed nitrogen was limiting in the early Earth [Kasting & Siefert, 2001], a condition that might have favored an early appearance of microbial N_2 fixation [Raymond *et al.*, 2004]. Schematically, the microbial driven nitrogen cycle comprises three steps (Figure 5.1).: i) the fixation of the atmospheric N_2 to ammonia (NH_4^+); ii) the stepwise oxidation of ammonia to nitrite and of nitrite to nitrate; iii) the denitrification of nitrite and nitrate to gaseous dinitrogen through anaerobic respiration in anoxic environment (complete denitrification) or the detoxifying reduction of nitrite to NO in aerobic environment (incomplete or nitrifier denitrification). Nevertheless, a complete picture of the microbial nitrogen cycle must take into account other relevant processes and the list of prokaryotic players in the biogeochemical N fluxes is continuously increasing.

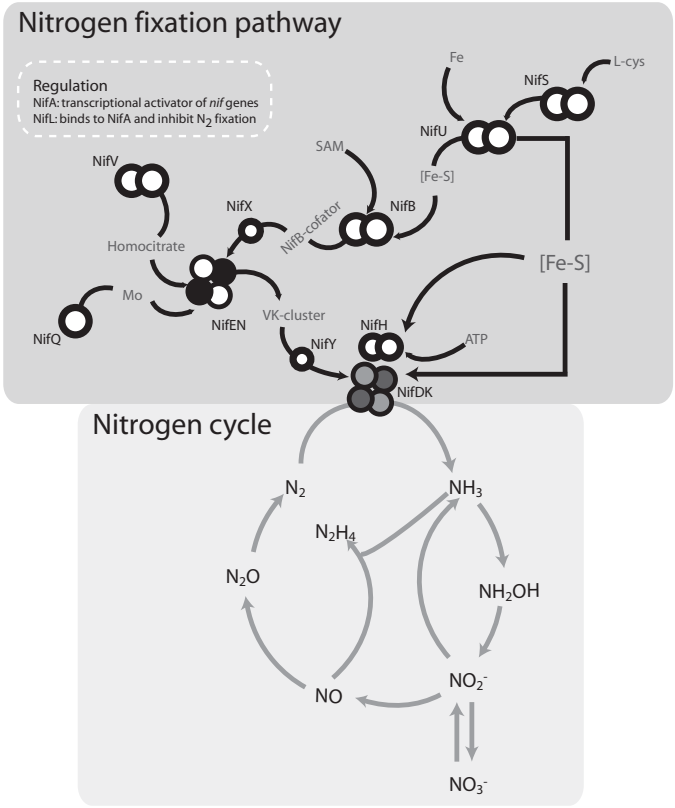


Figure 5.1: Schematic representation of the nitrogen fixation process together with the whole nitrogen cycle.

5.1.1 Nitrification

Nitrification, the stepwise oxidation of ammonia to nitrite, NO_2^-, via hydroxylamine and the successive oxidation of NO_2^- to NO_3^- (nitrate) is a catabolic O_2 dependent process that evolved after the oxygenation of the atmosphere and it is considered as the last step of nitrogen cycle appeared on Earth [Klotz & Stein, 2008]. Such process, enabling chemolithoautotrophic growth, (even though several heterotrophic bacteria can perform the same reaction) is performed by different players of the "nitrifying community" (Arp et al., 2007). The ammonia oxidizing bacteria use ammonia as an energy source for carbon assimilation. Nitrite oxidizers bacteria catalyzes the second step of nitrification and are so far restricted to five bacterial genera [Alawi *et al.*, 2007]. These microorganisms are able to catalyze the oxidation of nitrite in the reaction $NO_2^- + H_2O \rightarrow NO_3^- + 2H^+ 2e^-$ with the activity of nitrite oxidoreductase (NXR).

5.1.2 Denitrification

The denitrification process, the dissimilatory reduction of nitrate and nitrite to gaseous nitrogen, proceeds stepwisely following the reactions $NO_3^- \rightarrow NO_2^- \rightarrow NO \rightarrow N_2O \rightarrow N_2$ and is an anaerobic or microaerophilic process performed by denitrifying (facultative) heterotrophic soil and water bacteria using organic carbon source and nitrate as electron acceptor.

5.1.3 ANAMMOX

The recent discovery of anaerobic ammonia oxidation has been regarded as one of main advancement in the comprehension of nitrogen cycle [Jetten, 2008]. Microorganisms with this metabolic pathway are able to couple nitrification (oxidation of ammonia) and denitrification (until N_2 production) in anaerobic environments. The exact enzymology and genetic inventory of such process are still unsettled [Strous *et al.*, 2006]. The importance of ANAMMOX in the global nitrogen cycle is striking [Jetten, 2008] and its evolutionary origin intriguing. It is in fact proposed that ANAMMOX evolved soon after the incomplete denitrification pathway (in absence of the copper dependent NOS) [Strous *et al.*, 2006] and provided the first metabolic pathway to resupply the atmospheric N_2 pool and performed this role until the evolutionary origin of the full denitrification pathway.

5.1.4 Ammonification

Ammonification, the dissimilatory electrogenic reduction of nitrate to ammonia via formate or H_2 in oxygen limited conditions, is performed by many facultative and obligate chemolithotrophic proteobacteria [Simon, 2002]. Interestingly, since this process does not require oxygen and needs iron it is proposed that this pathway evolved very early and was responsible for fixed nitrogen resupply from abiotic formed NO_2^- before the advent of N_2 fixation [Mancinelli & McKay, 1988b].

5.2 Nitrogen Fixation: A Paradigm For The Evolution Of Metabolic Pathways

Nnitrogen fixation, the biological conversion of atmospheric dinitrogen to ammonia, represents an excellent model for studying the evolutionary interconnections linking different pathways and the functional divergence of paralogs Figure 5.1. Nitrogen fixation is the most important input of biologically available nitrogen in Earth ecosystems. It is a metabolic ability possessed only by some Bacteria (Green Sulphur Bacteria, Firmicutes, Actinomycetes, Cyanobacteria and Proteobacteria) and Archaea, where it is mainly present in methanogens [Dixon & Kahn, 2004]. Nitrogen fixation is compatible with different microbial lifestyles: aerobic, anaerobic and facultative anaerobic heterotrophs, anoxygenic and oxygenic photosynthetic bacteria and chemolithotrophs. Diazotrophs inhabit many ecological niches, marine and terrestrial environments as free living or plant symbiotic or endophytic microorganisms [Raymond, 2005]. The correlation between nitrogen fixation - that is poisoned by O_2 - and oxygen rich environment or oxygenic (photosynthetic) metabolism is particularly intriguing from an evolutionary viewpoint (see below). Nitrogen fixation is a complex process with a high energetic cost and requiring the activity of several genes Figure 5.1. In the free-living diazotroph *Klebsiella pneumoniae* 20 genes involved in nitrogen fixation (*nif* genes) have been identified Tab.5.1. The enzyme

Gene name	Poduct function	Source of reference
nifH	structural dinitrogenase reductase Fe protein	Mevarech et al. 1980
nifY	involved in nitrogenase maturation	Homer et al. 1993
nifT	involved in nitrogenase maturation	Simon et al. 1996
nifD	structural component of dinitrogenase (FeMo protein)	Lammers and Haselkorn 1984
nifK	structural component of dinitrogenase (FeMo protein)	Mazur and Chui 1982
nifU	required for the activation of Fe and FeMo proteins	Jacobson et al. 1989; Dos Snatos et al. 2004
nifS	required for the activation of Fe and FeMo proteins	Jacobson et al. 1989; Dos Snatos et al. 2004
nifM	required for accumulation of active FeMo protein	Jacobson et al. 1989; Howard et al. 1986; Paul and Merrick 1989
nifZ	acts as a chaperone in the assembly of the FeMo protein	Hu et al. 2004
nifW	protect the MoFe protein from oxygen damage	Kim et al.1996
nifN	scaffold for the FeMo and FeVn cofactor biosynthesis	Roll et al. 1995
nifE	scaffold for the FeMo and FeVn cofactor biosynthesis	Roll et al. 1995
nifO	involved in the biosynthesis of FeMo cofactor	Rodriguez-Quignones et al 1993; Shah et al. 1994
nifQ	involved in the biosynthesis of FeMo cofactor	Rodriguez-Quignones et al 1993; Shah et al. 1994
nifX	involved in FeMo-co biosynthesis (able to transfer NifB-co to nifEN)	Shah et al. 1999; Hernandez et al. 2006
nifB	crucial for FeMo cofactor biosynthesis	Bishop, P. E. & Joerger, R. D. (1990)
nifV	(homocitrate synthase) involved in the biosynthesis of FeMo cofactor	Filler et al. 1986
nifF	electron transport to the structural components	Hill and Kavanagh 1980
nifJ	electron transport to the structural components	Hill and Kavanagh 1980
nifA	(together with *rpoN*) activates trascription of all nitrogenase promoters	Dixon et al. 1980;Merrick 1983
nifL	modulates the activity of the transcriptional activator *NifA*	Hill et al.1981; Merrick et al. 1982; Blanco et al. 1993; Sidoti et al.1993

Table 5.1: The *nif* genes of *K. pneumoniae* together with their predicted functions

responsible for nitrogen fixation, the nitrogenase, shows high degree of conservation of structure, function and amino acid sequence across wide phylogenetic ranges [Fani *et al.*, 2000]. Nitrogenase contains an unusual metal clusters, the Iron-Molybdenum cofactor (FeMo-co), that is considered to be the site of dinitrogen reduction, and whose biosynthesis requires the products of *nifE*, *nifN*and several other *nif* genes Figure 5.1. All

known Mo-nitrogenases consist of two components, component I (dinitrogenase, or Fe-Mo protein), a $\alpha 2\beta 2$ tetramer encoded by *nifD* and *nifK*, and component II (dinitrogenase reductase, or Fe-protein) a homodimer encoded by *nifH*. In the last years some light has been shed on the molecular mechanisms responsible for the evolution of *nif* genes and the interconnections of nitrogen fixation with other metabolic pathways, such as bacteriochlorophyll biosynthesis [Xiong *et al.*, 2000]. In spite of this, many questions remain still open: 1) Is nitrogen fixation an ancestral character, arising prior to the appearance of LUCA? 2) How many genes were involved in the ancestral nitrogen fixation process? 3) How did the *nif* genes originate and evolve? 4) How and at what extent was nitrogen fixation correlated to other metabolic processes in the earliest cells? 5) Which were the molecular mechanisms involved in the origin, evolution and spreading of nitrogen fixation?

5.2.1 Is nitrogen fixation an ancestral character?

The time and order of appearance of nitrogen fixation in relation to the other nitrogen related metabolic pathways is still under debate. However, it is generally thought that N_2 fixation represents an early invention of evolution since the biological importance of the elements and the rapid depletion of abiotically fixed nitrogen in the primordial metabolism [Falkowski *et al.*, 2008]. Such model is consistent with both geological evidence, for example the availability of molybdenum and iron in the Archaean [Canfield *et al.*, 2006], and phylogenomics analyses [Raymond *et al.*, 2004]. Nevertheless, since *nif* genes can be organized in (compact) operons that are prone to HGT, the presence of *nif* genes in Archaea and Bacteria is not considered a straightforward demonstration of the antiquity of the metabolic pathway [Raymond *et al.*, 2004; Shi & Falkowski, 2008]. Moreover Mancinelli and McKay [1988b], basing on the complexity of the pathway, the high energy costs of fixation, and the absence in eukaryotic organelles, suggested that these findings are not compatible with an early origin of N_2 fixation that they proposed evolved after denitrification when fixed nitrogen was available for early metabolism by abiotic reactions or ammonification. This model is in agreement with the lack of supporting data for a depletion of atmospheric N_2 in presence of coupling of early nitrogen fixation and absence of denitrification [Capone & Knapp, 2007]. However, this scenario has some pitfalls [Klotz & Stein, 2008], such as the absence, in the Archean and Proterozoic eras, of nitrous oxide reductase (NOS), an enzyme possessed by extant denitrifiers for the lack of its copper cofactor and the low concentration of nitrite that could had formed only in limited amounts by combustion in the early neutral to mildly reducing CO_2 depleting Archean atmosphere [Navarro-Gonzalez *et al.*, 2001].

5.2.2 How many genes were involved in the ancestral nitrogen fixation?

The phylogenetic distribution of *nif* genes was checked in completely sequenced prokaryotes. The analysis performed by probing 842 prokaryotic genomes (52 Archaea and 790 Bacteria) for the presence of *nifH* genes revealed that 124 possessed it. All these genomes were scanned for the presence of genes homologous to each of the twenty *K. pneumoniae*

nif genes. As shown in Figure 5.2, only six *nif* genes (*nifHDKENB)*, those responsible for the synthesis of nitrogenase, nitrogenase reductase and Fe-Mo Cofactor biosynthesis, were present in almost all the genomes. All the other *nif* genes have a patchy phylogenetic distribution revealing a complex evolutionary history. This finding strongly suggests that if nitrogen fixation is an ancestral metabolic trait possessed by LUCA, it is quite possible that only *nifHDKENB* genes were present in the genome of the LUCA community. Thus, if nitrogen fixation required other enzymes, their function might have been performed by enzymes with low substrate specificity (in agreement with the Jensen hypothesis on the origin and evolution of metabolic pathways, [Jensen, 1976]). According to this idea, the *nifHDKENB* might represent a "universal core" for nitrogen fixation, whereas the other genes might be differentially acquired during evolution in the different phylogenetic lineages.

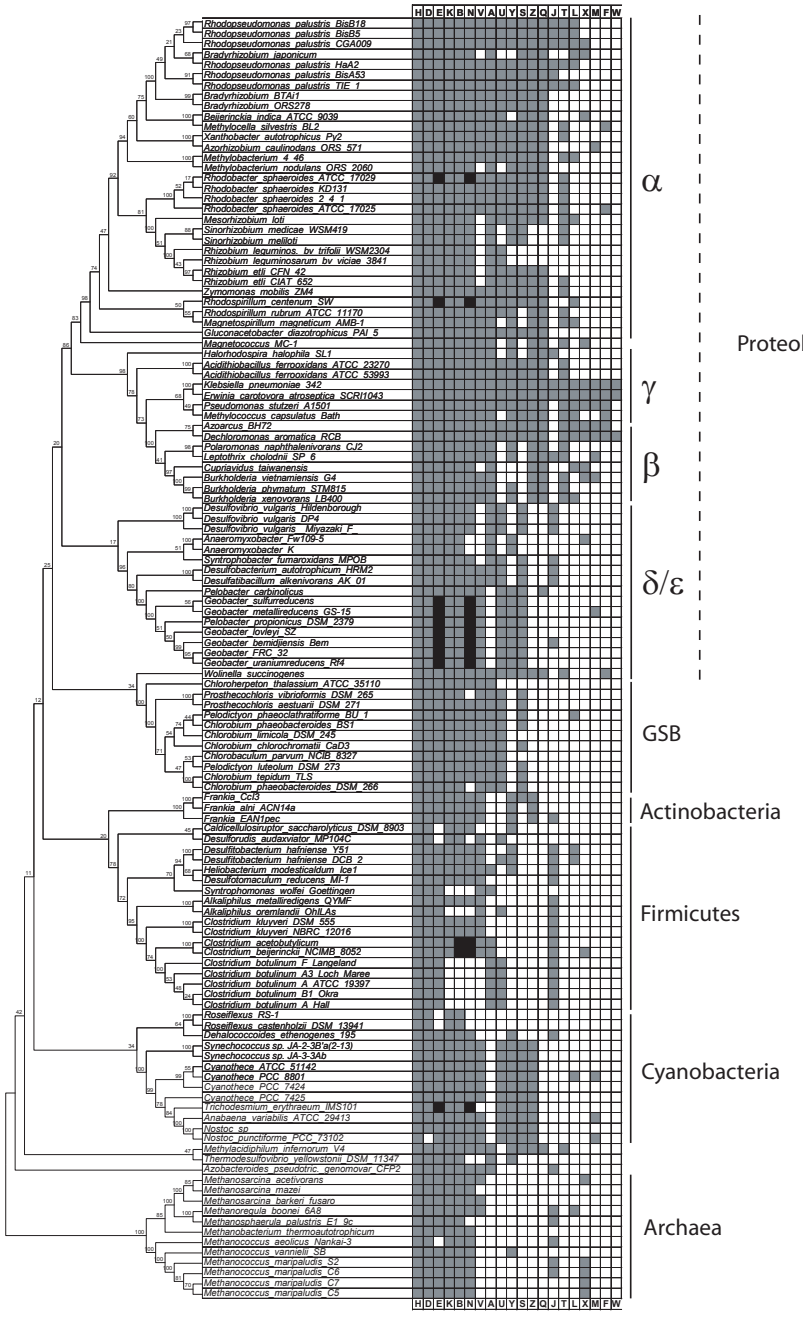


Figure 5.2: The distribution of *nif* genes within 124 diazotrophic Bacteria and Archaea (whose genomes were completely sequenced and available on NCBI). White and light grey boxes represent the absence or presence of the corresponding genes, respectively. Dark grey boxes represent fusions of the corresponding genes.

5.2.2.1 In/out - paralogs of *nif* genes

The hypothesis proposed in the previous paragraph implies that during evolution some genes might have been recruited from other metabolic pathways through duplication and divergence of genes coding for enzymes with a low substrate specificity. This idea is supported by the finding that most of *nif* genes have in-paralogs (i.e. paralogs involved in the same pathway) and/or out-paralogs (i.e. paralogs involved in different pathways) as pointed out by Fondi et al. (unpublished data) using a Psi-blast analysis using each Nif protein as query (Figure 5.3). The analysis did not retrieve any known paralogs for *nifW* (*nifO*), *nifT* (*fixU*), *nifQ* and *nifZ* which are also missing from a large fraction of diazotrophs genomes (Figure 5.2). Eight *nif* genes (*nifAFHJLMSU*) are related at a different extent to proteins involved in other metabolic pathways (out-paralogs). NifS is related to a number of paralogs mainly involved in amino acid and carbon metabolisms. NifJ, a multidomain pyruvate:ferredoxin (flavodoxin) oxidoreductase, exhibited a large number of paralogs. Several of the proteins involved in Fe-Mo cofactor biosynthesis have paralogs in other cofactor biosyntheses. Eight Nif proteins share a significant degree of sequence similarity with proteins involved in other metabolic routes, and also with other *nif* genes products; this group can be further separated into two different clusters, the first of which includes *nifDKEN*, and the second being composed by *nifBXY* and *nifV*. Actually, NifBXY are related through a common domain of about 90 aminoacids; moreover, *nifB* has an additional domain belonging to the S:-adenosylmethionine (SAM) family, found in proteins that catalyze diverse reactions, including unusual methylations, isomerisation, sulphur insertion, ring formation, anaerobic oxidation and protein radical formation. Evidence exists that these proteins generate a radical species by reductive cleavage of SAM through an unusual Fe-S centre. The genes *nifV* and *nifB* are not directly linked and their connection is due to multidomain proteins sharing homology with NifV and NifB in different domains. As expected, NifDKEN showed sequence similarity with Bch proteins involved in bacterial photosynthesis.

5.2.2.2 Nitrogen fixation and bacterial photosynthesis: an ancestral interconnections through a cascade of gene and operon duplication.

The two gene pairs *nifD-nifK* and *nifE-nifN*, coding for nitrogenase and the tetrameric complex Nif N2E2, form a paralogous gene family, and arose through duplications of an ancestral gene, by a two-step model in which an ancestor gene underwent an in-tandem duplication event giving rise to a bicistronic operon; this, in turn, duplicated leading to the ancestors of the present-day *nifDK* and *nifEN* operons [Fani *et al.*, 2000]. The model proposed is in agreement with the Retrograde Hypothesis [Jensen, 1976] but also fits the Jensen's hypothesis of the metabolic pathways assembly. Accordingly, the ancestor of the *nif* gene family encoded a protein which might assemble to give a homotetrameric (or a homomultimeric) complex with a low substrate specificity able to catalyse more than one enzymatic reactions [Fani *et al.*, 2000]. By assuming that the ability to fix nitrogen was a primordial property dating back to LUCA [Fani *et al.*, 2000; Zillig *et al.*, 1992], then the duplication events leading to the two operons predated the appearance of LUCA and the

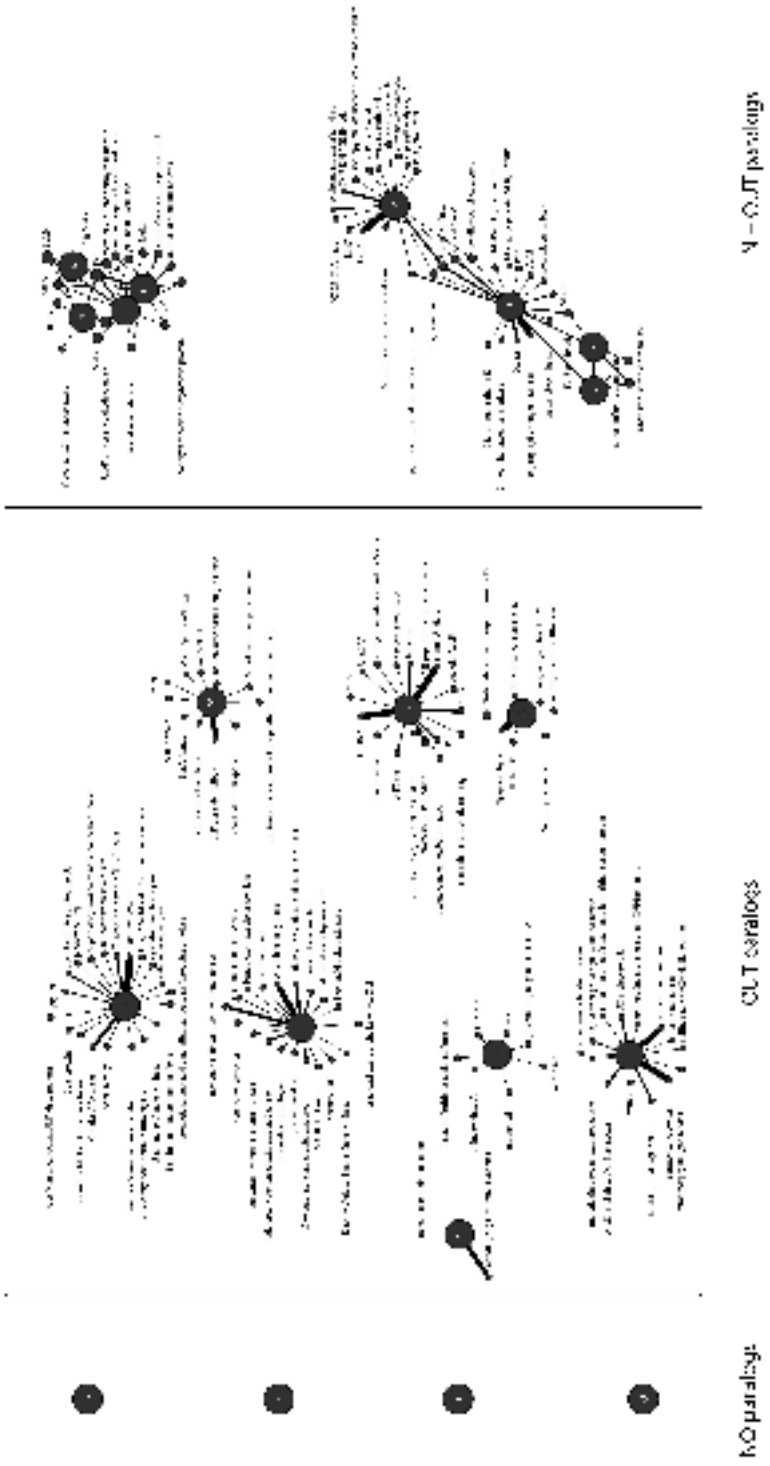


Figure 5.3: In- and Out-paralogs network of *nif* genes. Nodes represent protein, links represent similarity values.

function(s) performed by this primordial enzyme might have depended on the composition of the early atmosphere. There is no agreement on the composition of the primitive atmosphere, but it is generally accepted that O_2 was absent and this represents an essential prerequisite for the appearance of (an ancestral) nitrogenase, which is inactivated by free oxygen [Fay, 1992]. The appearance of nitrogenase on the primitive Earth would have represented a necessary event for the first cells, living in a planet whose atmosphere was neutral, containing dinitrogen, but not ammonia (Figure 5.4, Scenario 1). In fact, if ammonia was required by the primitive micro-organisms for their syntheses, then its absence must have imposed a selective pressure favouring those cells that had evolved a system to synthesise ammonia from atmospheric dinitrogen. Therefore, according to this scenario, the function of the ancestral enzyme might have been that of a "nitrogenase", slow, inefficient and with low substrate specificity able to react with a wide range of compounds with a triple bond. According to a second theory, the early atmosphere was a reducing

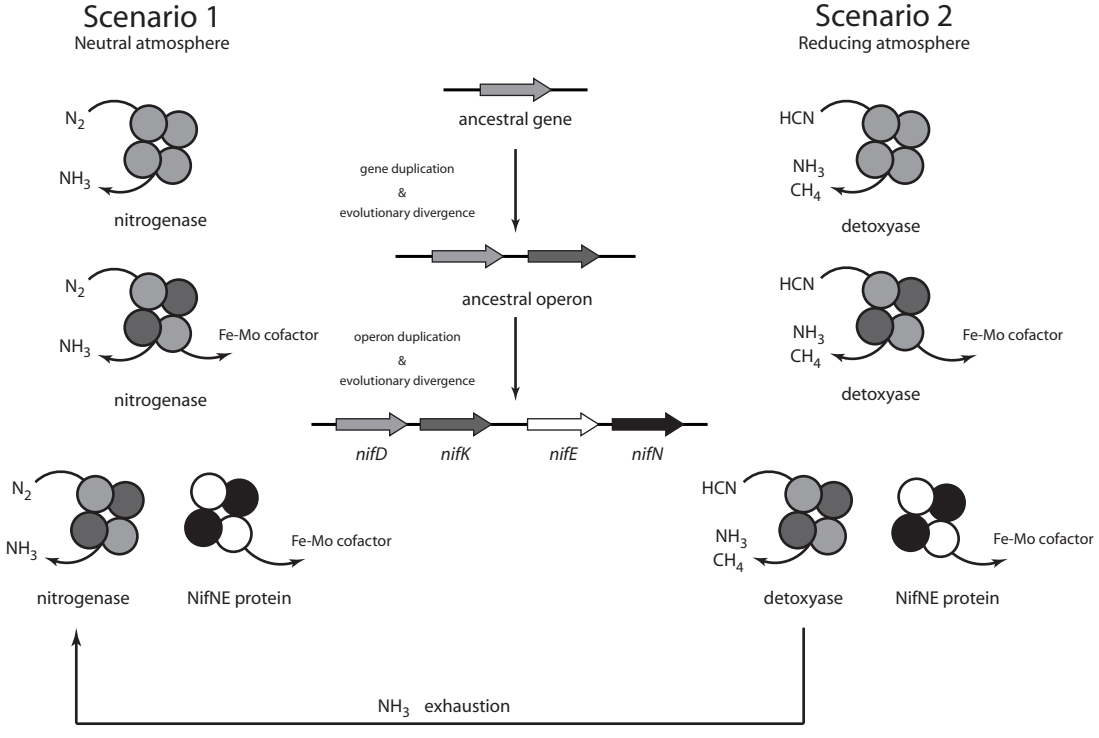


Figure 5.4: Two possible scenarios depicted for the original function performed by the nifDKEN genes and their ancestor(s) gene(s).

one and contained free ammonia (Figure 5.4). In those conditions, the evolution of a nitrogen fixation system was not a prerequisite because of the abundance of abiotically produced ammonia. Hence, why a nitrogenase in those days? The answer to this question relies in the catalytic properties of nitrogenase. In fact the enzyme is able to reduce also other molecules such as acetylene, hydrogen azide, hydrogen cyanide, or nitrous oxide, all of which contain a triple bond. Therefore, according to this second scenario (Figure

5.4, Scenario 2), the primitive enzyme encoded by the ancestor gene, would have been a detoxyase, an enzyme involved in detoxifying cyanides and other chemicals present in the primitive reducing atmosphere [Silver & Postgate, 1973]. This scenario implies that the progressive exhaustion of combined nitrogen would have imposed the refinement of the enzyme specificity which very likely modified and adapted to another triple-bond substrate, dinitrogen, and was selected for, and retained by some bacterial and archaeal lineages to enable survival in nitrogen-deficient environments. Finally, the decreasing of free ammonia and cyanides in the atmosphere triggered the evolution of the detoxyase toward nitrogenase, that might have been a common feature of all microbial life until photosynthesising cyanobacteria largely increased the oxygen concentration and burned cyanides. Particularly intriguing is the finding that genes coding for nitrogenase (*nifDK*) and nitrogenase reductase (*nifH*) are evolutionary related to the genes involved in bacteriochlorophyll biosynthesis (see below). Chlorophyll (Chl) and bacteriochlorophyll (Bchl) are the photochemicallly active reaction centre pigments for most of the extant photosynthetic organisms. During the synthesis of both Chl and Bchl, reduction of the tetrapyrrole ring system converts protochlorophyllide (Pchlide), into a chlorin. A second reduction that is unique to the synthesis of Bchl converts the chlorin into a bacteriochlorin. There are two mechanisms for reducing the double bond in the fourth ring of protochlorophyllide. One enzyme complex functions irrespective of the presence or absence of light and is thus termed "light-independent protochlorophyllide reductase". The second is a light-dependent reaction that utilizes the enzyme NADPH-protochlorophyllide oxidoreductase [Suzuki *et al.*, 1997]. In *Rhodobacter capsulatus*, the products of three genes are required for each reduction: *bchL*, *bchN*, and *bchB* for the Pchlide reductase and *bchX*, *bchY*, and *bchZ* for the chlorin reductase [Burke *et al.*, 1993]. Both enzymes are three-subunit complexes. Burke et al. [1993] detected a significant degree of sequence similarity between BchlL, BchN, BchB, and BchX, BchY and BchZ, respectively, suggesting that the six genes represent two triads of paralogs and that the two enzymes are derived from a common three-subunit ancestral reductase. It was also found that the so-called "chlorophyll iron protein" subunits encoded by *bchX*, *bchL*, and *chlL* shared a remarkable sequence similarity with the nitrogenase Fe proteins [Burke *et al.*, 1993]. Burke et al. (1993) suggested that genes involved in bacteriochlorophyll biosynthesis and nitrogen fixation were related mechanistically, structurally and evolutionarily. Similarly to NifH protein, which serves as the unique electron donor for the nitrogenase complex, the products of *bchL* and *bchX* could serve as the unique electron donor into their respective catalytic subunits (BchB-BchN and BchY-BchZ). The idea of a common ancestry of *nifH*, *bchL* and *chlL* genes [Burke *et al.*, 1993; Fujita *et al.*, 1993] has had an elegant experimental support by [Cheng *et al.*, 2005] who demonstrated in the photosynthetic eukaryote *Chlamydomonas reinhardtii* that NifH is able to partially complement the function of ChlL in the dark-dependent chlorophyll biosynthesis pathway. Nitrogenases and carboxylases might have represented bacterial preadaptations, multigenic traits that were retained because of new selective advantages in altered environments. As abiotically produced organic matter became depleted, competition for the organic prerequisites for reproduction ensued. As the carboxylation and nitrogen-fixing functions were achieved, a new, abundant, and direct

source of carbon and nitrogen for organic synthesis became available in the atmosphere. The ability to take up atmospheric carbon and nitrogen would be of great selective advantage [Margulis, 1993]. It is possible to propose a model (Figure 5.5) for the origin and evolution of nitrogen fixation and bacterial photosynthesis based on multiple and successive paralogous duplications of an ancestral operon encoding an ancient reductase. The eight genes (*nifDKEN* and *bchYZNB*) are members of the same paralogous gene family, in that that all of them are the descendant of a single ancestral gene. The model proposed posits the existence of an ancestral three-cistronic operon Figure 5.5 coding for an unspecific reductase. One might assume that this complex was (eventually) able to perform both carboxylation and nitrogen fixation. The following evolutionary steps might have been

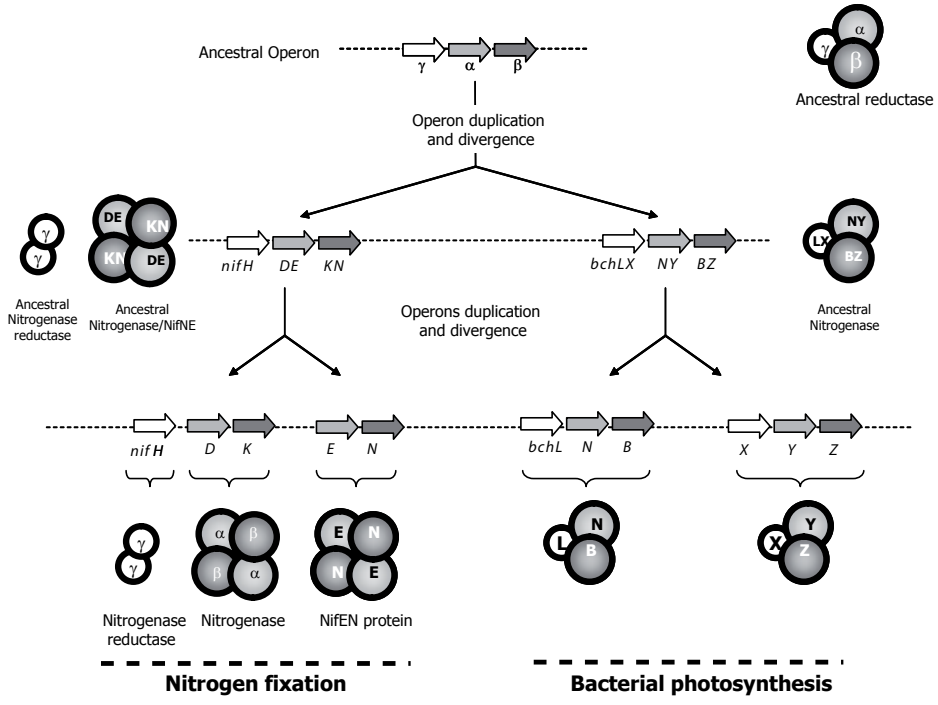


Figure 5.5: Possible evolutionary model accounting for the evolutionary relationships between *nif* and *bch* genes.

the duplication of the ancestral operon followed by an evolutionary divergence that led to the appearance of the ancestor of *nifH*, *nifDE*, and *nifKN* on one side, and *bchLX*, *bchNY* and *bchBZ* on the other one (Figure 5.5). In this way the two reductases narrowed their substrate specificity with one of them channelled toward nitrogen fixation and the other one toward photosynthesis. However, each of the two multicomplex proteins was able to perform at least two different reactions: 1) the ancestor of *nifDKEN*, was likely able to carry out the reduction of dinitrogen to ammonia and the synthesis of Fe-Mo cofactor [Fani *et al.*, 2000]; 2) The ancestor of protochlorophyllide reductase and chlorin reductase performed both of the reactions that in the extant photosynthetic bacteria are carried out by two triads (BchN and BchLX, respectively). The complete diversification of the function of the two heteromeric complexes was likely achieved thorough duplication of *nifDE*

nifKN ancestors and by the duplication of the three-cistronic operon *bch(LX)(NY)(NZ)* followed by evolutionary divergence (Figure 5.5). In our opinion, this idea may perfectly fit the Jensen's hypothesis [Jensen, 1976]. Concerning the timing of the above reported evolutionary events [Fani *et al.*, 2000] the two paralogous duplication events leading to *nifDK* and *nifEN* likely predated the appearance of the LUCA. Conversely, other authors [Raymond *et al.*, 2004] have proposed a different scenario, according to which nitrogen fixation per se was invented by methanogenic Archaea and subsequently transferred, in at least three separate events, into bacterial lineages. Differently from nitrogen fixation, tetrapyrrole-based photosynthesis occurs only in bacteria and bacterially derived chloroplasts, therefore it can be surmised that the appearance of photosynthesis should have not predated the divergence of Archaea and Bacteria.

5.2.3 Which were the molecular mechanisms involved in the spreading of nitrogen fixation?

The phylogenetic analysis performed using a concatenation of NifHDKEN proteins (Figure 5.6) may help to shed light on the main evolutionary steps leading to the extant distribution of nitrogen fixation in Prokaryotes. As shown in Figure 5.6, a group of bacteria (including representatives from Green Sulphur Bacteria (GSB) δ-proteobacteria and Chloroflexi) are strongly supported as sister groups of a cluster embedding *Methanosarcina* (Euryarchaea). Similarly, some Firmicutes (mainly *Clostridium* species) cluster as a sister clade with the Euryarchaea *Methanoregula boonei*. Their position in the phylogenetic tree suggests that these bacteria might have acquired nitrogen fixation via HGT from an archaeon. It is worth of noticing that all the microorganisms embedded in this clade are frequently found among syntophic consortia in anaerobic environment, providing a viable environment for gene sharing [Garcia *et al.*, 2000]. All the other bacterial sequences are embedded in a single monophyletic group. Interestingly, the sequences from Cyanobacteria, Firmicutes and Actinobacteria form three monophyletic clades that emerge as sister groups of α-, γ-, β-, δ- and ϵ-proteobacteria, respectively. The monophyly of the three groups that are surrounded by proteobacteria, points toward a later acquisition of nitrogen fixation in these bacteria from a proteobacterium; hence, HGT appears to have played a key role in spreading nitrogen fixation within the different bacterial lineages. The phylogenetic analyses also suggested that the ancestor of extant proteobacteria was a diazotroph. An evolutionary model for origin and spreading of nitrogen fixation is shown in (Figure 5.7). The available data do not permit to discern whether LUCA was a diazotroph or not. If we assume that LUCA already possessed the set of genes necessary for nitrogen fixation (the LCA hypothesis, Figure 5.7a) then gene loss should have played a major role in the evolution of nitrogen fixation pathway. Conversely, if we assume that nitrogen fixation was not present in LUCA but was later "invented" by methanogenic Archaea (Raymond et al. 2004), extensive HGT must be invoked to account for the distribution and the phylogeny that we observe in present-day prokaryotes Figure 5.7b. Finally, phylogenetic data suggest that, once appeared in bacteria, *nif* genes flowed through the ancestral prokaryotic communities by vertical inheritance and HGT events.

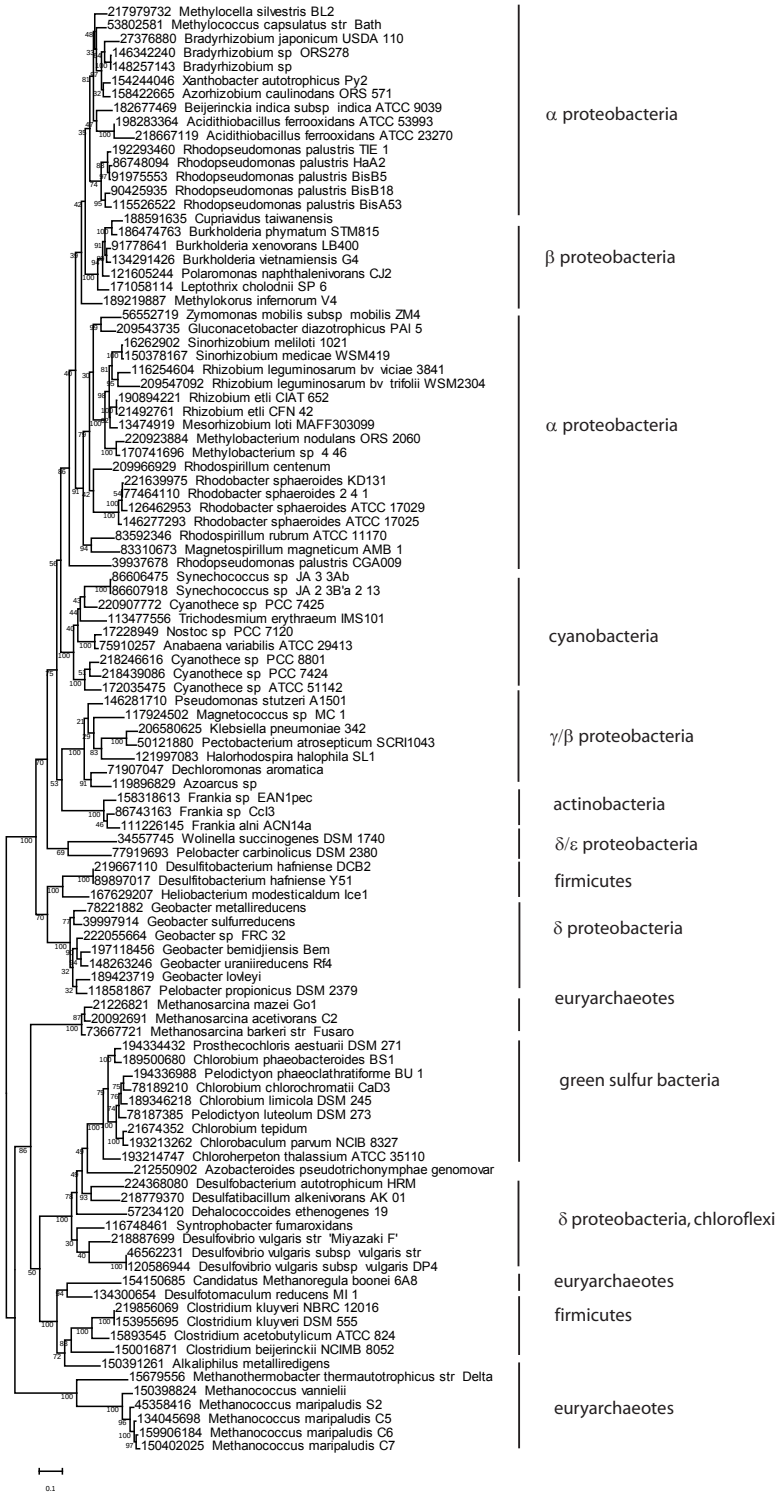


Figure 5.6: Maximum Likelihood phylogenetic tree of concatenated NifHDKEN sequences from 105 representative microorganisms.

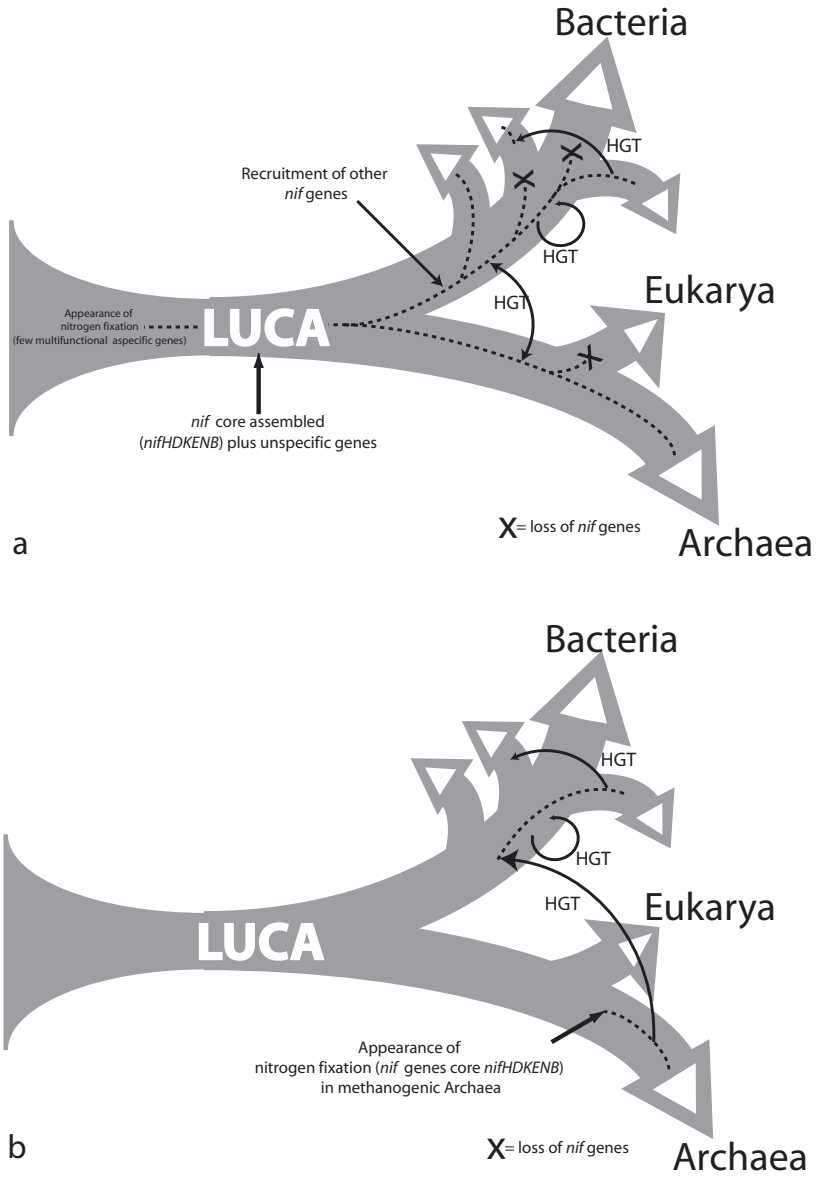


Figure 5.7: : Schematic representation of the origin, evolution and spreading of nif genes in Bacteria and Archaea assuming a) the presence of a core of *nif* gene in LUCA or b) the appearance of Nitrogen Fixation in methanogenic Archaea.

5.3 Conclusions

Data reported in this work confirm that in the course of molecular evolution different mechanisms might have concurred in the acquisition of new metabolic abilities and that gene duplication is a major force in genome evolution and extend the framework of genome evolution to operon duplication. Duplication may concern gene portions, coding for protein domains and motifs, entire genes, operons, part of genomes and entire chromosomes.

The antiquity of the above reported paralogous genes are in agreement with the idea that functional duplications of DNA stretches may have played an essential role in shaping the main metabolic pathways during the early stages of molecular evolution. Nitrogen fixation has been an ancient innovation that played a critical role during the early expansion of microbial life (being still crucial for extant life). Both i) the presence of the *nif core* in all the scanned prokaryotic genomes and ii) the phylogeny constructed using the concatenation of their sequences (consistent with their species phylogeny) speak towards the presence of these genes in the LUCA, although the possibility that nitrogen fixation was invented in a further stage of evolution and then spread through HGT cannot be, at present, completely ruled out. Lastly, the applied strategy allowed to map on the species phylogeny tree the appearance of several genes related to nitrogen fixation in several different bacterial lineages. This, in turn, suggests that, their appearance (and/or recruitment) during microbial evolution, probably allowed the refinement of nitrogen fixation process, initially carried out by a limited number of genes.

References

Alawi, M., Lipski, A., Sanders, T., Pfeiffer, E.M. & Spieck, E. (2007). Cultivation of a novel cold-adapted nitrite oxidizing betaproteobacterium from the siberian arctic. *The ISME J.*, **1**, 256264.

Burke, D., Hearst, J. & Sidow, A. (1993). Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins. *Proc. Natl. Acad. Sci. USA*, **90**, 7134–7138.

Canfield, D.E., Rosing, M.T. & Bjerrum, C. (2006). Early anaerobic metabolisms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **361**, 1819–1836.

Capone, D.G. & Knapp, A.N. (2007). Oceanography: a marine nitrogen cycle fix? *Nature*, **445**, 159–60, capone, Douglas G Knapp, Angela N Comment News England Nature Nature. 2007 Jan 11;445(7124):159-60.

Cheng, Q., Day, A., Dowson-Day, M., Shen, G.F. & Dixon, R. (2005). The klebsiella pneumoniae nitrogenase fe protein gene (nifh) functionally substitutes for the chll gene in chlamydomonas reinhardtii. *Biochem. Biophys. Res. Commun.*, **329**, 966–975.

Deutsch, C., Sarmiento, J.L., Sigman, D.M., Gruber, N. & Dunne, J.P. (2007). Spatial coupling of nitrogen inputs and losses in the ocean. *Nature*, **445**, 163–7, deutsch, Curtis Sarmiento, Jorge L Sigman, Daniel M Gruber, Nicolas Dunne, John P Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England Nature Nature. 2007 Jan 11;445(7124):163-7.

Dixon, R. & Kahn, D. (2004). Genetic regulation of biological nitrogen fixation. *Nat Rev Micro*, **2**, 621–631, 10.1038/nrmicro954.

Falkowski, P.G. (1997). Evolution on the nitrogen cycle and its influence on the biological sequestration of co2 in the ocean. *Nature*, **387**, 272–275.

Falkowski, P.G., Fenchel, T. & Delong, E.F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, **320**, 1034–9.

Fani, R., Gallo, R. & Lio, P. (2000). Molecular evolution of nitrogen fixation: the evolutionary history of the nifd, nifk, nife, and nifn genes. *J Mol Evol*, **51**, 1–11.

Fay, P. (1992). Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol. Rev.*, **56**, 340–73.

Fujita, Y., Matsumoto, H., Takahashi, Y. & Matsubara, H. (1993). Identification of a nifdk-like gene (orf467) involved in the biosynthesis of chlorophyll in the cyanobacterium plectonema boryanum. *Plant Cell Physiol.*, **34**, 305–314.

Galloway, J.N., Townsend, A.R., Erisman, J.W., Bekunda, M., Cai, Z., Freney, J.R., Martinelli, L.A., Seitzinger, S.P. & Sutton, M.A. (2008). Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. *Science*, **320**, 889–92.

Garcia, J.L., Patel, B.K.C. & Ollivier, B. (2000). Taxonomic phylogenetic and ecological diversity of methanogenic archaea. *Anaerobe*, **6**, 205226.

Gruber, N. & Galloway, J.N. (2008). An earth-system perspective of the global nitrogen cycle. *Nature*, **451**, 293–6, gruber, Nicolas Galloway, James N Research Support, Non-U.S. Gov't England Nature Nature. 2008 Jan 17;451(7176):293-6.

Houlton, B.Z., Wang, Y.P., Vitousek, P.M. & Field, C.B. (2008). A unifying framework for dinitrogen fixation in the terrestrial biosphere. *Nature*, **454**, 327–30, houlton, Benjamin Z Wang, Ying-Ping Vitousek, Peter M Field, Christopher B Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England Nature Nature. 2008 Jul 17;454(7202):327-30. Epub 2008 Jun 18.

Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, **30**, 409–25.

Jetten, M.S. (2008). The microbial nitrogen cycle. *Environ. Microbiol.*, **10**, 2903–9.

Kasting, J.F. & Siefert, J.L. (2001). Biogeochemistry: the nitrogen fix. *Nature*, **412**, 26–27.

Klotz, M.G. & Stein, L.Y. (2008). Nitrifier genomics and evolution of the nitrogen cycle. *FEMS Microbiol. Lett.*, **278**, 146–56.

Mancinelli, R.L. & McKay, C.P. (1988a). The evolution of nitrogen cycling. *Orig. Life Evol. Biosph.*, **18**, 311–325.

Mancinelli, R.L. & McKay, C.P. (1988b). The evolution of nitrogen cycling. *Orig. Life Evol. Biosph.*, **18**, 311–325.

Margulis, L. (1993). *Symbiosis in cell evolution: microbial communities in the archean and proterozoic eons*. WH Freeman and Company, New York.

McLain, J.E.T. & Martens, D.A. (2005). Nitrous oxide flux from soil amino acid mineralization. *Soil. Biol. Biochem.*, **37**, 289–299.

Navarro-Gonzalez, R., McKay, C.P. & Mvondo, D.N. (2001). A possible nitrogen crisis for archaean life due to reduced nitrogen fixation by lightning. *Nature*, **412**, 61–64, 10.1038/35083537.

Raymond, J. (2005). The evolution of biological carbon and nitrogen cycling -a genomic perspective. *Reviews in Mineralogy and Geochemistry*, **59**, 211–231.

Raymond, J., Siefert, J.L., Staples, C.R. & Blankenship, R.E. (2004). The natural history of nitrogen fixation. *Mol. Biol. Evol.*, **21**, 541–54.

Shi, T. & Falkowski, P. (2008). Genome evolutuon in cyanobacteria: The stable core and the variable shell. *Proc. Natl. Acad. Sci. USA*, **107**, 2510–2515.

Silver, V. & Postgate, J. (1973). Evolution of asymbiotic nitrogen fixation. *J. Theor. Biol.*, **56**, 340–373.

Simon, J. (2002). Enzymology and bioenergetics of respiratory nitrite ammonification. *FEMS Microbiol. Reviews*, **26**, 285–309.

Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M.W., Horn, M., Daims, H., Bartol-Mavel, D., Wincker, P., Barbe, V., Fonknechten, N., Vallenet, D., Segurens, B., Schenowitz-Truong, C., Medigue, C., Collingro, A., Snel, B., Dutilh, B.E., Op den Camp, H.J.M., van der Drift, C., Cirpus, I., van de Pas-Schoonen, K.T., Harhangi, H.R., van Niftrik, L., Schmid, M., Keltjens, J., van de Vossenberg, J., Kartal, B., Meier, H., Frishman, D., Huynen, M.A., Mewes, H.W., Weissenbach, J., Jetten, M.S.M., Wagner, M. & Le Paslier, D. (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, **440**, 790–794.

Suzuki, J., Bollivar, D. & Bauer, C. (1997). Genetic analysis of chlorophyll biosynthesis. *Annu. Rev. Genet.*, **31**, 61–89.

Wachtershauser, G. (2007). On the chemistry and evolution of the pioneer organism. *Chem. Biodivers.*, **4**, 584–602.

Xiong, J., Fischer, W.M., Inoue, K., Nakahara, M. & Bauer, C.E. (2000). Molecular evidence for the early evolution of photosynthesis. *Science*, **289**, 1724–1730.

Zillig, W., Palm, P. & Klenk, H. (1992). A model for the early evolution of organisms: the arisal of the three domains of life from a common ancestor. In H. Hartman & K. Matsuno, eds., *The origin and evolution of the cell*, 163–182, World Scientific, Singapore.

Chapter 6

The origin of Plant phenylpropanoid metabolism

The appearance of land plants was a key step towards the development of modern terrestrial ecosystems. Fossil data indicate that the first land plants appeared around 500 million years ago, from a pioneer green algal ancestor probably related to Charales. Early terrestrial environments were harsh. The ancestor of land plants that conquered emerged lands had to face important stresses including desiccation, UV radiation (not anymore shielded by water), as well as attack by already diversified microbial soil communities. This drove a number of key adaptations, including the emergence of specialized secondary metabolic pathways. Among them, the phenylpropanoid pathway (Figure 6.1) was crucial. It is in fact a ubiquitous and specific trait of land plants, and provides vital compounds such as lignin -essential for vascularization (xylem) and stem rigidity out of water-, and flavonoids -essential for reproductive biology (flower and fruit colors), protection against UV (pigments) and microbial attack (phytoalexins), and plant-microbe interaction (flavonoids). Three steps constituting the general phenylpropanoid pathway provide the precursors for the flavonoid and lignin branches. Phenylalanine ammonia-lyase (PAL) transforms phenylalanine into trans-cinnamic acid, which leads to p-coumaric acid by the action of cinnamic acid 4-hydrolase (CH4), which is then transformed into p-coumaroyl-CoA by p-coumaroyl:CoA ligase (4CL). It can be inferred that the origin of PAL, the first enzyme and the entering point of the whole phenylpropanoid metabolism, was a key evolutionary event, since it provided the initial step from which the rest of the pathway was assembled. Indeed, PAL is a key regulator of the phenylpropanoid pathway and any inhibition of PAL blocks the whole pathway. Given the clear importance of PAL in the emergence of the phenylpropanoid pathway and adaptation of plants to land, we sought to get more insight into the origin of this enzyme by carrying out an extensive search of PAL homologs in current sequence databases and by analyzing their phylogeny.

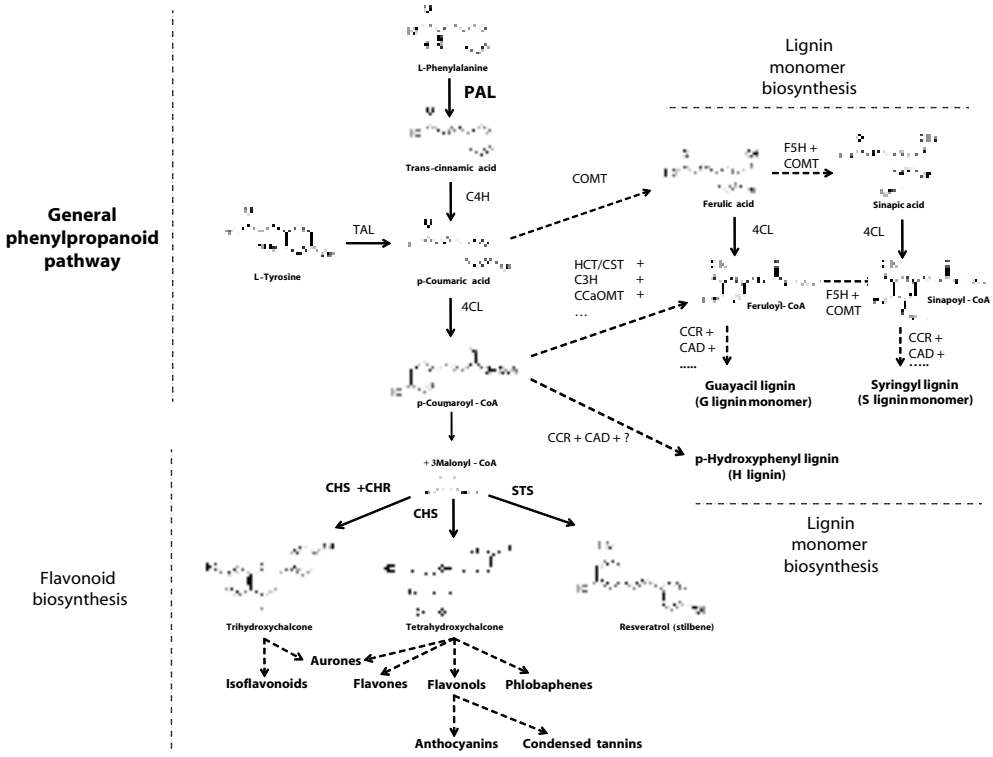


Figure 6.1: The Plant phenylpropanoid metabolism.

6.1 A horizontal gene transfer at the origin of Plant phenyl-propanoid metabolism

In this work we have performed an extensive phylogentic analysis of Phenylalanine Ammonia Lyase (PAL), which catalyses the first and essential step of the general plant phenylpropanoid pathway. This metabolic step leads from phenylalanine to p-Coumaric acid and p-Coumaroyl-CoA, the entry points of the flavonoids and lignin routes. We obtained robust evidence that the ancestor of land plants acquired a PAL via horizontal gene transfer (HGT) during symbioses with soil bacteria and fungi that are known to have established very early during the first steps of land colonization. This horizontally acquired PAL represented then the basis for further development of the phenylpropanoid pathway and plant radiation on terrestrial environments.

Biology Direct

BioMed Central

Research

Open Access

A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land

Giovanni Emiliani[1], Marco Fondi[2], Renato Fani[2] and Simonetta Gribaldo*[3]

Address: [1]Department of Environmental and Forestry Sciences and Technologies, University of Florence, via S. Bonaventura, 13, 50145, Florence, Italy, [2]Department of Evolutionary Biology, University of Florence, via Romana 19, 50125, Florence, Italy and [3]Institut Pasteur, Unité de Biologie Moléculaire du gène chez les Extrêmophiles, 25 rue du Docteur Roux, 75724, Paris Cedex 15, France

Email: Giovanni Emiliani - giovanni.emiliani@unifi.it; Marco Fondi - marco.fondi@unifi.it; Renato Fani - renato.fani@unifi.it; Simonetta Gribaldo* - simonetta.gribaldo@pasteur.fr

* Corresponding author

Published: 16 February 2009

Biology Direct 2009, **4**:7 doi:10.1186/1745-6150-4-7

Received: 21 January 2009
Accepted: 16 February 2009

This article is available from: http://www.biology-direct.com/content/4/1/7

Abstract

Background: The pioneering ancestor of land plants that conquered terrestrial habitats around 500 million years ago had to face dramatic stresses including UV radiation, desiccation, and microbial attack. This drove a number of adaptations, among which the emergence of the phenylpropanoid pathway was crucial, leading to essential compounds such as flavonoids and lignin. However, the origin of this specific land plant secondary metabolism has not been clarified.

Results: We have performed an extensive analysis of the taxonomic distribution and phylogeny of Phenylalanine Ammonia Lyase (PAL), which catalyses the first and essential step of the general phenylpropanoid pathway, leading from phenylalanine to p-Coumaric acid and p-Coumaroyl-CoA, the entry points of the flavonoids and lignin routes. We obtained robust evidence that the ancestor of land plants acquired a PAL *via* horizontal gene transfer (HGT) during symbioses with soil bacteria and fungi that are known to have established very early during the first steps of land colonization. This horizontally acquired PAL represented then the basis for further development of the phenylpropanoid pathway and plant radiation on terrestrial environments.

Conclusion: Our results highlight a possible crucial role of HGT from soil bacteria in the path leading to land colonization by plants and their subsequent evolution. The few functional characterizations of sediment/soil bacterial PAL (production of secondary metabolites with powerful antimicrobial activity or production of pigments) suggest that the initial advantage of this horizontally acquired PAL in the ancestor of land plants might have been either defense against an already developed microbial community and/or protection against UV.

Reviewers: This article was reviewed by Purificación López-García, Janet Siefert, and Eugene Koonin.

Background

The appearance of land plants was a key step towards the development of modern terrestrial ecosystems. Fossil data indicate that the first land plants appeared around 500 million years ago, from a pioneer green algal ancestor probably related to Charales [1,2].

Early terrestrial environments were harsh. The ancestor of land plants that conquered emerged lands had to face important stresses including desiccation, UV radiation (not anymore shielded by water), as well as attack by already diversified microbial soil communities [1,3]. This drove a number of key adaptations, including the emergence of specialized secondary metabolic pathways. Among them, the phenylpropanoid pathway was crucial. It is in fact a ubiquitous and specific trait of land plants, and provides vital compounds such as lignin -essential for vascularization (xylem) and stem rigidity out of water-, and flavonoids -essential for reproductive biology (flower and fruit colors), protection against UV (pigments) and microbial attack (phytoalexins), and plant-microbe interaction (flavonoids) [4,5]. Three steps constituting the general phenylpropanoid pathway provide the precursors for the flavonoid and lignin branches (Figure 1). Phenylalanine ammonia-lyase (PAL) transforms phenylalanine into trans-cinnamic acid, which leads to p-coumaric acid by the action of cinnamic acid 4-hydrolase (C4H), which is then transformed into p-coumaroyl-CoA by p-coumaroyl:CoA ligase (4CL) (Figure 1). Either p-coumaric acid and p-coumaroyl-CoA can enter the lignin monomer pathway, while p-coumaroyl-CoA is the precursor of the flavonoid pathway (Figure 1). Lignin monomer and flavonoid biosynthesis then involve complex highly branched pathways (Figure 1)[4,6].

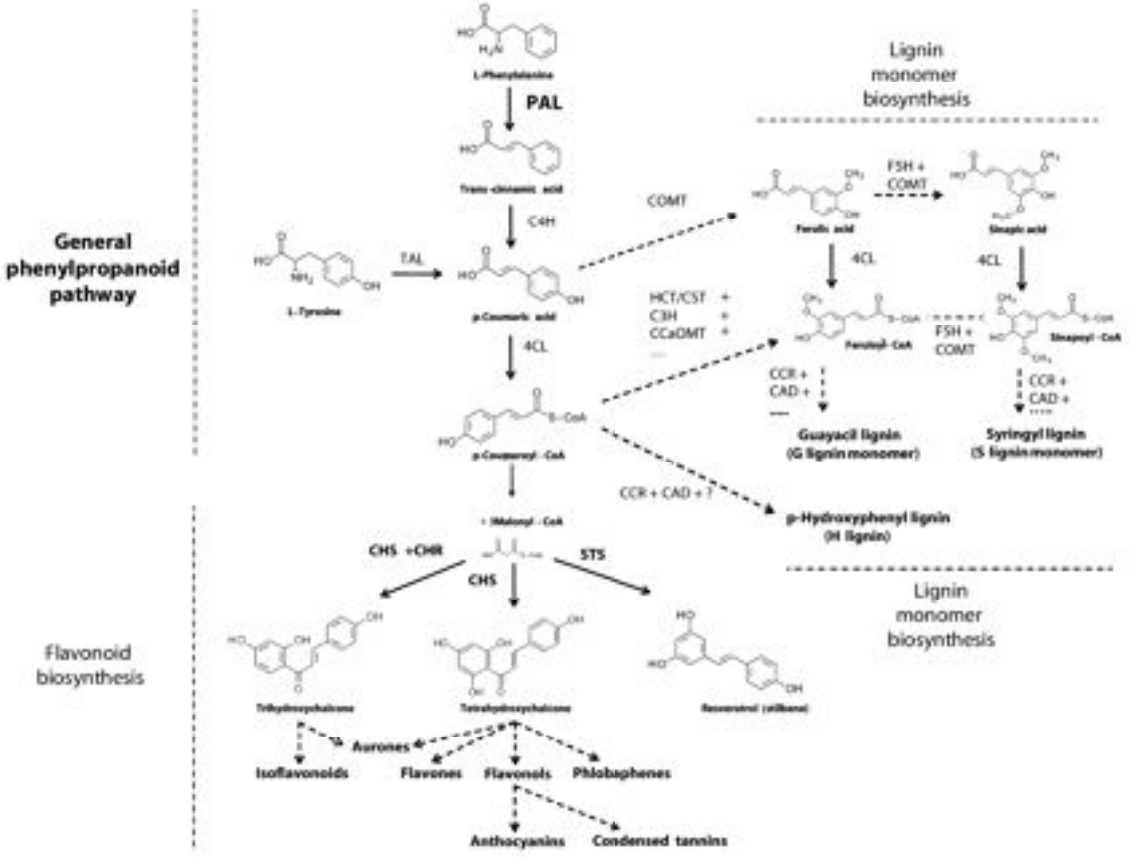


Figure 1

A schematic representation of phenylpropanoid metabolism. From the general phenylpropanoid pathway (top left, reactions from L-phenylalanine to p-Coumaroyl-CoA) two separated branches lead to the production of lignin monomers (right) and of flavonoids (bottom). Solid arrows indicate a single step enzymatic reaction, dashed arrows multiple sequential enzymatic reactions. Enzymes are reported with a three letter code: PAL, phenylalanine ammonia lyase; TAL, tyrosine ammonia lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate CoA ligase; COMT, caffeic acid/5-hydroxyferulic acid O-methyltransferase; HCT/CST, hydroxycinnamoyl CoA:shikimate/quinate hydroxycinnamoyltransferase; C3H, p-coumaroyl shikimate/quinate 3-hydroxylase; CCoAOMT, caffeoyl CoA O-methyltransferase; CCR, (hydroxy)cinnamoyl CoA reductase; CAD, (hydroxy)cinnamyl alcohol dehydrogenase; F5H ferulate 5-hydroxylase; CHS, chalcone synthase; STS, stilbene synthase.

The initial physiological advantage of phenolic compounds is not clear. In fact, flavonoids are not thought to have been immediately effective as UV protection before the emergence of complex structures allowing for their accumulation in large quantities, and it has been proposed that they were initially used as internal signaling molecules [7]. Lignin-like polymers have been identified in the cell walls of the charalean alga *Nitella* and in bryophytes (mosses, liverworts, and hornworts), early branching lineages of land plants that do not harbor a developed vascular system such as that found in Tracheophytes (Ferns, Gymnosperms and Angiosperms) [8]. Because these lignin monomers in non vascular plants do not fulfill structural functions it has been proposed that may principally serve as a defense mechanism against microorganisms or UV radiation [8]. To date, there is no evidence for the presence of a full phenylpropanoid metabolism in organisms other than land plants, although some bacteria and fungi harbor homologues of a few enzymes of the pathway [9,10].

The phenylpropanoid pathway likely evolved progressively in land plants by the recruitment of enzymes from the primary metabolism (for a recent review see 4). However, the origin of PAL was a key event, since it provided the initial step from which the rest of the pathway was assembled. Indeed, PAL is a key regulator of the phenylpropanoid pathway [11] and any inhibition of PAL blocks the whole pathway. Probably due to its essentiality, land plants harbor multiple copies of PAL [5] and no complete null mutant is available in the literature. De novo synthesis of PAL is induced in response to different stress stimuli such as UV irradiation, pathogenic attack, low levels of nitrogen, phosphate, or iron [6]. Although PAL enzymes have been extensively characterized in all land plants lineages, including the early emerging bryophytes (mosses, liverworts, and hornworts), their distribution in other organisms is limited. PAL are known to be present in fungi, in particular Basidiomicetes yeasts such as *Rhodotorula*, but also Ascomycetes such as *Aspergillus* and *Neurospora*, where they participate to the catabolism of phenylalanine as a source of carbon and nitrogen [12-14].

The PAL of some plants and fungi also harbor a tyrosine ammonia lyase (TAL) activity that is responsible for the synthesis of p-coumaric acid directly from tyrosine, which in turn leads to the production of p-coumaroyl-CoA [4] (Figure 1). PAL enzymes have been functionally characterized from a few sediment/soil bacteria such as *Streptomyces maritimus* (Actinobacteria), where PAL is required to supply cinnamic acid for the production of benzoyl-CoA, the starter molecule for the biosynthesis of the bacteriostatic agent enterocin [15], and *Photorhabdus luminescens* (γ-Proteobacteria), where PAL is essential for the production of the powerful stilbene antibiotic through yet unknown intermediate steps [16,17]. More recently, PAL have also been identified and structurally characterized in two Cyanobacteria belonging to the order Nostocales, where they are involved in a pathway whose end product is yet unknown [18]. From functional studies, it has been proposed that these cyanobacterial PAL might represent an evolutionary intermediate towards plants PAL [18]. PAL homologues with TAL activity have also been identified in some bacteria such as the Actinobacterium *Saccharotrix espanaensis*, where they are used to produce the antibiotic saccharomicin [19], and in purple phototrophic a-Proteobacteria such as *Rhodobacter*, where they are involved in the synthesis of the chromophore of their photoactive yellow protein photoreceptor [20].

PAL is homologous to histidine ammonia lyase (HAL), which is involved in the catabolism of histidine and is widespread in prokaryotes and eukaryotes [21,22]. It has been proposed that "PAL developed from HAL when fungi and plants diverged from the other kingdoms" [4]. However, the current view of eukaryotic evolution based on phylogenetic analyses indicates that fungi and plants do not share an exclusive ancestor [23,24]. In fact, Fungi are more related to Animals than to land plants. Moreover, land plants belong to the phylum Plantae, which also includes Glaucocystophytes, red algae, and green algae [23,24].

Given the clear importance of PAL in the emergence of the phenylpropanoid pathway and adaptation of plants to land, we sought to get more insight into the origin of this enzyme by carrying out an extensive search of PAL/TAL/HAL homologues in current sequence databases and by analyzing their phylogeny.

Results

Based on preliminary exhaustive phylogenetic analyses, 160 representative sequences were chosen for final tree construction (i.e. a selection of bacterial homologues including all characterized PAL and their closest homologues, all archaeal homologues, a selection of plant and fungi homologues, and all homologues for the remaining eukaryotic phyla, see Methods for details. These sequences are very well conserved and allowed the selection of 369 unambiguously aligned amino acid positions for analysis. The resulting unrooted bayesian tree is shown in Figure 2 (see Additional file 1 and 2 also). The prokaryotic part of the tree is not congruent with species phylogeny, indicating extensive gene duplications, losses, and horizontal gene transfer (HGT) within bacteria as well as between bacteria and archaea, which makes it difficult to retrace the evolutionary history of PAL/TAL/HAL enzymes in prokaryotes. The few characterized bacterial PAL and TAL are not monophyletic (Figure 2, indicated in red and light blue font, respectively), although their close relatives may

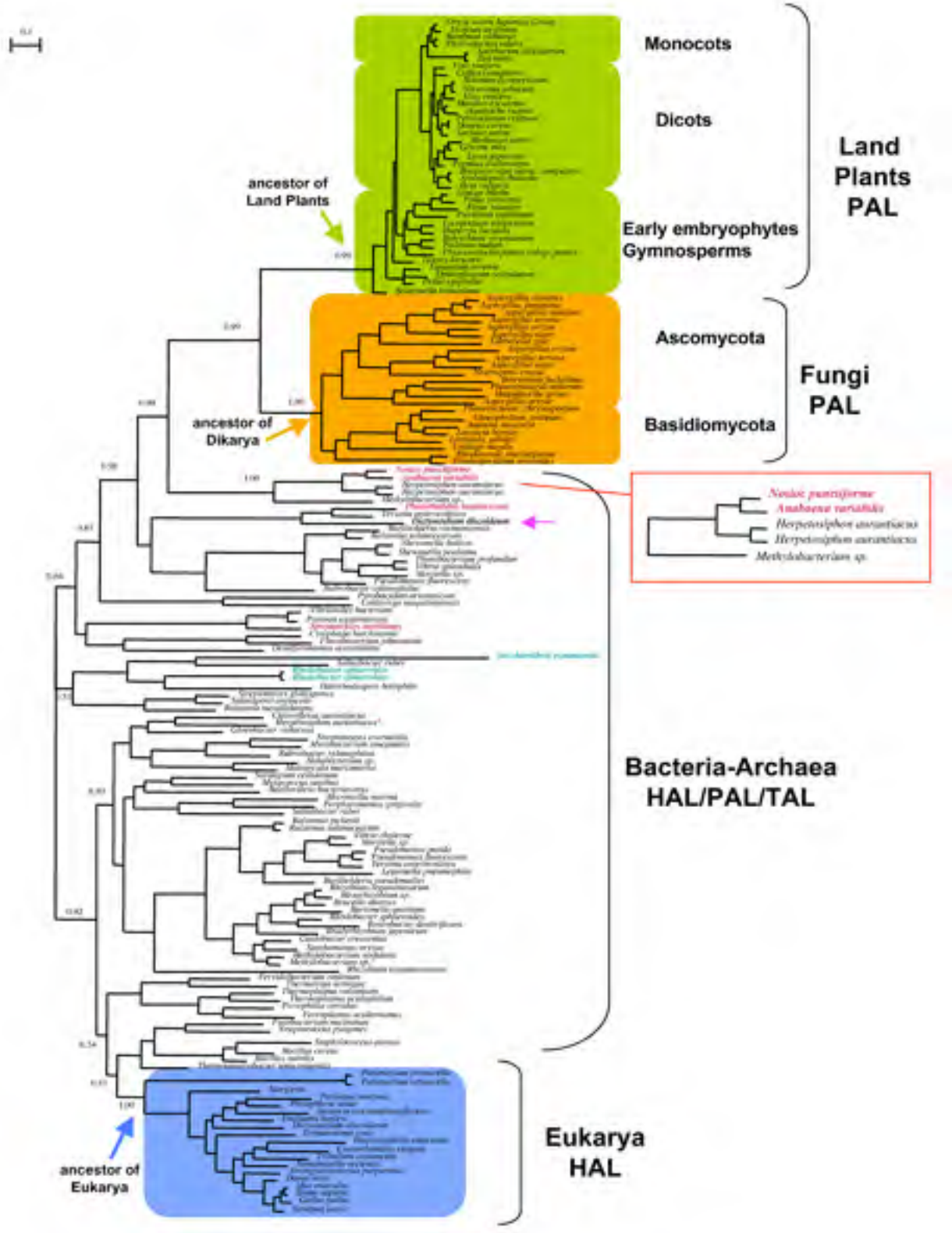


Figure 2 *(see legend on next page)*

Figure 2 *(see previous page)*
Phylogeny of PAL/TAL/HAL. Unrooted bayesian tree of a representative sampling of PAL/TAL/HAL homologues. Characterized bacterial PALs are shown in red font, while characterized bacterial TALs are shown in blue font. Although it is difficult to decide where the root lies, it is clear that eukaryotic HAL (blue square) and fungi/land plants PAL (orange and green squares, respectively) have distinct origins. Moreover, taxonomic distribution of HAL and PAL orthologues indicates that the ancestor of eukaryotes harbored a HAL (blue arrow) while a PAL was introduced by HGT in the ancestor of Dikarya fungi (orange arrow) and the ancestor of land plants (green arrow). The source of this HGT is likely in a group of sediment/soil bacteria including characterized cyanobacterial PAL and uncharacterized sequences from *Methylobacterium* sp. and *Herpetosiphon aurantiocus* (red square). Probable HAL orthologues of *Methylobacterium* sp. and *Herpetosiphon aurantiocus* are indicated by red asterisks. The amoebozoan *Dictyostelium discoideum* appear to have acquired a PAL in the course of a recent HGT from soil bacteria (pink arrow). Numbers at nodes represent posterior probabilities (for clarity only PP relevant for discussion are indicated). The scale bar represents the average number of substitutions per site. The same tree with full accession numbers and PP is provided as Additional file 1. A maximum likelihood analysis gave very similar results and is provided as additional file 2.

also be PAL or TAL and it would be interesting to characterize them.

Eukaryotic sequences form two distinct well-supported monophyletic clusters, one including characterized HAL (Figure 2, light blue rectangle, posterior probability PP = 1.00) and the other including characterized PAL (Figure 2, green and orange rectangles, PP = 0.99), separated from each other by bacterial/archaeal homologues. This clearly indicates that eukaryotic PAL and HAL have distinct evolutionary origins, since otherwise eukaryotic PAL should arise from within eukaryotic HAL. HAL homologues are not present in all complete eukaryotic genomes, indicating that catabolism of histidine is not an essential function. However, the cluster of eukaryotic HAL orthologues includes members of major phyla [23,24] such as Alveolates (*Paramecium tetraurelia*, *Perkinsus marinus*), Amoebozoa (*Dictyostelium discoideum*), Haptophytes (*Emiliania huxleyi*), Heterokonts (*Aureococcus anophagefferens*, *Phytophtora soyae*), Excavates (*Naegleria*, *Trypanosoma cruzi*), and Metazoans (Figure 2). This strongly suggests that a HAL was present in the most recent eukaryotic ancestor and the absence of HAL orthologues in some eukaryotic lineages has to be interpreted as a consequence of gene loss. For example, we found no HAL orthologues in any of the fungal lineages for which complete genome sequence data is currently available (i.e. Ascomycota and Basidiomycota, which form the Dikarya [25]), although we retrieved a orthologue in the EST database at NCBI from *Blastocladiella emersonii*, which belongs to the early emerging aquatic lineage Chytridiomycota [25]. This suggests that a HAL orthologue may have been present in the ancestor of Fungi and was secondary lost in Dikarya, but needs confirmation when complete genome sequences will be available from Chytridiomycota and other early emerging fungal lineages. We found no HAL orthologues in any member of the phylum Plantae for which complete genome data is available (i.e. the red algae *Cyanidioschyzon merolae*, the green algae *Chlamydomonas* and *Ostreococcus*, and the Angiosperms *Oryza sativa* and *Arabidopsis thaliana*) [23,24], indicating an early gene loss in this phylum. Intriguingly, *D. discoideum* harbors two additional homologues: one is very divergent and could not be included in the analysis, while the other lies outside of the eukaryotic HAL cluster and close to the characterized PAL of *P. luminescens* (Figure 2, indicated by a pink arrow) and may represent a recent acquisition by HGT. It would be interesting to investigate the role of this putative PAL homologue in *Dictyostelium*, in particular to verify whether it also has an antimicrobial defense role in this soil-dwelling eukaryote, which has been recently suggested to harbor a rudimentary immune system [26].

In contrast to the wide distribution of eukaryotic HAL orthologues, the eukaryotic PAL cluster contains exclusively orthologues from plants and fungi but no other eukaryotic lineage, and these form two well-supported monophyletic sister groups (Figure 2, green and orange rectangles PP = 0.99 and 1.00, respectively). However, the plant PAL cluster includes only members from land plants (including all early emerging lineages such as mosses, hornworts, and liverworts [1,2]), but we found no orthologues in available genomic data from the red and green algae lineages, which branch prior to the divergence of land plants within the phylum Plantae [23,24]. The monophyly of plants PAL orthologues strongly indicates that they have a single origin and derive from a gene that was already present in the ancestor of land plants [27] (Figure 2). Concerning Fungi, the PAL cluster contains the few characterized fungal enzymes (*Amanita*, *Rhodotorula*, *Aspergillus*) and is thus likely that the other orthologues have also PAL activity, although more functional data is needed to verify this. We found PAL orthologues in all complete genomes that are currently available (i.e. exclusively from Dikarya), to the exception of the late emerging lineages Saccharomycotina, Schizosaccharomycetes, and *Cryptococcus* [25], indicating secondary gene losses. We found no PAL orthologues in available genomic data of the fungal lineages Chytridiomycota and Zygomycota, which branch prior to the divergence of Dikarya in the

phylogeny of the phylum Fungi [25]. This may indicate absence of a PAL coding gene in these lineages, although this needs to be verified when complete genome data becomes available. The monophyly of the fungal PAL orthologues strongly indicates that they have a single origin and derive from a gene that was already present at least in the ancestor of Dikarya (Figure 2), and possibly earlier.

The evolutionary relatedness of PAL orthologues from land plants and fungi clearly indicates a common origin. However, the phylum Plantae does not share an exclusive ancestor with Fungi [23,24], i.e. the most recent common ancestor of these two eukaryotic lineages corresponds the most recent common ancestor of all eukaryotes (Figure 3). Consequently, if a PAL orthologue was present in the ancestor of all eukaryotes, it would have been subsequently lost in all eukaryotic lineages to the exception of land plants and fungi (Figure 3a). A more parsimonious scenario is one where a PAL originated either in the ancestor of land plants or in the ancestor of at least Dikarya fungi and then was transferred *via* HGT between these two phyla (Figure 3b, dotted arrows). Although the prokaryo-

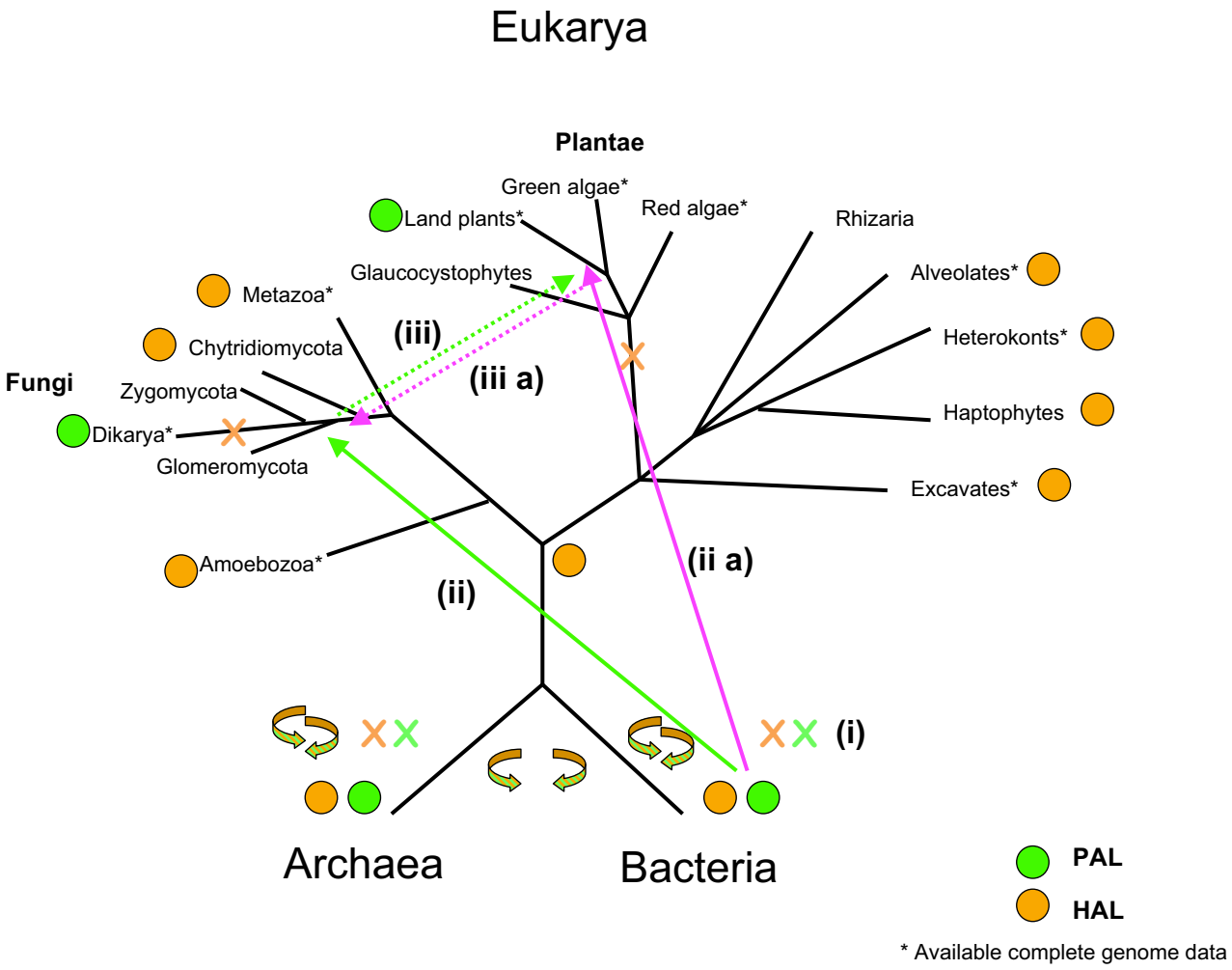


Figure 3
An evolutionary scenario for the origin of plant PAL. A HAL coding gene (orange circle) was present in the most recent eukaryotic ancestor, based on its presence in all major eukaryotic supergroups for which sequence data is available (indicated by an asterisk), and it was lost in the ancestor of Dikarya Fungi and in the ancestor of the phylum Plantae (orange crosses). In contrast, the origin of eukaryotic PAL is more recent: (1) origin of PAL in a bacterium (green circle), (2) HGT to fungi -Dikarya or possibly earlier (solid green arrow), (3) HGT from fungi to an ancestor of land plants (dashed green arrow). Alternatively: (1) origin of PAL in a soil bacterium (green circle), (2a) HGT to an ancestor of land plants (solid pink arrow), (3a) HGT from this ancestor to fungi (dashed pink arrow). Extensive HGT of PAL and HAL among and within bacteria and archaea are indicated by double rounded arrows and gene losses by green and orange crosses.

tic part of the tree is blurred by HGT, it is intriguing that a group of bacterial homologues including the two characterized cyanobacterial PAL [18] is robustly supported as sister of the land plants/fungi PAL cluster (PP = 0.99, Figure 2). Again, the heterogeneity of this bacterial group testifies for HGT between its members. In fact, it includes two distantly related sediment/soil bacteria, *Herpetosiphon aurantiacus* (Chloroflexi), and *Methylobacterium* sp. (α-Proteobacteria), a facultative methylotrophic pink pigmented relative of *Rhizobiales* [28,29]. These uncharacterized homologues may also be PAL since these bacteria harbor a second homologue that may be a *bona fide* HAL (indicated by a * symbol in Figure 2).

Discussion

During early colonization of emerged environments by pioneer land plant ancestors, beneficial associations with fungi and soil bacteria were likely crucial. In particular, it is known that N_2 fixing cyanobacteria formed symbioses with early fungal lineages (lichen-like or endocytobiotic symbioses, such as those between the glomeromycotan *Geosiphon pyriformis* and the cyanobacterium *Nostoc* [30]) as well as with land plants, and that fungi (Glomeromycota) started arbuscular-mycorrhizal (AM) symbioses with the first land plants [30-34].

The peculiar distribution and phylogeny of plant PAL suggests a plausible scenario for its origin: PAL emerged in bacteria (Figure 3(i)), likely with an antimicrobial role; a member of an early fungal lineage (i.e. at least before the divergence of Dikarya) obtained a PAL *via* HGT from a bacterium (possibly a Nostocale or another soil/sediment bacterium through an early symbiosis [30]) (Figure 3(ii)); this fungal PAL was transferred to an ancestor of land plants *via* an ancient AM symbiosis (Figure 3(iii)), where it paved the way for the development of the phenylpropanoid pathway, and the radiation of plants on terrestrial environments. The fact that land plants PAL do not appear to arise from within fungi PAL (Figure 2) can be explained by the fact that the donor was the ancestor of Dikarya, or by the fact that the donor belonged to a lineage predating the divergence of Dikarya and we either still lack complete genome sequence data from it or the lineage has gone extinct. Important insights to test this evolutionary scenario will be obtained by the future availability of genomic data from early emerging fungal lineages such as Glomeromycota, that possibly emerged before land plants [30] and were most likely the first fungi to form AM type symbioses with them [30,31].

We cannot exclude *a priori* a transfer in a different direction, i.e. from a soil bacterium to an ancestor of land plants *via* an ancient symbiosis (Figure 3(ii a)), then from this to an ancestor of Dikarya fungi (or an earlier branching lineage for which sequence data is not yet available) *via* an ancient AM symbiosis (Figure 3(iii a)). Nevertheless, we wish to stress that land plants PAL are unlikely to be of chloroplastic origin. In fact, since chloroplasts in the phylum Plantae derive from a single primary cyanobacterial endosymbiosis, if land plant PAL had been inherited from the cyanobacterial ancestor of the chloroplast this would imply at least two independent losses of PAL in red and green algal lineages, which postdate the acquisition of the primary chloroplast but emerged prior to the divergence of land plants (Figure 3) [1,2,23,24]. Moreover, only 3 out of the 36 currently available complete cyanobacterial genomes harbor a PAL/HAL homologue (i.e. only *Gloeobacter* in addition to the two Nostocales). PAL is a cytoplasmic enzyme and is not targeted to the chloroplast (we tried to assess the probability of plastid targeting of PAL using the predictions software Predotar V1.03 [35], obtaining no significant results (data non shown). The ancestor of the phylum Plantae likely preceded the ancestor of land plants of many millions of years. If a PAL was transferred by Endosymbiotic Gene Transfer from the cyanobacterial symbiont to the host nucleus in the ancestor of the phylum Plantae, it is not clear why it would have been lost multiple times independently in 2–3 algal lineages, indicating a lack of selective advantage, while being maintained only in the algal line leading to land plants up to around 500 million years ago. Finally, it is possible that the ancestor of land plants and the ancestor of fungi independently acquired their PAL from two different but related bacteria. Importantly, since the gene coding for HAL appears to have been lost early in the phylum Plantae, the phenylpropanoid pathway in land plants could not have been emerged without the acquisition of a PAL homologue by HGT.

Conclusion

The origin of land plants was a key event in the history of life on our planet since it played a fundamental role in the evolution of modern terrestrial ecosystems. The contribution of bacteria to eukaryotic innovations is considered important, but remains poorly explored. Our results highlight the crucial role of HGT from soil bacteria in the emergence of key metabolic pathways such as that of phenylpropanoids, and therefore in the path leading to land colonization by plants and their subsequent evolution.

Since it is likely that the phenylpropanoid pathway took some time to be fully assembled, it is intriguing to speculate about the original selective advantage to keep a horizontally acquired PAL in the first land plants. The direct products of PAL are cinnamate and p-coumarate. These might have been used as an antimicrobial, such as in some bacteria, and would have played a fundamental role as protection from attack by an already developed microbial soil community. Alternatively (or in combination with an antimicrobial role), they might have provided

protection against UV radiation, for example being the precursor of a light capturing pigment such as in modern purple bacteria. Moreover, cinnamate and p-coumarate are the precursors of benzoic acid and salicylic acid, which are known defense compounds [36,37] Finally, it is known that coumarins have appetite suppressing properties, suggesting that an initial role for PAL may have been to provide defense against grazing animals.

It would be interesting to know if fungi also use PAL for these purposes, and what are the corresponding mechanisms for UV shielding and antimicrobial defense in the green algae that are known to colonize soil habitats. To answer these questions, it will be important to investigate further the distribution of PAL enzymes in both bacteria and fungi, which may be more widespread than currently thought, as well as their role in still largely unexplored secondary metabolisms.

Methods

Exhaustive Blast searches were carried out by using different HAL and PAL sequences as seeds on the non-redundant sequence database and on the EST database at NCBI [38], on ongoing eukaryotic genomes at the DOE Joint Genome Institute [39], at the Broad Institute [40], and at the *Cyanidioschyzon merolae* Genome Project web service [41].

Based on exhaustive preliminary phylogenetic analyses, 160 representative taxa were chosen for final tree construction. From the global alignment, 369 unambiguously aligned amino acid positions were selected for analysis. Tree reconstruction was performed using the bayesian method implemented in MrBayes [42] with a mixed model of amino acid substitution and a gamma correction (eight discrete categories plus a proportion of invariant sites) to take into account among-site rate variations. MrBayes was run with four chains for 1 million generations and trees were sampled every 100 generations. To construct the consensus tree, the first 1500 trees were discarded as "burnin.". Maximum likelihood analysis of the same dataset was carried out by using Phyml [43], with a WAG model of amino acid substitution, including a gamma law with 4 categories to take into account differences in evolutionary rates at sites, and an estimated proportion of invariable sites.

Abbreviations

PAL: phenylalanine ammonia lyase; TAL: tyrosine ammonia lyase; HAL: histidine ammonia lyase; HGT: horizontal gene transfer; CH4: cinnamic acid 4-hydrolase; 4CL: p-coumaroyl:CoA ligase.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GE, MF, RF conceived the study, GE MF and SG performed the analyses and all authors drafted the manuscript. All authors read and approved the final manuscript.

Reviewers' comments

Reviewer's report 1

Purificación López-García

This article presents an extensive molecular phylogenetic analysis of phenylalanine ammonia lyase (PAL, many of which also use tyrosine as substrate), the enzyme catalyzing the first step of the phenylpropanoid pathway leading, in plants and some fungi, to the synthesis of flavonoid secondary metabolites and lignin monomers. The study includes also the related enzyme histidine ammonia lyase (HAL), widespread in the three domains of life. Since land plants and dikaryotic fungi PAL form two sister monophyletic clades clearly distinct from eukaryotic HAL and from their prokaryotic homologues, it is proposed that PAL was transferred horizontally from bacteria to land plants or to fungi and, subsequently, from land plants to fungi or viceversa. This is an interesting observation, well supported by the phylogenetic analysis presented, that leads the authors to hypothesize a key role of this enzyme for the adaptation of plants to land.

I have two major comments. First, the hypothesis that a horizontal gene transfer of PAL to the land plant ancestor is at the origin of the phenylpropanoid metabolism and of their adaptation to terrestrial ecosystems is appealing. However, a single enzyme does not make a pathway and, in the absence of data about the remaining genes involved in phenylpropanoid metabolism, this idea remains hypothetical. In this sense, the title of the article appears too conclusive (*A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land*). Have the authors tried to make preliminary phylogenetic analyses for other genes in the pathway or, at least, do they have an idea about their phylogenetic distribution? It would be interesting to compare the distribution of enzymes involved in flavonoid and lignin monomer biosynthesis with that of PAL.

AU: We clarified the text to explain that the whole pathway is a specificity of land plants, although a few bacteria and fungi harbor homologues of some enzymes of the pathway. The assembly of the pathway in land plants likely occurred stepwise by the recruitment of preexisting enzymes from other metabolic routes. Although it will be surely interesting to investigate further how this occurred, we now stress in a more clear way that we addressed specifically the very origin of the pathway, which could not have occurred without the acquisition of PAL, since this enzyme performs the first and essential step. Moreover, we precise that such HGT of PAL was essential, since a HAL homologue was likely lost early in the phylum Plantae and therefore

land plants could not have assembled the pathway by recruiting a preexisting HAL. Even if the other genes of the pathway were also acquired by HGT, this would have occurred either simultaneously or after the acquisition of PAL. For this reason, our title appears to us justified.

Nevertheless, we have performed preliminary analysis of the two enzymes following PAL in the general phenylpropanoid pathway (C4H and 4CL) and these are large gene families that do not appear to show a pattern similar to PAL, supporting the idea that they were recruited from preexisting pathways and strengthening the importance of the HGT of PAL.

My second comment relates to the primary selective advantage attributed to the acquisition of PAL from bacteria, which might have been the production of antimicrobial or pigmented metabolites that would allow the successful competition of land plants/fungi in soils or protection against UV light. Again, the idea is attractive but, to prove it, would require as a preliminary step to show that the whole flavonoid biosynthesis pathway emerged prior to that of lignin monomer biosynthesis.

If the latter appeared first, one could propose instead that the advantage of acquiring this pathway was to increase stiffness and developing the ability to construct rigid structures, an essential property of land plants and some stages of many fungal life cycles. Perhaps the authors can consider this possibility or discuss why they think it is unlikely. In addition, green algae, which also colonize soil surfaces, have also to compete with other members of the microbial community and to protect themselves from UV. They might have preferred to keep their own, non PAL-derived, protective systems against microbes and UV light.

AU: We now clarify in the text that early branching land plant lineages harbor the first enzymes of the two main branches of the pathway leading to lignin monomers and flavonoids. Unfortunately, the unavailability of genomic data from earlier lineages (e.g. Charales) prevents understanding for the time being which of the two branches emerged first. We now discuss briefly the production of lignin-like monomers in non-vascular early emerging land plants where these are likely used as defense against either UV or microorganism attack. To our knowledge fungi consume lignin but do not produce it, they construct rigid structures by using chitin.

We speculate that the initial selective advantage of PAL that would have lead to the fixation of the HGT may have had to do with the use of its direct products, cinnamic acid and p-coumarate, both involved in antimicrobial or anti UV functions in bacteria and possibly fungi. Moreover, cinnamate and p-coumarate are the precursors of benzoic acid and salicylic acid, which are known defense compounds.

The remark on green algae colonizing soil habitats is very interesting. We now discuss it in the text.

Alternatively, the authors might wish to consider the possibility that flavonoid synthesis did not confer a particular efficient protection against microorganisms, but against metazoan grazers, which constitute indeed the major threat for land plants.

AU: interesting point, although we do not address specifically the origin of flavonoid production (see above), coumarins have appetite suppressing properties, suggesting its widespread occurrence in plants, especially grasses, is because of its effect of reducing the impact of grazing animals. Thus an immediate advantage of PAL (TAL) might have also been defense against grazers. We now mention it in the text.

Reviewer's report 2

Eugene Koonin

This straightforward phylogenetic study of Phenylalanine Ammonia Lyase (PAL), the first committed enzyme of the phenylpropanoid pathway, reveals the monophyly of PALs from land plants and dikaryal fungi, with this eukaryotic branch embedded with a highly diverse bacterial tree. The interpretation of this result favored by the authors is that the ancestor of dikaryal fungi acquired the PAL gene from a soil bacterium and passed the gene to the ancestor of land plants. This conclusion implies a key role of HGT in the land colonization by plants.

I think this study highlights both the huge advantages and the considerable headaches that are associated with having numerous genome sequences from all walks of life. The conclusion made by the authors is, of course, interesting and plausible but it is by no means the only one that is possible to make from the tree shown in Figure 2. The main problem, as with most scenarios that involve HGT, is that we do not know the relative likelihoods of HGT and gene loss (but we do know that gene loss in many eukaryotic lineages was extensive). Therefore, arguments for HGT are doomed to remain mostly qualitative and often less than conclusive. With this in mind, potential alternative scenarios for PAL include (but are not necessarily limited to): i) presence in the last common ancestor of the extant eukaryotes with subsequent loss in all lineages (with sequenced genomes) except for land plants and dikaryal fungi; the authors briefly discuss this possibility and dismiss it as unlikely, and generally, one tends to agree (the number of required losses is quite large) but just how unlikely this possibility is, is still hard to tell;

AU It is indeed hard to tell, but we think that one HGT is a much more parsimonious scenario than massive independent losses in all eukaryotes apart from land plants and fungi. We now explain it more clearly in the text.

ii) HGT from the chloroplast to the common ancestor of all Plantae, with subsequent loss in algae, followed by HGT to dikaryal fungi; in the manuscript, this scenario is also dismissed as a highly unlikely one but, in this case, I am not sure I agree as the bacterial sister group of the eukaryotic PALs does include some cyanobacteria, and a loss of the gene in 2–3 algal lineages is not unlikely;

AU: at least two reasons make us think that this scenario is unlikely:

First, even if Nostoc is considered the extant cyanobacterium most similar to the first photosynthetic endosymbiont, only three cyanobacteria over 36 complete genomes harbor PAL/HAL homologues. No chloroplastic genomes harbor a PAL nor a HAL homologue. Furthermore, PAL is a cytoplasmic enzyme and is not targeted to the chloroplast (we tried to assess the probability of plastid targeting of PAL using the predictions software Predotar V1.03, obtaining a not significant result).

Second, the ancestor of the phylum Plantae likely preceded the ancestor of land plants of many millions of years. If a PAL was transferred by Endosymbiotic Gene Transfer from the cyanobacterial symbiont to the host nucleus in the ancestor of the phylum Plantae, it is not clear why it would have been lost multiple times independently in 2–3 algal lineages, indicating a lack of selective advantage, whereas it would have been maintained only in the algal line leading to land plants up to around 500 million years ago.

We therefore think that a pal EGT from the cyanobacterial endosymbiont to the host nucleus, although it cannot be excluded a priori, is not a scenario more supported than the one that we propose.

iii) independent HGTs from related (soil) bacterial to plants and fungi – a possibility that is not discussed in the manuscript but that, as far as I see, cannot be ruled out.

AU: We included this possibility in the text.

The above alternatives to the authors' conclusion do not invalidate the work but it must be admitted that, e.g., the chloroplast scenario is less surprising than the one presented by the authors, so much so that the advisability of dedicating a special papers to the origin of PAL in plants and fungi could be questioned. My disappointment with the manuscript is that the authors do not investigate the phylogeny of other enzymes of the phenylpropanoid pathway. Had this been done and had a coherent pattern been discovered, the conclusions could be much more convincing and exciting. If, on the other hand, such a coherent pattern does not exist, this also would be notable indicating that, like many other systems, this key pathway is a patchwork of genes of different origins. I understand, of course, that such a complete phylogenetic analysis requires a considerable amount of extra work, so the authors might prefer to highlight the PAL analysis separately, but I still think that a more comprehensive paper will be of greater value.

AU: As we explained in our answer to referee 1, we now clarified better in the text that in this report we wished to focus on the very first step in the origin of the pathway that was key to its further assembly. How the pathway was then assembled is surely an interesting question but we feel not directly relevant to our hypothesis. Since without the acquisition of PAL the pathway could not have been be assembled, in particular because of the absence of a preexisting HAL homologue from which a PAL may have been derived, we reckon that our analysis is not incomplete.

Indeed, as the referee points out, it would be more exciting not seeing the same pattern for the other genes, and this is what appears from preliminary analysis (see answer to referee 1).

At a more technical level, I think that it is highly desirable to also include result from a maximum likelihood analysis to buttress those obtained with the superoptimistic MrBayes. With just one family to analyze, this will not take too much effort.

AU: this analysis was in fact already done and gave very similar results and statistical support, we now mention it in the text and included the tree as supplementary material 2.

Reviewer's report 3

Janet Siefert

It's a beginning insight into land plant colonization. I think that other reviewers might have some issues with the argument being based on this one enzyme. I have to admit I did wonder myself about other key enzymes in the phenylpropanoid pathway. I think to help your cause in this regard, you should make a definition of what you mean by the 'first committed step' when you are speaking of the PAL enzyme. The team does a reasonably good job of speculating why the ancestor to land plants might have acquired this gene and it's beneficial use. In figure 2... you need a little bit more information on the methodology for this tree.. it is of course drawn as if it is rooted, is it? This time element to this tree that helps to support your argument presented, is stronger if it is.

AU: the referee is right, we added some clarifying comments in the legend to figure 2.

Additional material

Additional file 1

Unrooted bayesian tree of Figure 2 with full accession numbers and posterior probabilities.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-6150-4-7-S1.ppt]

Additional file 2

Unrooted ML tree of the same dataset. *Numbers at nodes represent non-parametric bootstrap values calculated on 100 bootstrapped samples of the original alignment calculated by Phyml (for clarity, not all are shown). For both trees, when no accession number is indicated, the corresponding sequence was retrieved from either the EST database at NCBI or from ongoing genome projects at JGI. EST_chimera indicates chimeric sequences obtained from two different EST sources of the same species.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-6150-4-7-S2.ppt]

Acknowledgements

SG wishes to thank the French Agence Nationale de la Recherche for support (ANR 07-JC-JCJC-0094-01).

References

1. Kenrick P, Crane PR: **The origin and early evolution of plants on land.** *Nature* 1997, **389**:33-39.
2. Bateman RM, Crane PR, Di Michele WA, Kenrick PR, Rowe NP, Speck T, Stein WE: **Early evolution of land plants, phylogeny, physiology, and ecology of the primary terrestrial radiation.** *Annu Rev Ecol Syst* 1998, **29**:263-292.
3. Waters ER: **Molecular adaptation and the origin of land plants.** *Mol Phylogenet Evol* 2003, **29(3)**:456-463.
4. Ferrer JL, Austin MB, Stewart C Jr, Noel JP: **Structure and function of enzymes involved in the biosynthesis of phenylpropanoids.** *Plant Physiol Biochem* 2008, **46**:356-370.
5. Dixon RA, Achnine L, Kota P, Liu C-J, Reddy MSS, Liangjiang Wang L: **The phenylpropanoid pathway and plant defence -a genomics perspective.** *Mol Plant Pathology* 2002, **3**:371-390.
6. Dixon RA, Paiva NL: **Stress-Induced Phenylpropanoid Metabolism.** *The Plant Cell* 1995, **7**:1085-1097.
7. Stafford HA: **Flavonoid Evolution: An Enzymic Approach.** *Plant Physiol* 1991, **96**:680-685.
8. Ligrone R, Carafa A, Duckett JG, Renzaglia KS, Ruel K: **Immunocytochemical detection of lignin-related epitopes in cell walls in bryophytes and the charalean alga Nitella.** *Plant Systematic and Evolution* 2008, **270**:257-272.
9. Moore BS, Hertweck C, Hopke JN, Izumikawa M, Kalaitzis JA, Nilsen G, O'Hare T, Piel J, Shipley PR, Xiang XL, Austin MB, Noel JP: **Plant-like Biosynthetic Pathways in Bacteria: From Benzoic Acid to Chalcone1.** *J Nat Prod* 2002, **65**:1956-1962.
10. Seshime Y, Juvvadi PR, Fujii I, Kitamoto K: **Genomic evidences for the existence of a phenylpropanoid metabolic pathway in *Aspergillus oryzae*.** *Biochem Biophys Res Commun* 2005, **337(3)**:747-751.
11. Ro D-K, Douglas CJ: **Reconstitution of the Entry Point of Plant Phenylpropanoid Metabolism in Yeast (*Saccharomyces cerevisiae*).** *The Journal of Biological Chemistry* 2004, **279**:2600-2607.
12. Moore K, Subba Rao PV, Towers GHN: **Degradation of Phenylalanine and Tyrosine by *Sporobolomyces roseus*.** *Biochem J* 1968, **106(2)**:507-514.
13. MacDonald MJ, D'Cunha GB: **A Modern view of phenylalanine ammonia lyase.** *Biochem Cell Biol* 2007, **85**:273-282.
14. Nehls U, Ecke M, Hampp R: **Sugar and nitrogen-dependent regulation of an *Amanita muscaria* phenylalanine ammonium lyase gene.** *J Bact* 1999, **181**:1931-1933.
15. Xiang L, Moore BS: **Biochemical characterization of a prokaryotic phenylalalnine ammonia lyase.** *J Bact* 2005, **187**:4286-4289.
16. Williams JS, Thomas M, Clarke DJ: **The gene stlA encodes a phenylalanine ammonia-lyase that is involved in the production of a stilbene antibiotic in *Photorhabdus luminescens* TT01.** *Microbiology* 2005, **151**:2543-2550.
17. Eleftherianos I, Boundy S, Joyce SA, Aslam S, Marshall JW, Cox RJ, Simpson TJ, Clarke DJ, ffrench-Constant RH, Reynolds SE: **An antibiotic prduced by an insect-pathogenic bacterium suppresses host defences through phenoloxidase inhibition.** *PNAS* 2007, **104**:2419-2424.
18. Moffitt MC, Louie GV, Bowman ME, Pence J, Noel JP, Moore BS: **Discovery of two cyanobacterial phenylalanine ammonia lyases: kinetic and structural characterization.** *Biochemistry* 2007, **46**:1004-1012.
19. Berner M, Krug D, Bihlmaier C, Vente A, Müller R, Bechthold A: **Genes and enzymes involved in caffeic acid biosynthesis in the actinomycete *Saccharothrix espanaensis*.** *J Bact* 2006, **188**:2666-2673.
20. Kyndt JA, Meyer TE, Cusanovich MA, Meyer JJ, Van Beeumen JJ: **Characterization of a bacterial tyrosine ammonia lyase, a biosynthetic enzyme for the photoactive yellow protein.** *FEBS Lett* 2002, **512**:240-244.
21. Rother D, Poppe L, Viergutz S, Langer B, Rétey J: **Characterization of the active site of histidine ammonia-lyase from *Pseudomonas putida*.** *Eur J Biochem* 2001, **268**:6011-6019.
22. Taylor RG, Lambert MA, Sexsmith E, Sadler SJ, Ray PN, Mahuran DJ, McInnes RR: **Cloning and Expression of Rat Histidase: homology to two bacterial histidases and four phenylalanine ammonia-lyases.** *The Journal of Biological Chemistry* 1990, **265**:18192-18199.
23. Baldauf SL: **The Deep Roots of Eukaryotes.** *Science* 2003, **300**:1703-1706.
24. Simpson AG, Roger AG: **Eukaryotic evolution: getting to the root of the problem.** *Curr Biol* 2002, **12(20)**:R691-R693.
25. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung G-H, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schußler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lucking R, Budel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JS, Vilgalys R: **Reconstructing the early evolution of Fungi using a six-gene phylogeny.** *Nature* 2006, **443**:818-822.
26. Chen G, Zhuchenko O, Kuspa A: **Immune-like phagocyte activity in the social amoeba.** *Science* 2007, **317**:678-681.
27. Rausher MD: **The evolution of flavonoids and their genes.** In *The Science of Flavonoids* Edited by: Grotewold E. Springer; 2006:175-211.
28. Madhaiyan M, Suresh Reddy BV, Anandham R, Senthilkumar M, Poonguzhali S, Sundaram SP, Tongmin Sa: **Plant growth-promoting *Methylobacterium* induces defense responses in groundnut (*Arachis hypogaea* L.) compared with root pathogens.** *Curr Microbiol* 2006, **53**:270-276.
29. Lidstrom ME, Chistoserdova L: **Plants in the pink: cytokinin production by *Methylobacterium*.** *J Bact* 2002, **184**:1818.
30. Parniske M: **Arbuscular mycorrhiza: the mother of plant root endosymbioses.** *Nature Reviews Microbiology* 2008, **6**:763-771.
31. Brundeztt MC: **Coevolution of roots and mycorrhizas of land plants.** *New Phytologist* 2002, **154**:275-304.
32. Raven JA: **The evolution of cyanobacterial symbioses.** *Biology and Environment: Proceedings of the Royal Irish Academy* 2002, **1**:3-6.
33. Yuan X, Xiao S, Taylor TN: **Lichen-Like Symbiosis 600 Million Years Ago.** *Science* 2005, **308**:1017-1020.
34. Selosse MA, Le Tacon F: **The land flora: a phototroph-fungus partnership?** *Tree* 1998, **13**:15-20.
35. Small I, Peeters N, Legeai F, Lurin C: **Predotar: A tool for rapidly screening proteomes for *N*-terminal targeting sequenze.** *Proteomics* 2004, **4**:1581-1590.

36. Krebs HA, Wiggins D, Stubbs M, Sols A, Bedoya F: **Studies on the mechanism of the antifungal action of benzoate.** *Biochem J* 1983, **214**:657-63.
37. Klessig DF, Durner J, Noad R, Navarre DA, Wendehenne D, Kumar D, Zhou JM, Shah J, Zhang S, Kachroo P, Trifa Y, Pontier D, Lam E, Silva H: **Nitric oxide and salicylic acid signaling in plant defense.** *Proc Natl Acad Sci USA* 2000, **97**:8849-55.
38. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/]
39. **DOE Joint Genome Institute** [http://www.jgi.doe.gov/]
40. **Broad Institute** [http://www.broad.mit.edu/annotation/genome/rhizopus_oryzae/Info.html]
41. **Genome Project web service** [http://merolae.biol.s.u-tokyo.ac.jp/]
42. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19**:1572-1574.
43. Guindon S, Gascuel O: **PhyML: A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Systematic Biology* 2003, **52**:696-704.

6.2 Conclusion

The origin of land plants was a key event in the history of life on our planet since it played a fundamental role in the evolution of modern terrestrial ecosystems. The contribution of bacteria to eukaryotic innovations is considered important, but remains poorly explored. Our results highlight the crucial role of HGT from soil bacteria in the emergence of key metabolic pathways such as that of phenylpropanoids, and therefore in the path leading to land colonization by plants and their subsequent evolution. Since it is likely that the phenylpropanoid pathway took some time to be fully assembled, it is intriguing to speculate about the original selective advantage to keep a horizontally acquired PAL in the first land plants. The direct products of PAL are cinnamate and p-coumarate. These might have been used as an antimicrobial, such as in some bacteria, and would have played a fundamental role as protection from attack by an already developed microbial soil community. Alternatively (or in combination with an antimicrobial role), they might have provided protection against UV radiation, for example being the precursor of a light capturing pigment such as in modern purple bacteria. Moreover, cinnamate and p-coumarate are the precursors of benzoic acid and salicylic acid, which are known defense compounds. Finally, it is known that coumarins have appetite suppressing properties, suggesting that an initial role for PAL may have been to provide defense against grazing animals. It would be interesting to know if fungi also use PAL for these purposes, and what are the corresponding mechanisms for UV shielding and antimicrobial defense in the green algae that are known to colonize soil habitats. To answer these questions, it will be important to investigate further the distribution of PAL enzymes in both bacteria and fungi, which may be more widespread than currently thought, as well as their role in still largely unexplored secondary metabolisms.

Part II

Comparative Evolutionary Genomics

Chapter 7

Analysis of plasmids sequences

While bacterial chromosomes show a relatively high conservation of their architecture, plasmid molecules are more variable concerning gene content and/or organization, even at short evolutionary distances. Indeed, plasmid genes can be considered to be under differential selection, while moving around the bacterial community. Moreover they have a dynamic structure, i.e. genes can be gained or lost from the plasmid molecule. Actually, the same plasmid can be hosted by different organisms inhabiting different environments (e.g.: pH, temperature and chemical composition) and cohabiting with different genetic backgrounds. These factors may shape both the functional role(s) of the proteins, and the compositional features of plasmid DNA, such as GC or oligomers contents, some of the last being a very specific signature even at close phylogenetic distances. Despite their key role in the microbial world, at least two main issues concerning plasmids remain poorly investigated, that is the function of proteins they code for and their (sometimes complex) evolutionary dynamics. To overcome these limitations we have developed a bioinformatic package (Blast2Network, B2N) having three main aims:

1. to reconstruct the evolutionary history of plasmids molecules by identifying those having the most similar gene content.

2. To assign a putative function to previously uncharacterized proteins. This task is fulfilled in two ways: by means of sequence similarity of unknown or hypothetical proteins to known ones and through a phylogenetic profiling approach.

3. To provide an immediate visualization of the similarities existing among sequences. In fact, one of the outputs of the program is a network of sequence similarities, where proteins are represented by nodes and the shared identity values by links connecting them.

This approach (and/or some its implementations) was use to analyze:

1. plasmids harbored by members of the Enterobacteriaceae family of γ-Proteobacteria, which is one of the most studied divisions of bacteria and includes *Escherichia*, *Shigella*, and *Salmonella* genera, whose biomedical importance has allowed to record a relatively high number of completely sequenced plasmids in a few species (*this Chapter*).

2. the reticulate evolution of plasmids and chromosomes, focusing on the *Acinetobacter* genes (*Chapter 8*).

3. Analyze the horizontal flow of plasmids encoded resistome (i.e. genes involved in conferring resistance to antibiotics, *Chapter 9*).

This last issue is related to the more general issue of bacterial antibiotic resistance. This argument is related also to the last work of this dissertation (*Chapter 10*) were we analyzed the HAE1 and HME efflux systems in the *Burkholderia* genus.

7.1 *In silico* tools for plasmid sequences analysis: Blast2Network

Phylogenetic methods are well-established bioinformatic tools for sequence analysis, allowing to describe the non-independencies of sequences because of their common ancestor. However, the evolutionary profiles of bacterial genes are often complicated by hidden paralogy and extensive and/or (multiple) horizontal gene transfer (HGT) events which make bifurcating trees often inappropriate. In this context, plasmid sequences are paradigms of networklike relationships characterizing the evolution of prokaryotes. Actually, they can be transferred among different organisms allowing the dissemination of novel functions, thus playing a pivotal role in prokaryotic evolution. However, the study of their evolutionary dynamics is complicated by the absence of universally shared genes, a prerequisite for phylogenetic analyses. To overcome such limitations we developed a bioinformatic package, named Blast2Network (B2N), allowing the automatic phylogenetic profiling and the visualization of homology relationships in a large number of plasmid sequences. The software was applied to the study of 47 completely sequenced plasmids coming from *Escherichia*, *Salmonella* and *Shigella* spps. The tools implemented by B2N allow to describe and visualize in a new way some of the evolutionary features of plasmid molecules of Enterobacteriaceae; in particular it helped to shed some light on the complex history of *Escherichia*, *Salmonella* and *Shigella* plasmids and to focus on possible roles of unannotated proteins.

BMC Bioinformatics

BioMed Central

Research article

Open Access

Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network

Matteo Brilli[1,3], Alessio Mengoni[1], Marco Fondi[1], Marco Bazzicalupo[1], Pietro Liò[2] and Renato Fani*[1]

Address: [1]Department of Evolutionary Biology, University of Florence, via Romana 17, I-50125 Florence, Italy, [2]Computer Laboratory, University of Cambridge, 15 JJ Thompson Avenue, Cambridge, CB3 0FD, UK and [3]Laboratoire de Biometrie et Biologie Evolutive, UMR CNRS 5558, Lyon, France

Email: Matteo Brilli - brilli@biomserv.univ-lyon1.fr; Alessio Mengoni - alessio.mengoni@unifi.it; Marco Fondi - marco.fondi@unifi.it; Marco Bazzicalupo - marco.bazzicalupo@unifi.it; Pietro Liò - pl219@cam.ac.uk; Renato Fani* - renato.fani@unifi.it

* Corresponding author

Published: 21 December 2008

BMC Bioinformatics 2008, **9**:551 doi:10.1186/1471-2105-9-551

Received: 18 June 2008
Accepted: 21 December 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/551

Abstract

Background: Phylogenetic methods are well-established bioinformatic tools for sequence analysis, allowing to describe the non-independencies of sequences because of their common ancestor. However, the evolutionary profiles of bacterial genes are often complicated by hidden paralogy and extensive and/or (multiple) horizontal gene transfer (HGT) events which make bifurcating trees often inappropriate. In this context, plasmid sequences are paradigms of network-like relationships characterizing the evolution of prokaryotes. Actually, they can be transferred among different organisms allowing the dissemination of novel functions, thus playing a pivotal role in prokaryotic evolution. However, the study of their evolutionary dynamics is complicated by the absence of universally shared genes, a prerequisite for phylogenetic analyses.

Results: To overcome such limitations we developed a bioinformatic package, named Blast2Network (B2N), allowing the automatic phylogenetic profiling and the visualization of homology relationships in a large number of plasmid sequences. The software was applied to the study of 47 completely sequenced plasmids coming from *Escherichia*, *Salmonella* and *Shigella* spps.

Conclusion: The tools implemented by B2N allow to describe and visualize in a new way some of the evolutionary features of plasmid molecules of Enterobacteriaceae; in particular it helped to shed some light on the complex history of *Escherichia*, *Salmonella* and *Shigella* plasmids and to focus on possible roles of unannotated proteins.

The proposed methodology is general enough to be used for comparative genomic analyses of bacteria.

Background

Despite the huge amount of available sequences, few papers reported comparative analyses of entire plasmids with the aim of a complete classification of the functions they code for [1-4], and none considered all the sequences coming from entire genera or more inclusive taxonomic groups.

Nevertheless, plasmids are extremely important in microbial evolution, because they can be transferred between organisms, representing natural vectors for the transfer of genes and functions they code for [[5,6] and references therein]. In medical epidemiology and microbial ecology plasmids are thoroughly investigated because they often carry genes encoding adaptive traits such as antibiotic resistance, pathogenesis or the ability to exploit new environments or compounds [[7-9] and references therein].

While bacterial chromosomes show a relatively high conservation of their architecture, plasmid molecules are more variable concerning gene content and/or organization, even at short evolutionary distances. Indeed, plasmid genes can be considered to be under differential selection, while moving around the bacterial community. Moreover they have a dynamic structure, i.e. genes can be gained or lost from the plasmid molecule. Actually, the same plasmid can be hosted by different organisms inhabiting different environments (e.g.: pH, temperature and chemical composition) and cohabiting with different genetic backgrounds. These factors may shape both the functional role(s) of the proteins, and the compositional features of plasmid DNA, such as GC or oligomers contents, some of the last being a very specific signature even at close phylogenetic distances [10].

Despite their key role in the microbial world, at least two main issues concerning plasmids remain poorly investigated: i) the function of proteins they code for (see Additional file 1, more than 25% of proteins do not have assigned COG) and ii) the evolutionary dynamics of plasmids including their importance in bacterial evolution [11].

This latter point is often analyzed using phylogenetic methods that make use of rigorous statistical approaches to model the evolution of sequences (such as Maximum Likelihood or Bayesian inference). However, such methods are of restricted use in the case of plasmid molecules: they are computationally expensive when thousands of amino acid or nucleotide sequences are analyzed, and, moreover, require a set of homologous and universally shared sequences, that could be unavailable when studying plasmids.

To overcome these limitations we have developed a bioinformatic package (Blast2Network, B2N) having three main aims:

1) to reconstruct the evolutionary history of plasmids molecules by identifying those having the most similar gene content;

2) to assign a putative function to previously uncharacterized proteins. This task is fulfilled in two ways: by means of sequence similarity of unknown or hypothetical proteins to known ones and through a phylogenetic profiling approach. In this case the function of a protein is inferred by observing co-occurrence patterns. This is based on the idea that proteins involved in the same metabolic process or macromolecular complex tend to be maintained (or lost) together and that proteins which often occur together are likely to be functionally linked [12].

3) to provide an immediate visualization of the similarities existing among sequences. In fact, one of the outputs of the program is a network of sequence similarities in a format readable by the visualization software Visone http://visone.info/.

To test the package, we focused the attention on plasmids harbored by members of the Enterobacteriaceae family of γ-Proteobacteria, which is one of the most studied divisions of bacteria and includes *Escherichia*, *Shigella*, and *Salmonella* genera, whose biomedical importance [13] has allowed to record a relatively high number of completely sequenced plasmids in a few species. Moreover, horizontal transfer of plasmids between them has been described [14], complicating the phylogenetic information on plasmids; lastly, several pathogenesis-associated phenotypes are plasmid-borne [15]. Consequently, the application of B2N to this dataset could allow to reveal the presence of relationships between known pathogenesis-associated proteins and those which have not been characterized yet.

Methods

Description of the program

The procedure implemented in B2N is schematically reported in Figure 1a, but several tasks can be performed separately because of the modular nature of our software. The main workflow starts from a file containing protein or nucleic acid sequences in standard NCBI fasta format. This is used as an input to gather information on source sequences from the NCBI website. Several files are automatically generated for reference along with the corresponding nucleotide sequences for both genes and source sequences (e.g. the genome or the plasmid encoding the proteins used as input). Input sequences are then screened one against each other using BLAST [16]. The resulting output is parsed in the form of an adjacency matrix that describes the global sequence similarities in the dataset where each entry w_{ij} reflects the similarity existing between protein i and j. The user is initially prompted to choose two different selection criteria for alignments: an E-value threshold and an alignment length cut-off; after setting these parameters, all alignments passing the selection criteria are inserted in the matrix. Moreover, the user can specify the nature of the similarity score to be used, i.e. identity percentage or bit score; the bit score can also be normalized using the score of the alignment of the query with itself obtaining a value which is normalized on the

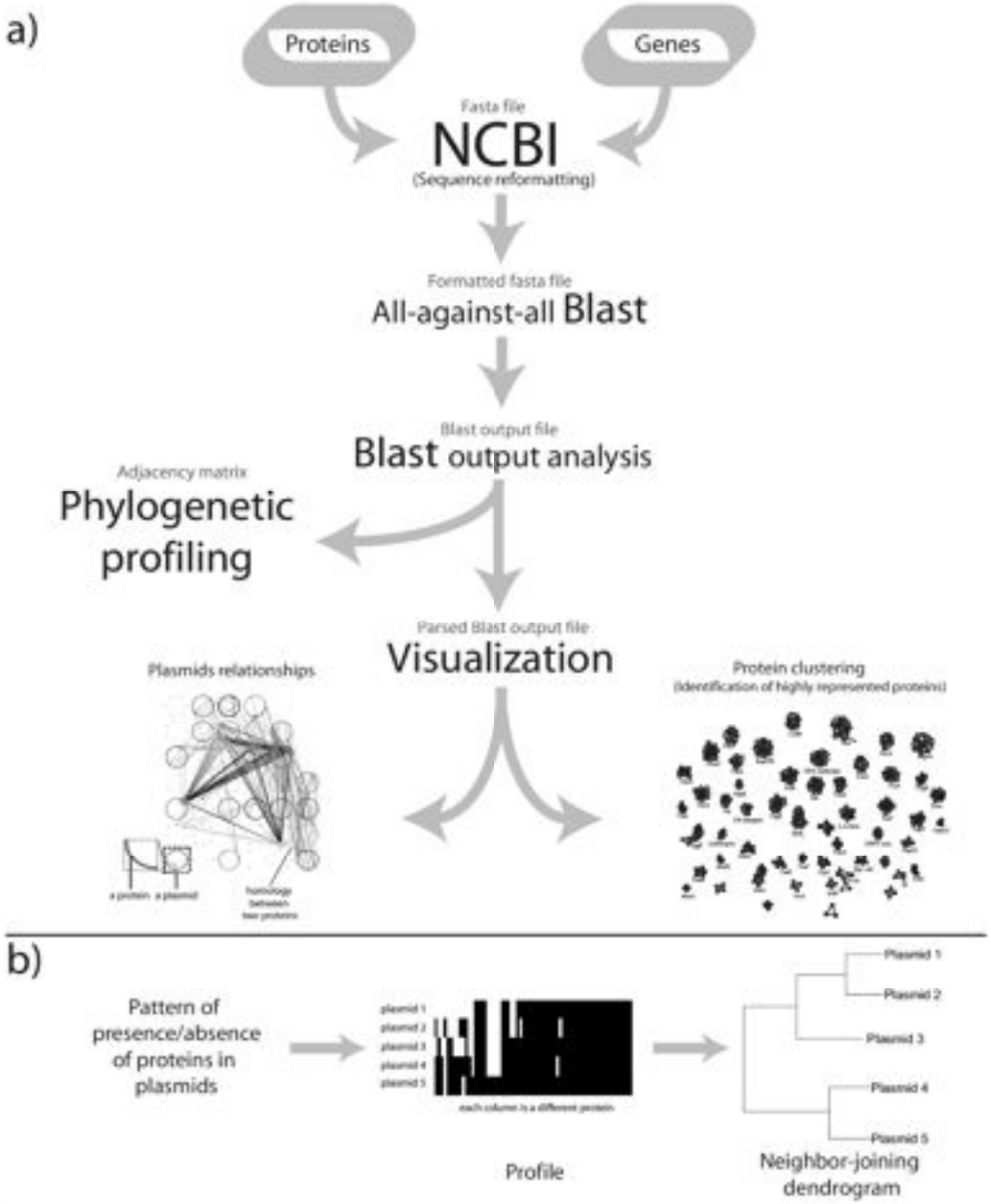


Figure 1
B2N workflow and analysis. a) Scheme of the data workflow in B2N. The visualization is realized using Visone software. The input of each module (i.e. the output of the previous one) is shown in red fonts. b) Phylogenetic profiling of molecules in the dataset. Using the matrix of occurrence patterns, groups of proteins are identified at different threshold values. A new matrix is obtained composed of a row for each plasmid in the dataset and a number of columns corresponding to the number of groups in the network. Each entry *i*, *j* of the matrix contains 1 if at least one protein from plasmid *i* is present in cluster *j*, 0 if no protein from plasmid *i* is present cluster *j*. This matrix is used for calculating distances using the Jaccard metric and dendrogram construction. This analysis identifies those plasmids that contain similar proteins. By applying the same workflow in the second dimension of the phylogenetic profiles matrix, it is possible to find those protein clusters having similar occurrence patterns.

alignment length. The weighted link values can be useful when comparing sequences from different species searching for those having the highest rate of horizontal transfer. This can be done in B2N specifying a distance matrix of house-keeping genes in Phylip format. The adjacency matrix obtained by parsing the BLAST output is the input for the phylogenetic profile method.

Phylogenetic Profiling

This approach allows the analysis of co-occurrence patterns, metabolic reconstruction and so on. In details, by taking as input the adjacency matrix storing the sequence similarity values, B2N produces a rectangular matrix (as described in the central part of Figure 1b) composed by all the plasmids under analysis (rows) and all the protein clusters (columns) identified through a depth-first search of the adjacency matrix. Each position of the phylogenetic profile matrix will be "1" in the case a given plasmid (row) possesses (at least) one protein in the corresponding protein cluster (column), whereas it is filled with "0" in the opposite case.

One of the commonly used metrics for binary data comparison is the Jaccard similarity coefficient. Given two vectors of phylogenetic profiles in binary form (A and B in this case, with n observations), the Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the sample sets: $J(A, B) = |A \cap B|/|A \cup B|$. The 'Jaccard distance', which measures dissimilarity between sample sets, is obtained by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union: $J_\delta(A, B) = |A \cup B| - |A \cap B|/|A \cup B| = 1 - J(A, B)$.

The Jaccard coefficient is a useful measure of the overlap that the attributes of 'A' and 'B' share. Each attribute of 'A' and 'B' can either be 0 ('absence') or 1 ('presence'). The total number of each combination of attributes for both 'A' and 'B' are specified as follows: M_{11} (M_{00}) represents the total number of attributes where 'A' and 'B' both have a value of 1 (0). M_{01} (M_{10}) represents the total number of attributes where the attribute of 'A' is 0 (1) and the attribute of 'B' is 1 (0). Each attribute must fall into one of these four categories, meaning that their sum equals n. The Jaccard similarity coefficient is $J = M_{11}/(M_{01} + M_{10} + M_{11})$. Blast2Network calculates the Jaccard distance for both dimensions of the phylogenetic profiles matrix, which corresponds to the distance between plasmids in term of shared genes, and the distance between occurrence patterns of clusters in plasmids. The Jaccard distance matrices are then used for the construction of two neighbor-joining dendrograms (Figure 1b). The first one describes similarities in gene content of the plasmids, the other one groups together those protein clusters with the most similar occurrence pattern within plasmids. Random permutations of the original data allows to compute the statistical significance of the Jaccard distances.

Network construction

B2N also outputs the BLAST post processing results as a network in Visone format http://visone.info/, a freely available software for network visualization and analysis. In doing so, it takes advantage of several information: the position and the color of the nodes (proteins) in the network correspond to the plasmid source, whereas the links indicate the existence of a given degree of sequence similarity between nodes. To reduce the dimensionality of the networks it is possible to use the Jaccard distance matrices to construct two hypergraphs where each plasmid or protein cluster, respectively, are collapsed to single nodes connected by edges whose values reflect the significance of the Jaccard distance calculated (see below and in Additional file 2).

Additional tools

B2N can include additional information in the network, assigning to each node a numerical (or binary) value which can be visualized in Visone as the size of the node: this node-associated value might be a compositional measure, such as the GC content and/or the codon adaptation index [17,18] of the corresponding gene. To this purpose, B2N has two methods but the user can input its own list of values as a text file. The first built-in method writes node values corresponding to the GC content of a sequence, while the other one implements the dinucleotide analysis derived from [10] and [19], obtaining a composition-based dissimilarity index of a gene sequence with respect to the source plasmid (or genome). Considering each possible dinucleotide, say *xy*, and a gene *s*, $_{xy(s)}^{m}$ $(f_{xy(s)}/f_{x(s)} * f_{y(s)})$. From this value the program obtains $_{(s,g)}$ $= 1/16 * \left| _{xy(s)} - _{xy(g)} \right|$ over all 16 dinucleotides, that is a measure of the compositional bias of a given sequence *(s)* with respect to a reference sequence *(g)* i.e. the genome or the entire plasmid. The can be used to detect genes that have been recently transferred and have since then maintained the compositional properties of the original plasmid.

Sequence data source and software availability

The dataset used in this work is composed by all the proteins encoded by the available completely sequenced plasmid sequences from *Escherichia*, *Shigella*, and *Salmonella* genera (Table 1). Complete plasmid sequences were downloaded from the NCBI ftp website ftp://ftp.ncbi.nih.gov/refseq/release/plasmid.

The software B2N with the user's manual can be directly requested to the authors and is also available as Additional file (Additional file 3).

Table 1: Plasmids analyzed

Plasmid	Organism	Length (nt)	# ORF	Accession Number
R721	*Escherichia coli*	75582	91	NC_002525
p9123	*Escherichia coli*	6222	8	NC_005324
pC15-1a	*Escherichia coli*	92353	100	NC_005327
pCol-1et	*Escherichia coli*	5847	7	NC_002487
pAPEC-O2-R	*Escherichia coli*	101375	119	NC_006671
pColK-K235	*Escherichia coli*	8318	7	NC_006881
pRK2	*Escherichia coli*	5360	6	NC_005970
pECO29	*Escherichia coli*	3895	2	NC_001537
ClaDF13	*Escherichia coli*	9957	8	NC_002119
pBHRK18	*Escherichia coli*	5721	4	NC_005568
pBHRK19	*Escherichia coli*	5721	4	NC_005569
pFL129	*Escherichia coli*	6464	4	NC_005923
pAPEC-O2-ColV	*Escherichia coli*	184501	209	NC_007675
pCoo	*Escherichia coli*	98396	94	NC_007635
pB171	*Escherichia coli*	68817	80	NC_002142
pO113	*Escherichia coli*	165548	155	NC_007365
pLG13	*Escherichia coli*	6293	7	NC_005019
pIGAL1	*Escherichia coli*	8145	3	NC_005248
p1658/97	*Escherichia coli*	125491	141	NC_004998
pKL1	*Escherichia coli* KL4	1549	1	NC_002145
pO157	*Escherichia coli* O157:H7 str. Sakai	92077	184	NC_002128
pOSAK1	*Escherichia coli* O157:H7 str. Sakai	3306	3	NC_002127
pSFD10	*Salmonella choleraesuis*	4091	6	NC_003079
pOU1113	*Salmonella enterica*	80156	89	NC_007208
pC	*Salmonella enterica* serovar Enteritidis	5269	4	NC_003457
pBERT	*Salmonella enterica* subsp.*enterica* serovar Berta	4656	9	NC_001848
pKDSC50	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	49503	48	NC_002638
cryptic_plasmid	*Salmonella enterica* subsp.*enterica* serovar Choleraesuis	6066	7	NC_005862

Table 1: Plasmids analyzed *(Continued)*

pSCV50	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	49558	51	NC_006855
pSC138	*Salmonella enterica* subsp.*enterica* serovar Choleraesuis	138742	170	NC_006856
pHCM1	*Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18	218160	235	NC_003384
pHCM2	*Salmonella enterica* subsp.*enterica* serovar Typhi str. CT18	106516	105	NC_003385
pP	*Salmonella enteritidis*	4301	3	NC_003455
pK	*Salmonella enteritidis*	4245	3	NC_003456
pB	*Salmonella enteritidis*	1983	1	NC_005002
R27	*Salmonella typhi*	180461	207	NC_002305
pSC101	*Salmonella typhimurium*	9263	6	NC_002056
R64	*Salmonella typhimurium*	120826	135	NC_005014
pU302S	*Salmonella typhimurium*	3208	4	NC_006815
pU302L	*Salmonella typhimurium*	84514	103	NC_006816
pSLT	*Salmonella typhimurium*	93939	112	NC_003277
pSB4_227	*Shigella boydii* Sb227	126697	148	NC_007608
pSD1_197	*Shigella dysenteriae* Sd197	182726	223	NC_007607
pWR501	*Shigella flexneri*	221851	293	NC_002698
pCP301	*Shigella flexneri* 2a str. 301	221618	261	NC_004851
ColJs	*Shigella sonnei*	5210	3	NC_002809
pSS_046	*Shigella sonnei* Ss046	214396	238	NC_007385

The table lists the plasmids used in this work, their accession numbers, their host organism, their length, and the number of proteins their genes code for.

Results and discussion

Visual representation of sequence homology network

B2N was used to study the relationships existing between homologous proteins from all the completely sequenced plasmids available from three γ-Proteobacterial genera: *Escherichia, Shigella, Salmonella*. The dataset contains a total of 3701 ORFs, from 47 different plasmids (Table 1). To our knowledge, no attempt was made to describe in a meta analysis the overall body of plasmid sequence data in these species. Figure 2 shows the graphical representation of two networks generated with B2N using protein sequences in our dataset and using an aminoacid sequence identity threshold of 90% or 100% (Figure 2a and 2b respectively, where the thresholds are particularly high and the number of plasmids reduced to 39 out of 47 for clarity purposes). Proteins from the same plasmid are circularly arranged around the same centre and share the same color; proteins from the same genus are represented by the same shape (Figure 2c). The networks, obtained choosing an E-value threshold of 0.0001 and a minimum alignment length of 70 residues, have been visualized using the software Visone. The size of the nodes is proportional to the number of links they have. The analysis of Figure 2 revealed that most plasmids are strongly connected to others, but there are also plasmids exhibiting just few connections (see the section Phylogenetic profiling).

Focusing on protein clusters instead of plasmids, we can arrange nodes in an uniform visualization, where nodes are clustered together if they directly or indirectly share at least one link (Figure 3, with a threshold of 40% identity).

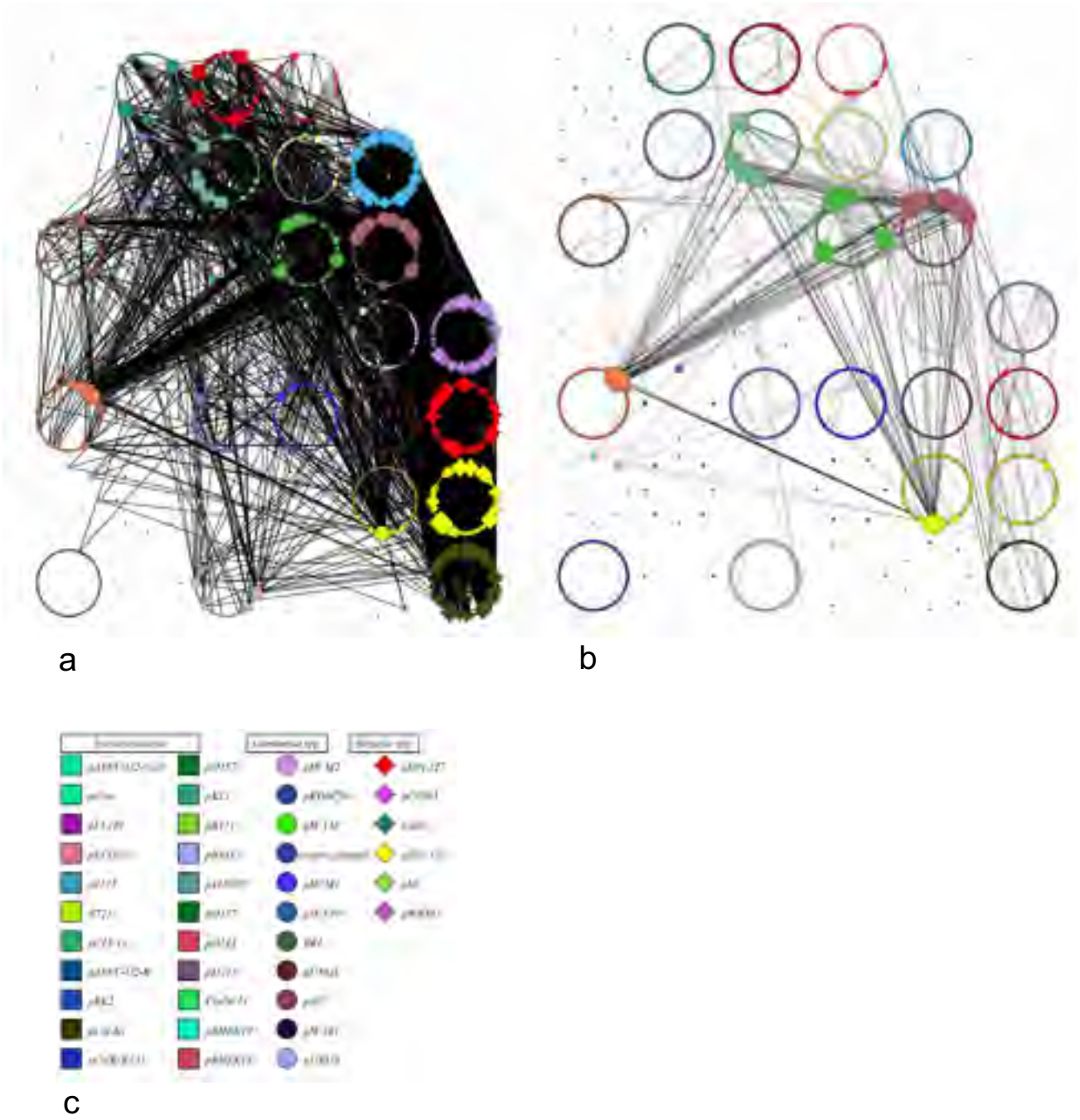


Figure 2
Plasmid homology networks. The output of B2N launched on the proteins encoded by 39 plasmids of three enterobacteria. Each protein in the dataset (see Table 1) is arranged circularly with proteins from the same source plasmid; proteins from the same plasmid are shown the same colour. Links connecting different nodes represent alignments found by BLAST (length > 70 amino acids and E-value<0.0001); consequently they describe the relationships existing between plasmids with a 90% (a) or 100% (b) identity cut-off; c) graphical legend. Symbols: squares, circles, and diamonds represent *E. coli*, *Salmonella* and *Shigella* plasmid proteins, respectively.

Quite interestingly, clustering of similar sequences at lower thresholds permits to assign a putative function to unknown or hypothetical proteins, and to discover the presence (if any) of functional classes or metabolic pathways that are more common in the network.

One of the problems faced with such complex data is the reduction of the dimensionality, so that important relationships can be more easily identified. Similarities in gene content between different plasmids can be better visualized by collapsing all the proteins belonging to the same plasmid in a single node. In this way a hypergraph is obtained where each node represents a single plasmid. The connection can be obtained from the plasmid vs plasmid Jaccard distance matrix or better, they can reflect the p-values matrix, so that each link in the hypergraph quantifies the significance of a given association between plasmids (showed in Additional file 2) and a simple hard thresholding allows changing the stringency for the inclusion of edges in the hypergraph.

Network data analysis

The analysis of the network data represented in Figures 2 and 3 revealed several interesting features of the relationships among the sequenced plasmids of the three genera under investigation:

1) Out of a total of 3701 proteins in the dataset, we found 1633 (44%) and 2471 (66.7%) isolated nodes at a threshold of 90% or 100% of identity for links, respectively (Figure 2a and 2b).

2) Most plasmids contain at least some gene coding for highly interconnected proteins; however, some of them (e.g. pRK2, ColJs CjI, pLG13, CloDF13) exhibited only few connections. Hence, these plasmids share few genes with the other members of the dataset at these threshold levels. This, in turn, may suggest that they might have experienced less recombination events than others.

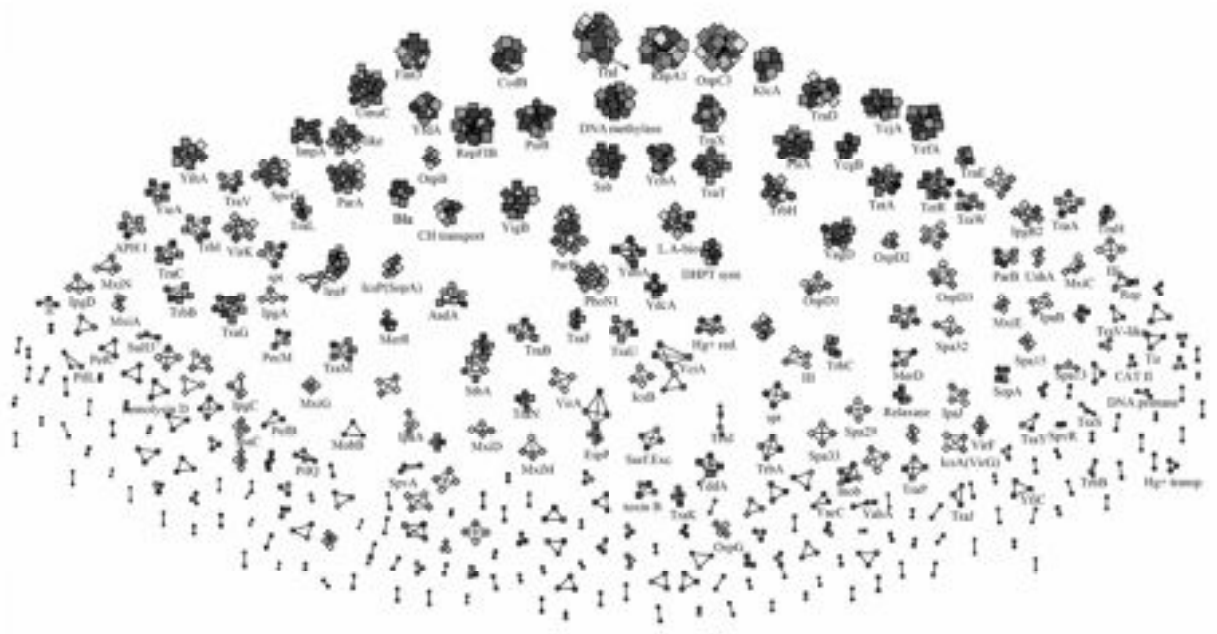


Figure 3

Uniform visualization of protein clustering. Uniform visualization of the similarity network for all of the 3701 proteins, displayed using a threshold identity for links of 40% (a degree of amino acid sequence identity sufficiently high to cluster together proteins that should perform the same function, and also allowing a better defined separation of all the main protein clusters [29,30]). Groups of homologous proteins are separated, allowing the identification of proteins that very likely share an identical/similar function. The labels for some groups of proteins discussed in the text or very common are shown: KlcA, antirestriction protein involved in the broad-host range of IncP plasmids; FinO, RNA chaperone related to repression of sex pilus formation; CcdB, protein involved in plasmid stability by killing bacteria that lose the plasmid during cell division; TetA and TetR, proteins responsible for resistance to tetracycline; Bla, β-lactamases; AadA, and DHPT synthase, proteins involved in resistance to aminoglycosides or sulfonamides, respectively. Tra and Trb, proteins requested for sex pilus formation; Mxi, Spa, Ipa, Ipg and Osp, proteins that are part of the type III secretion system.

3) Several proteins (about 40% of all the connected nodes) were found to be mobile elements (transposases, IS and transposons -related sequences), representing the most highly connected proteins in the network.

4) As shown in Figure 3, proteins shared by *Escherichia*, *Salmonella* and *Shigella* plasmids included: a) the antirestriction protein KlcA involved in the broad-host range of IncP plasmids [20]; b) the RNA chaperone FinO, related to repression of sex pilus formation [21,22]; c) the CcdB protein, which is involved in plasmid stability by killing bacteria that lose the plasmid [23].

5) Several clusters were composed by proteins shared by *Shigella* spp. and *Escherichia coli*; this finding is in agreement with the notion that they are considered to belong to the same species [24]. Moreover, several proteins were shared only by *E. coli* and *Salmonella* plasmids, including: the genetic determinants for antibiotic resistance such as TetA and TetR [25], β-lactamases (Bla) [25,27], genes for resistance to amino glycosides (AadA) and sulphonamides (DHPT synthase). A similar scenario was observed for sex pilus related proteins, such as Tra and Trb proteins: out of 22 different Tra groups, 21 contain proteins coming from *E. coli* and *Salmonella*, but 3 groups only (TraDI for DNA transport and TraX for pilin acetylation) have *Shigella* sequences. Likewise, out of 5 different Trb groups, we observed *Shigella* plasmid sequences in a single cluster (TrbH). Moreover, the proteins TraP, TrbA and TrbJ seem to be only present in plasmids from *E. coli*, while all the other sex pilus related proteins are shared with *Salmonella*. These data are in agreement with evidences for recent transfer of plasmid genes between enteroinvasive *Escherichia* and *Salmonella* [26,27].

Concerning the pathogenesis-related genes, *Shigella* plasmids seem to have a specific set of these genes, comprising at least some of the proteins of the type III secretion system (TTS), e.g.: Mxi, Spa, Ipa, Ipg and Osp proteins.

Finally, on the overall observation it appeared that besides the closer phylogenetic relationships existing between *E. coli* and *Shigella*, plasmid content appeared more similar among *E. coli* and *Salmonella* for what is concerned with antibiotic resistance and sex pilus formation.

Phylogenetic Profiling

Data discussed in the previous paragraphs, that is which proteins join a given cluster, were stored by B2N into a text file, which represent the phylogenetic profile of the dataset used; this can be further used by the program to calculate two matrices storing the distances between profiles in the two dimensions (i.e. for plasmids and for proteins), as described in Methods. The corresponding neighbor-joining dendrograms, that describe the similarity in gene content of the plasmids and protein co-occurrence patterns are shown in Figure 4, Figure 5 and Additional file 4. Data reported in Figure 4 revealed that most of plasmids does not form tight clusters coherent with the taxonomic status of their respective host species (*E. coli*, *Salmonella* or *Shigella*). This finding suggests a complex evolutionary history of such plasmid replicons with massive horizontal transfer and gene rearrangements. In particular, plasmid pSFD10 from *Salmonella* grouped with two *E. coli* plasmids (pRK2 and pLG13).

A relevant exception is represented by five Shigella plasmids (pCP301, pSB4 227, pSD1 197, pWR501, and pSS 046) that form a unique clade (which, however, also includes pC plasmid from Salmonella enterica).

Figure 5 and Additional file 3 report the co-occurrence clustering for the protein dataset of the selected plasmids. In general, plasmids are believed to share very few common functions (mainly related to their replication and mobility), several accessory genes and a complex history of recombination events among either them or the host chromosome(s) [28]. Here, we actually show that most of the co-occurrence clusters are due to protein related to plasmid transfer (e.g. Trb and Tra proteins). Nevertheless, several clusters are present showing the co-occurrence of hypothetical proteins with proteins with predicted functions such as type II secretion proteins and pilins (BfpK), or with proteins involved in mobilization (MobA, MbkC) and virulence factors (IroN). These analyses may help in addressing experimental analyses for elucidating the functional role of these proteins.

Conclusion

In conclusion, we report that the tools implemented by B2N allow to describe and to visualize in a new way some of the evolutionary features of plasmid molecules of Enterobacteriaceae; the most important results obtained by B2N on the Enterobacteriaceae dataset are related to the possibility, by means of phylogenetic profiling and network relationships of proteins, to uncover some of the molecular history, which shaped the evolution of this group of plasmids. In particular, data obtained suggested a large amount of horizontal transfer and rearrangement of plasmid molecules between *E. coli*, *Salmonella* and *Shigella*. Moreover, interestingly some plasmids from *Shigella* share a common history with *Salmonella* and several hypothetical proteins form co-occurrence clusters, suggesting possible roles in plasmid maintenance and/or pathogenesis, which could be investigated by conventional genetic techniques.

The proposed method is general enough to be proposed as a new tool for comparative genomic analyses of bacteria and can work at least within the range of phylogenetic

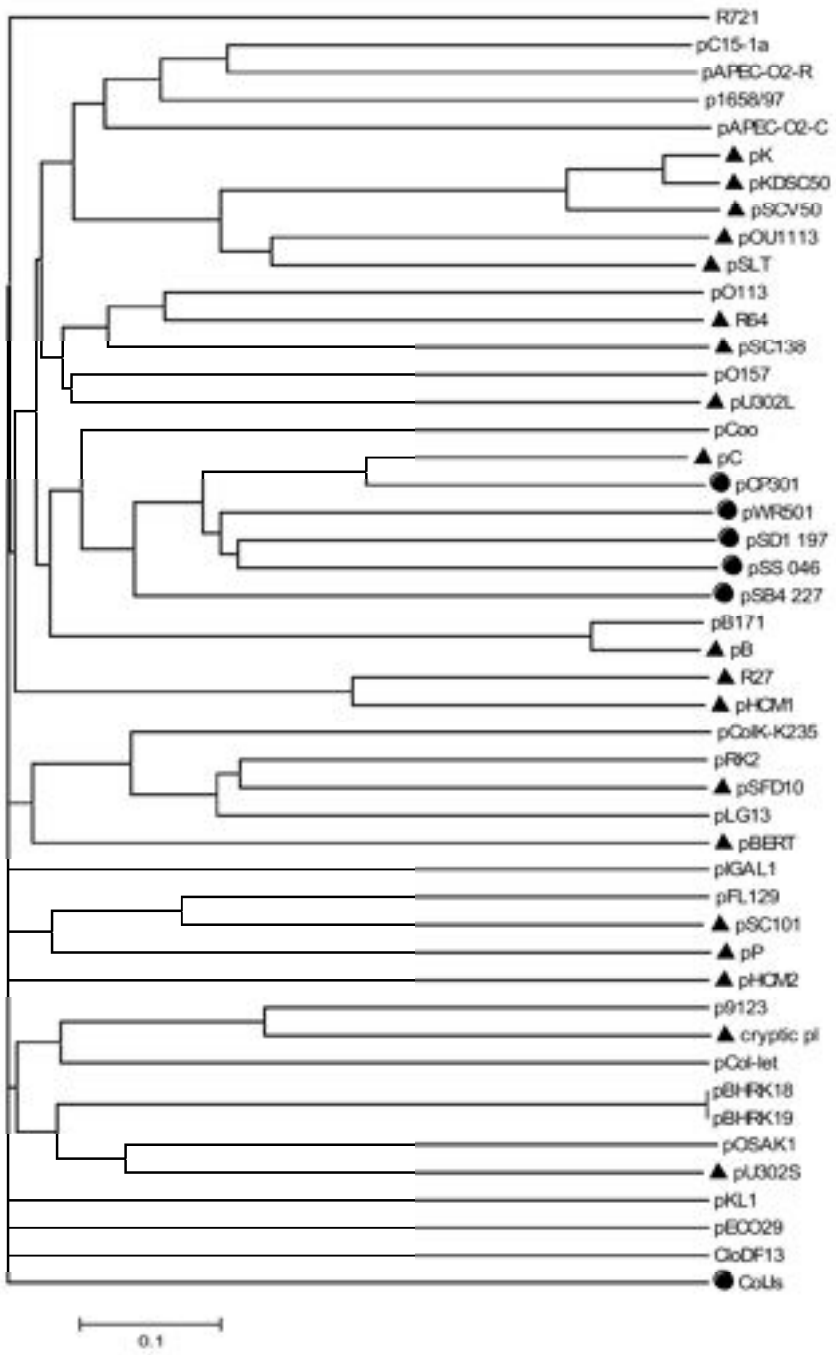


Figure 4
Neighbor-Joining dendrogram of the plasmids from phylogenetic profiling. Clustering of similarities in gene content of the plasmids obtained from their phylogenetic profile is reported (see text for details). Black circles or triangles before plasmid name refer to *Shigella* spp. or *Salmonella* spp. plasmids, respectively; *Escherichia coli* plasmids are not labeled.

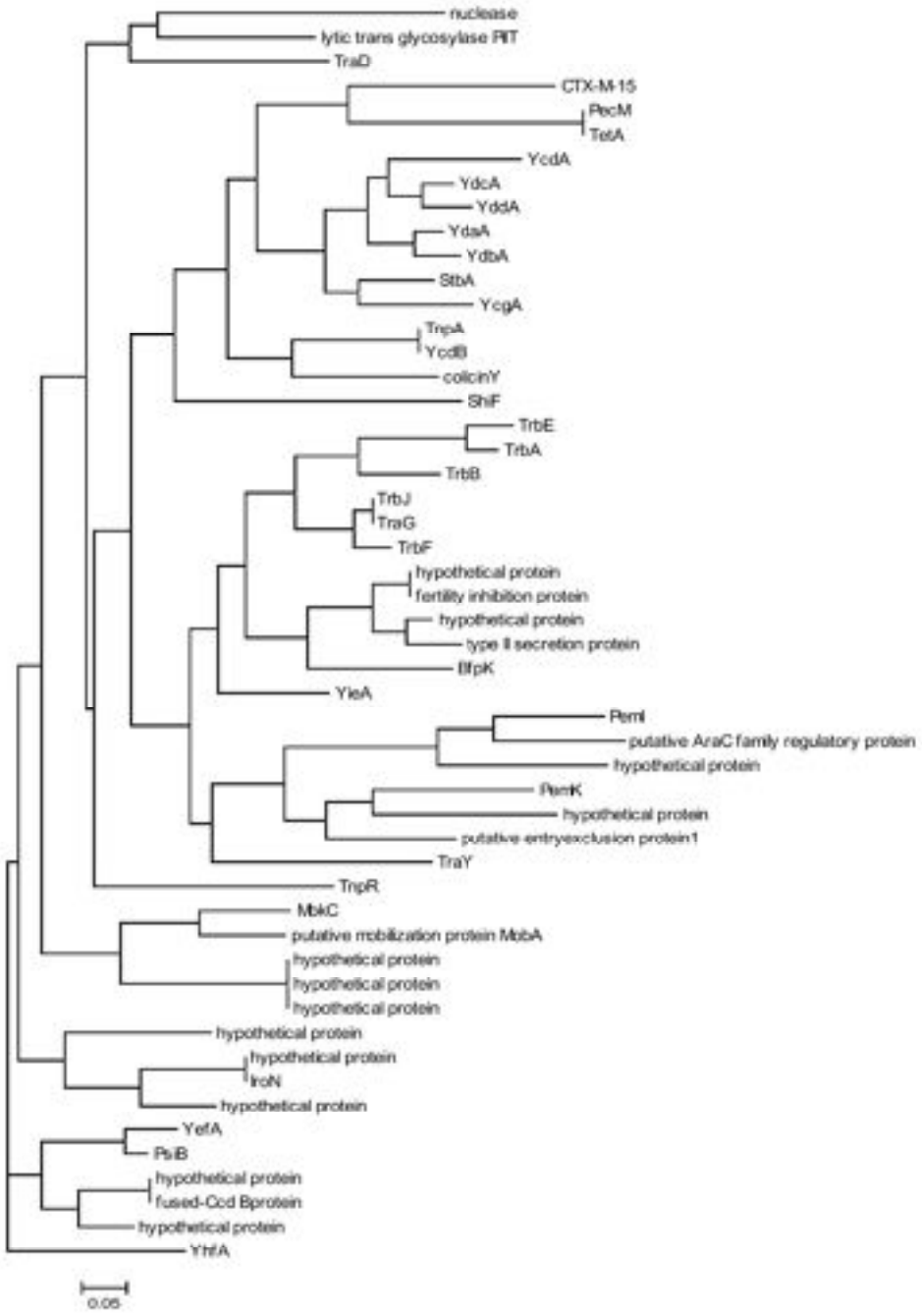


Figure 5
Neighbor-Joining dendrograms of protein co-occurrence pattern from phylogenetic profiling. Each cluster of this dendrogram includes those proteins that are commonly found together in the plasmids of the dataset reported in Table 1. Each hypothetical protein is associated with the GI of one representative of its corresponding protein cluster.

distances enabling Blast to find homologs. For this reason, the B2N approach could help solving some questions linked to the presence of (few) well conserved functions within plasmid datasets from wide taxonomic ranges (e.g. functions related to transfer or replication). Moreover, possible applications of the method could include also chromosomal replicons, trying to depict histories of gene rearrangement and integration from plasmid to chromosomes and *viceversa*.

Abbreviations

B2N: Blast2Network; TTS: type III secretion system.

Authors' contributions

MBr participated in conceiving the idea, wrote the program and performed part of the analyses. AM, PL and RF participated in conceiving the idea. MF performed part of the analyses. MBa participated in discussing results. All authors contributed to draft the paper. All authors read and approved the final manuscript.

Additional material

Additional file 1

Figure 1S – The functional activity of proteins from plasmid molecules present in GenBank database (as on March 2008). *Histogram showing the putative roles of all the proteins (73909) encoded by the plasmids present in the NCBI repository. Each of the 73909 proteins was probed against the COG database http://www.ncbi.nlm.gov/COG and its function was inferred according to the one assigned to the first BLAST hit of COG database. Data show that about 45% of the plasmid proteins deposited in the NCBI plasmids database have only a "general function" assignment or do not have any functional assigment at all.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-551-S1.pdf]

Additional file 2

Figure 2S – Hypergraph of plasmid sequences. *Similarity network showing the relationships existing among plasmids listed in Table 1, shown at two distinct thresholds. Differently from Figure 2, now each node represents a single plasmid and links the overall protein content shared among entire plasmids. In details, the size of nodes is proportional to the number of links possessed by a given plasmid, whereas the thickness of links was computed using p-values of the Jaccard distance calculated in phylogenetic profiling analysis (see text), hence accounting for an overall estimation of the shared proteins by each plasmids in respect to the others.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-551-S2.pdf]

Additional file 3

Software availability, requirements and user manual. *Project name: Blast2Network. Project home page: http://www.unifi.it/dblesson/CMpro_e-p-34.html. Operating system(s): Platform independent. Programming language: Java. Licence: GNU GPL. Any restrictions to use by non-academics: no restriction.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-551-S3.rar]

Additional file 4

Figure 3S – Phylogenetic profile with GI numbers of represented proteins as in Figure 5. *Protein co-occurrence patterns (see text for details) including the GI numbers of those proteins taken as representatives of each single cluster of Figure 3.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-551-S4.pdf]

Acknowledgements

This work was supported by the Italian Ministry of Research, FISR founding "Soil sink". MBr was supported by a post doc fellowships of the University of Firenze.

References

1. Gilmour MW, Thomson NR, Sanders M, Parkhill J, Taylor DE: **The complete nucleotide sequence of the resistance plasmid R478: defining the backbone components of incompatibility group H conjugative plasmids through comparative genomics.** *Plasmid* 2004, **52:**182-202.
2. Guerrero G, Peralta H, Aguilar A, Diaz R, Villalobos MA, Medrano-Soto A, Mora J: **Evolutionary, structural and functional relationships revealed by comparative analysis of syntenic genes in *Rhizobiales*.** *BMC Evol Biol* 2005, **5:**55.
3. Johnson TJ, Siek KE, Johnson SJ, Nolan LK: **DNA sequence and comparative genomics of pAPEC-O2-R, an avian pathogenic *Escherichia coli* transmissible R plasmid.** *Antimicrob Agents Chemother* 2005, **49:**4681-4688.
4. Tauch A, Puhler A, Kalinowski J, Thierbach G: **Plasmids in *Corynebacterium glutamicum* and their molecular classification by comparative genomics.** *J Biotechnol* 2003, **104:**27-40.
5. Kohiyama M, Hiraga S, Matic I, Radman M: **Bacterial sex: playing voyeurs 50 years later.** *Science* 2003, **301:**802-803.
6. Thomas CM, Nielsen KM: **Mechanisms of, and barriers to, horizontal gene transfer between bacteria.** *Nat Rev Microbiol* 2005, **3:**711-721.
7. Burrus V, Waldor MK: **Shaping bacterial genomes with integrative and conjugative elements.** *Res Microbiol* 2004, **155:**376-386.
8. Espinosa-Urgel M: **Plant-associated *Pseudomonas* populations : molecular biology, DNA dynamics, and gene transfer.** *Plasmid* 2004, **52:**139-150.
9. Dennis JJ: **The evolution of IncP catabolic plasmids.** *Curr Opin Biotechnol* 2005, **16:**291-298.
10. Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature.** *Trends Genet* 1995, **11:**283-290.
11. Fernández-López R, Garcillán-Barcia MP, Revilla C, Lázaro M, Vielva L, de la Cruz F: **Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution.** *FEMS Microbiol Rev* 2006, **30:**942-966.
12. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.
13. Linton A, Hinton MH: **Enterobacteriaceae associated with animals in health and disease.** *Soc Appl Bacteriol Symp Ser* 1988, **17:**715-855.
14. Slater FR, Bailey MJ, Tett AJ, Turner SL: **Progress towards understanding the fate of plasmids in bacterial communities.** *FEMS Microbiol Ecol* 2008, **66:**3-13.
15. Lavigne JP, Blanc-Potard AB: **Molecular evolution of *Salmonella enterica* serovar Typhimurium and pathogenic *Escherichia coli*: from pathogenesis to therapeutics.** *Infect Genet Evol* 2008, **8:**217-226.
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25:**3389-3402.
17. Sharp PM, Li WH: **The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281-1295.

18. Ramazzotti M, Brilli M, Fani R, Manao G, Degl'Innocenti D: **The CAI Analyser Package: inferring gene expressivity from raw genomic data.** *In Silico Biol* 2007, **7**:507-526.
19. van Passel MW, Bart A, Luyf AC, van Kampen AH, Ende A van der: **Compositional discordance between prokaryotic plasmids and host chromosomes.** *BMC Genomics* 2006, **7**:26.
20. Larsen MH, Figurski DH: **Structure, expression, and regulation of the kilC operon of promiscuous IncP alpha plasmids.** *J Bacteriol* 1994, **176**:5022-5032.
21. Dionisio F, Matic I, Radman M, Rodrigues OR, Taddei F: **Plasmids spread very fast in heterogeneous bacterial communities.** *Genetics* 2002, **162**:1525-1532.
22. Arthur DC, Ghetu AF, Gubbins MJ, Edwards RA, Frost LS, Glover JN: **FinO is an RNA chaperone that facilitates sense-antisense RNA interactions.** *EMBO J* 2003, **22**:6346-6355.
23. Aguirre-Ramirez M, Ramirez-Santos J, Van Melderen L, Gomez-Eichelmann MC: **Expression of the F plasmid ccd toxin-antitoxin system in *Escherichia coli* cells under nutritional stress.** *Can J Microbiol* 2006, **52**:24-30.
24. Escobar-Paramo P, Giudicelli C, Parsot C, Denamur E: **The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised.** *J Mol Evol* 2003, **57**:140-148.
25. Hartman AB, Essiet II, Isenbarger DW, Lindler LE: **Epidemiology of tetracycline resistance determinants in *Shigella* spp. and enteroinvasive *Escherichia coli*: characterization and dissemination of tet(A)-1.** *J Clin Microbiol* 2003, **41**:1023-1032.
26. Boyd EF, Hartl DL: **Recent horizontal transmission of plasmids between natural populations of *Escherichia coli* and *Salmonella enterica*.** *J Bacteriol* 1997, **179**:1622-1627.
27. Call DR, Kang MS, Daniels J, Besser TE: **Assessing genetic diversity in plasmids from *Escherichia coli* and *Salmonella enterica* using a mixed-plasmid microarray.** *J Appl Microbiol* 2006, **100**:15-28.
28. Thomas CM: **Paradigms of plasmid organization.** *Mol Microbiol* 2000, **37**:485-491.
29. Tian W, Skolnick J: **How well is enzyme function conserved as a function ofpairwise sequence identity?** *J Mol Biol* 2003, **333**:863-882.
30. Friedberg I: **Automated protein function prediction – the genomic challenge.** *Brief Bioinform* 2006, **7**:225-242.

Chapter 8

Exploring plasmids evolutionary dynamics: the *Acinetobacter* pan-plasmidome

8.1 Introduction

Plasmids are among the most important players in the evolution of prokaryotes and in their adaptation to fluctuating environmental conditions [Eberhard, 1990; Frost *et al.*, 2005; Slater *et al.*, 2008]. They are actually involved in many accessory functions and constitute, together with "not essential" chromosomal regions, what is referred to as the "dispensable genome" in the microbial pan-genome concept [Medini *et al.*, 2005]. Typically, a plasmid includes one or more essential genes encoding replicative functions. In addition, it may harbor one or more genes coding for a variable panoply of accessory metabolic processes and functions that are, in general, different from those encoded by chromosome(s) [Frost *et al.*, 2005; Khomenkov *et al.*, 2008; Tsuda *et al.*, 1999]. Actually, plasmid architecture is more flexible than the chromosomal one, concerning both gene content and gene organization, even within members of the same bacterial genus. Plasmids genes are in fact under differential selection while moving through the prokaryotic community [Eberhard, 1990], and consequently, they frequently gain and lose genes, revealing a very dynamic organization [Dutta & Pan, 2002; Frost *et al.*, 2005; Osborn & Boltner, 2002]. This flexibility is mostly due to the abundance of transposable elements they harbor and that facilitate intra- and intermolecular recombination by creating homology regions. Moreover, plasmids can be both vertically and horizontally inherited in a prokaryotic community, giving rise to the possibility that the very same plasmid molecule can be hosted in different genomic contexts, boosting the rearrangement of their functions and of gene organization [Bergstrom *et al.*, 2000; Davison, 1999; Zaneveld *et al.*, 2008]. Despite the key-role of plasmids in the prokaryotic world, the evolutionary dynamics of plasmids have been poorly explored, mainly because of the lack of extensive similarities between them, except for genes involved in replication and transfer functions [Cevallos *et al.*, 2008; Fernandez-Lopez *et al.*, 2006] which hampers classical phylogenetic analyses based on gene genealogy and synteny [Bentley & Parkhill, 2004]. However, a compu-

tational biology approach (Blast2Network) based on similarity networks reconstruction and phylogenetic profiling has been recently proposed and applied in a study-case to depict the similarities among plasmids from Enterobacteriaceae [Brilli *et al.*, 2008]. The bioinformatic package Blast2Network (hereafter designated B2N) provides an immediate visualization of the similarities, existing among aminoacidic or nucleic sequences [Brilli *et al.*, 2008]. This, in turn, opens the possibility to trace the evolutionary dynamics and history of entire plasmids and not only of single genes and/or operons harbored by them. In this context, bacteria belonging to the genus *Acinetobacter* may represent an excellent study-case, because strains of this genus are commonly found in soil, water and in association with animals [Juni, 1972; Peleg *et al.*, 2008]. Besides, some of them are well-known human pathogens, often responsible for opportunistic infections in hospitalized patients [Chen *et al.*, 2008; Dijkshoorn *et al.*, 2007; Peleg *et al.*, 2008]. A striking recent manifestation of *A. baumannii* is the occurrence in severely wounded soldiers coming back from Iraq [Davis *et al.*, 2005]. Currently, the genus *Acinetobacter* comprises 19 species with valid names and at least 13 putative species [Nemec *et al.*, 2009]. More than 975 strains have been recorded in the Taxonomy Browser of NCBI at July, 2 2009, but the precise taxonomy of these strains is not always clear since many have not been identified by unambiguous genotypic identification methods [Dijkshoorn *et al.*, 2008; Nemec *et al.*, 2009]. *Acinetobacter* strains are of special interest for the huge variety of environments they can colonize and the diverse metabolic abilities they display, as inferred from the occurrence of, e.g., hydrocarbon degrading strains in oil spills, human pathogens resistant to a plethora of antibiotics, rhizospheric bacteria and strains inhabiting bioreactors or insect guts [Bach & Gutnick, 2006; Hawkey & Munday, 2004; Iacono *et al.*, 2008; Khomenkov *et al.*, 2008; Marti *et al.*, 2008; Morales-Jimenez *et al.*, 2009; Mugnier *et al.*, 2008; Reams & Neidle, 2003]. Moreover, a special interest for members of this genus also relies on the ability of some strains, i.e. those belonging to the species *A. baylyi*, to undergo natural transformation [Davison, 1999; de Vries & Wackernagel, 2002]. This attribute has made the *A. baylyi* strain ADP1 (also named BD413) an exceptional tool for genetic analysis and engineering [Young *et al.*, 2005]. It has been reported that several *Acinetobacter* strains, especially those sharing particular ecological niches that require specific adaptations, like polluted environments and bioreactors, harbor plasmid molecules of different sizes undergoing frequent molecular rearrangements [Barberio & Fani, 1998; Decorosi *et al.*, 2006]. Particularly interesting among *Acinetobacter* plasmids is the pKLH2 family [Osborn *et al.*, 1997], a group of evolutionary related plasmids harboring mercury resistance genes (*mer*) embedded in a single compact operon that, in turn, has been suggested to represent an aberrant mercury resistance transposon (namely TndPKHLK2) that has lost genes responsible for transposition [Kholodii *et al.*, 2003]. Recently, some *Acinetobacter* genomes and plasmids have been completely sequenced. On March 31, 2009, the sequences of 7 genomes and 29 plasmids were available (Table 8.1). The *Acinetobacter* "pan-plasmidome", that is the complete set of plasmids harbored by members of this genus (comprising plasmids isolated from both pathogenic and environmental strains), is then particularly attractive to study its evolutionary dynamics because of the eclectic lifestyle of their host strains and the possible frequent genetic ex-

changes between its members. Therefore, in this work, a detailed comparative analysis of the completely sequenced *Acinetobacter* plasmids, available in public databases, was performed with the aim to i) reconstruct their evolutionary dynamics and ii) investigate the evolutionary cross-talk between them and the chromosomes of *Acinetobacter* strains.

Strains		Plasmids				Chromosomes		
Species and/or designation	Origin	n.	Name	Length (bp)	ORF(s)	n.	Length (bp)	ORF(s)
Acinetobacter baumannii	Clin.	1	pABIR	29823	26	n.d.		
Acinetobacter baumannii ATCC19606^{T}	Clin.	1	pMAC	9540	11	n.d.		
Acinetobacter baumannii	Clin.	1	pAB02	4162	6	n.d.		
Acinetobacter baumannii ACICU	Clin.	2	pACICU1 pACICU2	28279 64366	28 64	1	3904116	3667
Acinetobacter baumannii ATCC 17978	Clin.	2	pAB1 pAB2	13408 11302	11 5	1	3976747	3351
Acinetobacter baumannii AYE	Clin.	4	p1ABAYE p2ABAYE p3ABAYE p4ABAYE	5644 9661 94413 2726	7 11 82 5	1	3936291	3607
Acinetobacter baumannii SDF	Body lice	3	p1ABSDF p2ABSDF p3ABSDF	6106 25014 24922	8 30 24	1	3421954	2913
Acinetobacter baumannii AB0057	Clin.	1	pAB0057	8729	11	1	4050513	3790
Acinetobacter baumannii AB307-0294	Clin.	0				1	3760981	3451
Acinetobacter sp. EB104	Unknown	1	pAC450	4379	4	n.d.		
Acinetobacter sp. SUN	Clin.	1	pRAY	6076	10	n.d.		
Acinetobacter venetianus	Env.	2	pAV1 pAV2	10820 15135	11 16	n.d.		
Acinetobacter sp LUH5605	Env.	1	Ptet5605	3727	4	n.d.		
Acinetobacter sp BW3	Env.	1	pKLH207	9910	16	n.d.		
Acinetobacter calcoaceticus KHW14	Env.	1	pKLH201	11191	14	n.d.		
Acinetobacter calcoaceticus KHP18	Env.	1	pKLH2	6838	12	n.d.		
Acinetobacter sp ED23-35	Env.	1	pKLH208	9435	15	n.d.		
Acinetobacter sp ED45-25	Env.	1	pKLH205	8561	13	n.d.		
Acinetobacter junii	Env.	1	pKLH203	7195	12	n.d.		
Acinetobacter sp LS56-7	Env.	1	pKLH204	9489	15	n.d.		
Acinetobacter lwoffii	Env.	1	pKLH202	9471	17	n.d.		
Acinetobacter sp YAA**	Env.	1	pYA1	7407	5	n.d.		
Acinetobacter baylyi ADP1	Env.	0				1	3598621	3307
Tot.		29			493	7		24086

Table 8.1: List of completely sequenced *Acinetobacter* plasmids and chromosomes used in this work. Sequences were downloaded from the NCBI web-site (see Methods section, as on 31st March 2009). N.d. stands for not determined.

8.2 Methods

8.2.1 Sequence data source

The dataset used in this work is composed of all the proteins encoded by all the available completely sequenced *Acinetobacter* plasmids and chromosomes, downloaded from the NCBI ftp websites ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, (Table 8.1). On March 31 2009, 29 completely sequenced plasmids (whose lengths range between 2,726 and 94,413 bp) were available for a total of 493 amino acid sequences encoded. In addition, the genomic sequences of 7 *Acinetobacter* strains were also available, encoding for a total of 24,086 putative proteins.

8.2.2 Network construction and phylogenetic profiling

Similarity, identity based, networks were constructed using the tools implemented in the software B2N [Brilli *et al.*, 2008]. Networks, whereby the nodes represent the proteins and the links connecting them represent the shared identity values, were visualized and analyzed using the software Visone (http://visone.info/). Phylogenetic profiling dendrograms were constructed taking as input the matrix composed by all the plasmids under analysis (rows) and all the protein clusters (columns) identified [Brilli *et al.*, 2008]. Each position of the phylogenetic profile matrix will be "1" in the case a given plasmid (row) possesses (at least) one protein in the corresponding protein cluster (column), whereas it is filled with "0" in the opposite case. B2N calculates the Jaccard distance for both dimensions of the phylogenetic profiles matrix, which corresponds to the distance between plasmids in term of shared genes, and the distance between occurrence patterns of protein clusters in plasmids. The Jaccard distance matrices are then used for the construction of two neighbor-joining dendrograms. The first one describes similarities in gene content of the plasmids, the other one groups together those protein clusters with the most similar occurrence pattern within plasmids. Finally, random permutations of the original data allow to compute the statistical significance of the Jaccard distances.

8.2.3 Functional Assignment

The putative functional role of unassigned proteins was automatically retrieved according to the first best hit (FBH) in a similarity search (using Blast algorithm [Altschul *et al.*, 1997]) in the COG (http://www.ncbi.nlm.nih.gov/COG/) and in the PFAM database (http://pfam.sanger.ac.uk/). In both cases the standalone version of the databases was used, using default parameters.

8.3 Results

8.3.1 Plasmid networks

The first aim of the work was the identification and the analysis of the possible evolutionary relationships existing among the 29 *Acinetobacter* plasmids. To this purpose, all the 493 retrieved sequences of *Acinetobacter* plasmid-encoded proteins were used as input for the B2N software (see Methods section), generating a set of networks showing all the sequence identities existing among these proteins. In these networks nodes represent proteins, whereas links indicate the existence of sequence identity among them (Fig8.1 and Supplementary Material S1). The degree of sequence identity threshold is *a priori* selected. In principle, the higher the threshold used, the lower the number of links existing between proteins encoded by different plasmids. In addition, it can be assumed that the higher the degree of aminoacid identity between two proteins, the more recent would be the event (recombination/transposition/duplication/vertical transmission) responsible for the presence of the two orthologous/paralogous coding genes in different plasmids. We selected a minimum of 50% identity threshold since this degree of sequence identity

is sufficiently high to guarantee that in most cases the interconnected proteins perform the same function (i.e., they are encoded for by orthologous genes) [Gonzalez *et al.*, 1989; Tian & Skolnick, 2003]. The three networks shown in Figure 8.1 (at 100%, 90%, and 50% identity thresholds) and the other three reported in Supplementary Material S1 (at 60%, 70%, and 80% identity thresholds) were obtained by reiterating the analysis using different identity thresholds.

8.3.1.1 Analysis of links

The analysis of the networks reported in Figure 8.1 revealed that:

1. as expected, the number of links and interconnected nodes decreased with the increase of identity threshold (Table 8.2). At 50% of sequence identity, 213 out of

Identity threshold (%)	Number of	
	Nodes	Links
50	213	534
60	201	501
70	193	471
80	187	462
90	174	384
100	133	228

Table 8.2: Number of nodes and links at different identity threshold between 29 different *Acinetobacter*1 plasmids

   the 493 plasmid-encoded proteins were linked together. The other 280 proteins remained isolated because each of them did not share any link with the others and these were excluded from further analysis (see the Phylogenetic profiling section). The number of linked proteins decreased to 133 at a 100% sequence identity threshold. Still this number is unexpectedly high and suggests that the plasmids sharing at least one gene underwent recombination events very recently.

2. Three main groups can be recognized (Figure 8.1): Cluster 1 includes the eight plasmids pKLH2, pKLH201, pKLH202, pKLH203, pKLH204, pKLH205, pKLH207, and pKLH208 (hereinafter designated as the pKLH-family plasmids) that are highly interconnected although they have been isolated from different *Acinetobacter* spp. strains (Table 8.1). Cluster 2 is constituted by fifteen plasmids isolated from 8 *A. baumannii* strains, and Cluster 3 includes the remaining six plasmids isolated from *Acinetobacter* species different from *A. baumannii*.

3. c) Four plasmids (pRAY, pAC450, pYA1, and p4ABAYE) harbor genes encoding proteins that do not share any link neither between them nor with other proteins in the network.

4. d) Concerning the connections within each group (intra-links), the analysis of Figure 8.1 reveals that plasmids of cluster 1 (pKLH family) maintain a very high number of

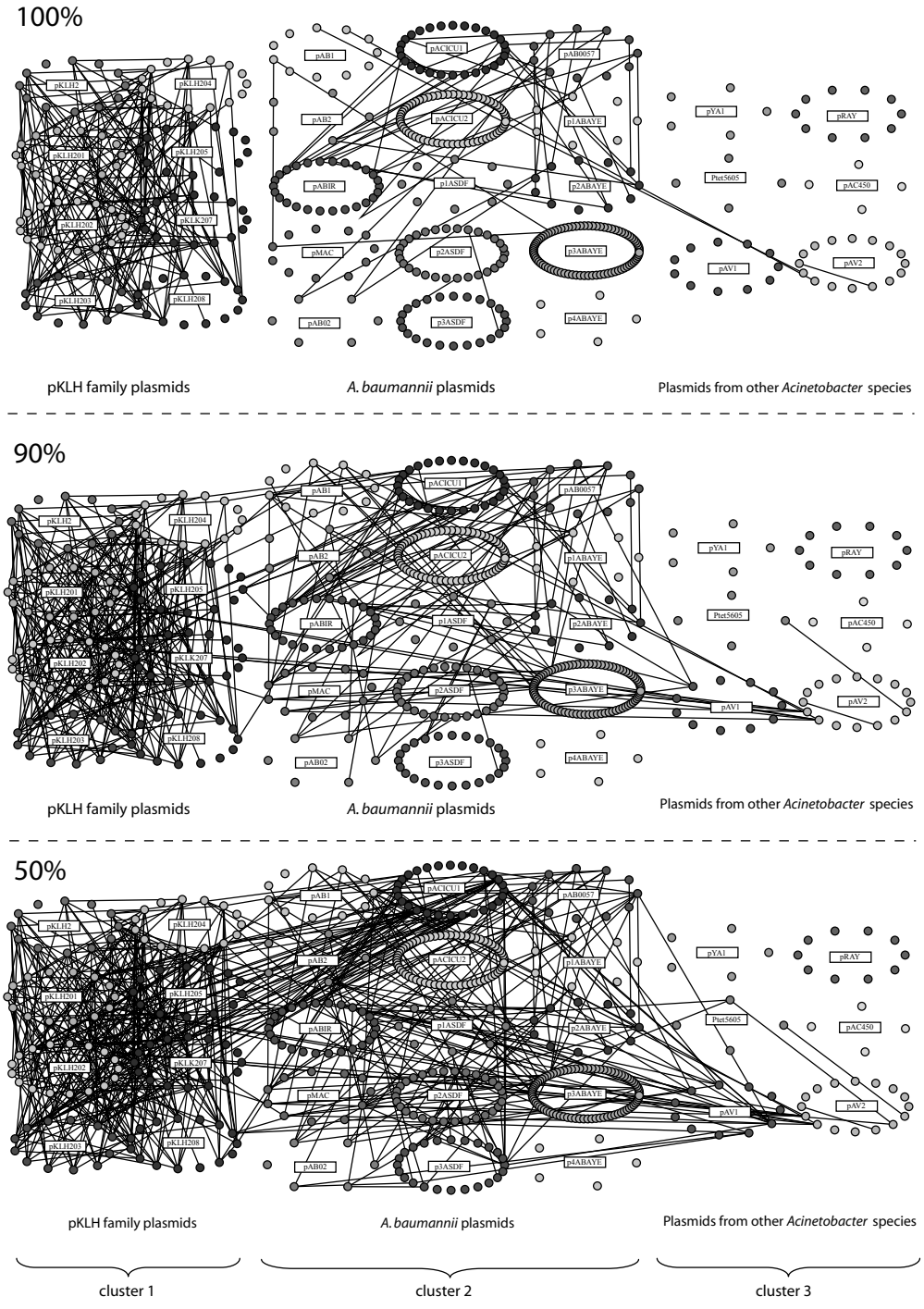


Figure 8.1: Identity based networks of the 493 *Acinetobacter* plasmid encoded proteins. All the proteins belonging to the same plasmid (nodes) are circularly arranged and are linked to the others according to their identity value. The resulting pictures for three different identity thresholds (100%, 90%, 50%) are shown.

links at the 100% threshold, whereas the number of intra-links of plasmids belonging to clusters 2 and 3 strongly decreases from the 50% to the 100% threshold. Overall, this finding suggests that pKLH plasmids share a more recent evolutionary pathway than that exhibited by the other plasmids. In addition to this, the finding that such pKLH plasmids have been isolated from different strains belonging to the same or to different *Acinetobacter* species suggests a high degree of horizontal flow of these plasmids (or at least of the shared genes). The biological significance of these data relies mainly on the fact that these plasmids harbor the genetic determinants for mercury resistance (*mer* genes, see Supplementary Material S2) that are positively favored in an environment under a strong selective pressure, i.e. in the presence of high mercury concentrations.

5. e) Regarding the connections between plasmids belonging to different clusters (inter-links), no link was observed between pKLH-family plasmids and those belonging to the other two clusters at the 100% threshold. However, at the 90% and 50% thresholds, many inter-links between cluster 1 and cluster 2 plasmids were observed. This interconnection was mainly due to plasmid pACICU1 and involves proteins predicted to be involved in DNA transposition, recombination and replication. Interestingly proteins assigned to OXA-58 oxacillinase and AraC binding protein were shared between pACICU1 and pABIR up to the 90% identity threshold. The connections existing between cluster 2 and some cluster 3 plasmids were in some cases retained also at higher thresholds (90-100%), for instance with proteins of pAV1 and pAV2 plasmids from *A. venetianus*. These links at high threshold between clusters suggest that the plasmid involved shared at least some common step in their evolutionary pathways.

6. f) In some cases, for instance plasmid pAV2, it is possible to recognize the traces of paralogous duplications within the same molecule.

7. g) The analysis of plasmids from the same strain revealed that, almost in all the cases, they did not share any link at the 100% threshold except for plasmids p2ABSDF and p3ABSDF (both from *A. baumannii* SDF), which had two links corresponding to two sequences assigned as 'hypothetical proteins'. The absence of links shared by these molecules may suggest the absence of recent genetic exchanges between plasmids in the same host.

8.3.1.2 Analysis of nodes

To analyze the functional classes of clustered/unclustered proteins, we made use of the uniform visualization output of the B2N software. The uniform visualization of the proteins involved in link formation obtained at 50% and 100% protein sequence identity is shown in Figure 8.2. At the 50% threshold the 213 proteins were clustered into 46 groups comprising at least 2 proteins per group. As might be expected, the number of clusters decreased to 32 at the 100% sequence identity threshold. Still, this number is surprisingly high and includes several cases (32) of genes shared at least by two plasmids of different

Acinetobacter species. On the basis of the above assumption, each cluster was numbered

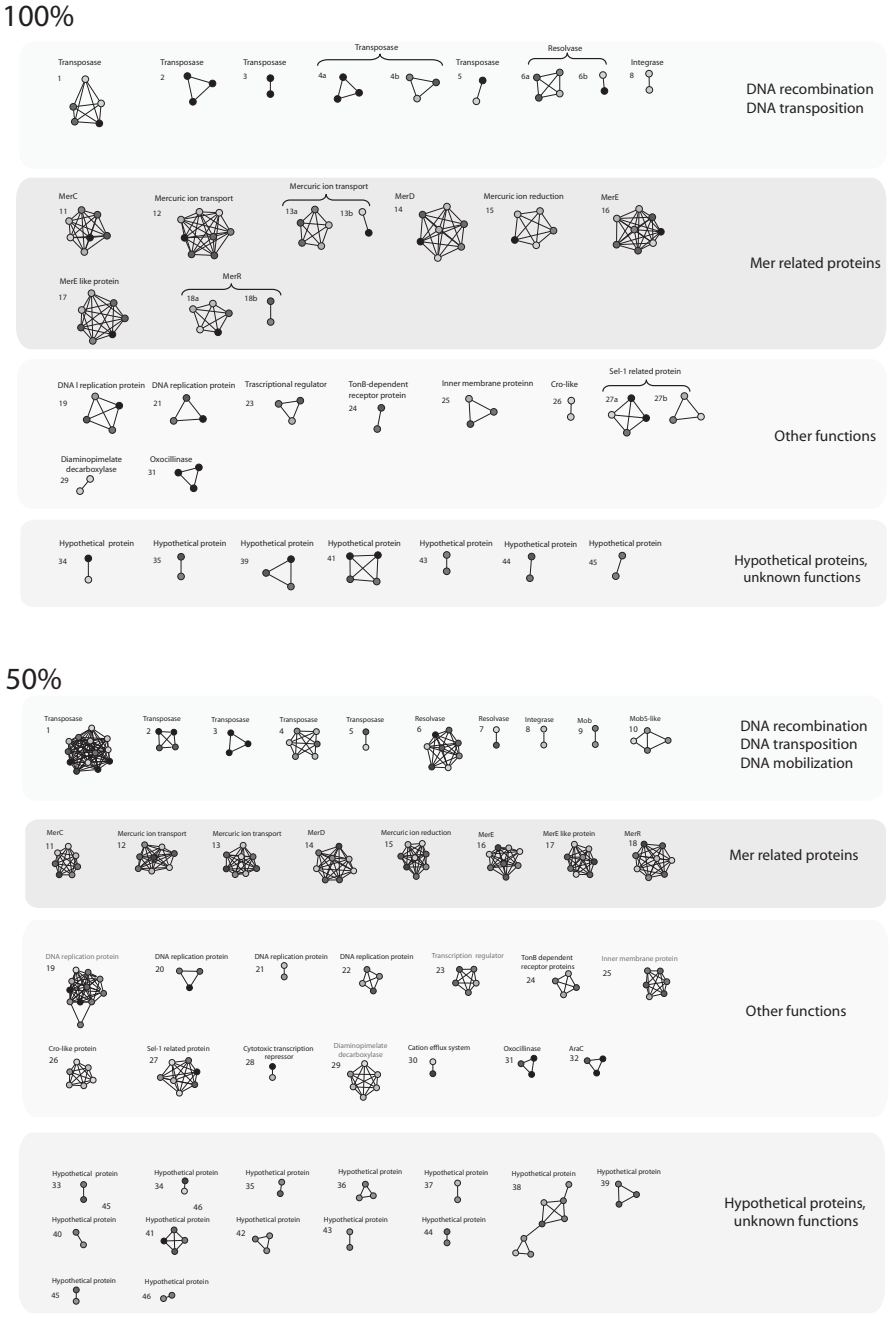


Figure 8.2: Uniform visualization of the networks shown in Figure 8.1. The different clusters embed proteins sharing 50% (below) and 100% (above) identity.

and named according to the functional assignation of the most represented proteins in the cluster (Table 8.3). The analysis of data reported in Table 8.3 revealed that:

1. a high number of protein clusters (1-8) were constituted by proteins involved in DNA transposition.

Protein cluster	Function	N° of nodes in the protein cluster	
		50%	100%
1	Transposase	13	9
2	Transposase	4	3
3	Transposase	3	2
4a	Transposase	6	3
4b			3
5	Transposase	2	2
6a	Resolvase	8	4
6b			2
7	Resolvase	2	0
8	Integrase	2	2
9	Mob	2	0
10	MobS-like	4	0
11	MerC	7	7
12	Mercuric ion transport	8	8
13a	Mercuric ion transport	8	5
13b			2
14	MerD	8	7
15	Mercuric ion reduction	8	5
16	MerE	8	8
17	MerE-like protein	8	7
18a	MerR	8	5
18b			2
19	DNA replication	11	4
20	DNA replication	3	0
21	DNA replication	2	3
22	DNA replication	4	0
23	Transcription regulator	5	3
24	TonB dependent receptor protein	4	2
25	Inner membrane protein	6	3
26	Cro-like protein	6	2
27a	Sel-1 related protein	7	4
27b			3
28	Cytotoxic transcription repressor	2	0
29	Diaminopimelate decarboxylase	7	2
30	Cation efflux system	2	0
31	Oxacillinase	3	3
32	AraC	3	0
33	Hypothetical protein	2	0
34	Hypothetical protein	2	2
35	Hypothetical protein	2	2
36	Hypothetical protein	3	0
37	Hypothetical protein	2	0
38	Hypothetical protein	8	0
39	Hypothetical protein	3	3
40	Hypothetical protein	2	0
41	Hypothetical protein	4	4
42	Hypothetical protein	3	0
43	Hypothetical protein	2	**2**
44	Hypothetical protein	2	2
45	Hypothetical protein	2	**2**

Table 8.3: Clusters of proteins exhibiting a link at 100% and/or 50% sequence identity.

2. Two protein clusters (9-10), comprising 2 and 4 nodes at the 50% threshold, respectively, included proteins encoded by *mob* genes, that is genes involved in plasmid transfer and/or mobilization. These two clusters disappeared at the 100% threshold.

3. Eight protein clusters (11-18), including a high number of plasmids (8), comprised proteins related to mercury resistance.

4. Fourteen protein clusters (19-32) included proteins whose function could not be assigned to a single cellular process.

5. Lastly, protein clusters numbered from 33 up to 46 (mainly composed by only two nodes) include only hypothetical proteins.

Overall, the number of nodes per cluster decreased from the 50% to the 100% threshold. However, clusters 11-18 (Mer-related proteins) maintained a high number of representatives (ranging from 5 to 8) also at the 100% threshold. This is responsible for the high number of connections existing between plasmids belonging to plasmid cluster 1 of Figure 8.1, which is (mainly) due to the sharing of genes involved in mercury resistance.

Concerning the 280 isolated proteins, a further analysis revealed that most of them (190) perform unknown functions and cannot be included in a functional category. Among the remaining 90 proteins, 46 of them are involved in a known information storage and processing and more precisely in translation and DNA replication, recombination and repair. The other functional categories are less frequent, except for ten sequences assigned to the COG database function "Energy production and conversion". The information stored in these networks was then used for the analyses presented in the next section.

8.3.2 Phylogenetic profiling

In order to try to depict the relationships existing between the *Acinetobacter* plasmids, phylogenetic profiles at the 100% and 50% identity thresholds were computed. Data obtained are reported in Figure 8.3 and Supplementary Material S3.

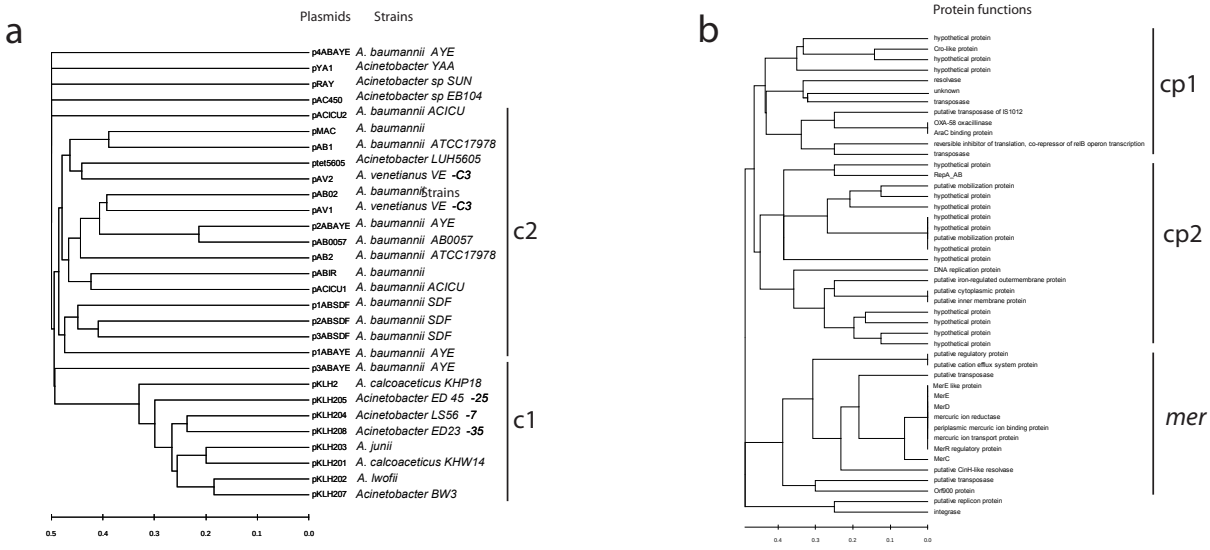


Figure 8.3: Neighbor joining dendrograms built using the Jaccard distance matrix values between phylogenetic profiles of the proteins in the dataset (see text for details) obtained with an identity threshold of 50% for plasmids (a) and protein clusters (b).

8.3.2.1 Analysis of plasmid dendrograms

The analysis of the phylogenetic profiles revealed that the branching order in the plasmid dendrogram at 50% identity is in partial agreement with the subdivisions reported in the similarity network shown in Figure 8.1. In details, (at least) two main clusters can be identified (c1 and c2 in Figure 8.3a), whereby the first one (c1) embeds all the pKLH-family plasmids. This clustering is in agreement with data presented in the previous

sections (Figure 8.1 and 8.2). The second cluster (c2) contains most of the plasmids belonging to *A. baumannii* strains, the two *A. venetianus* plasmids (pAV1 and pAV2) and the plasmid ptet5605 from *Acinetobacter* strain LUH5605. The remaining plasmids (those with no connection between them and with the other plasmids) and the plasmid pACICU2 and those present indifferent *Acinetobacter* strains, are not embedded in any of the two clusters. As expected, the dendrogram built when increasing the threshold up to 100% identity possessed both longer branches and a less defined clustering of plasmids (Supplementary Material S3). However, in agreement with the data shown in Figure 8.1 and Figure 8.2, the pKLH plasmids formed a coherent cluster clearly separated from the other plasmids. It is quite interesting that in most cases (5 out of 6) plasmids isolated from the same strain are more related to plasmids from other strains/species rather than to the other plasmids from the same strain. The only exception to this observation is represented by the three plasmids isolated from *A. baumannii* SDF. Indeed, all of them are embedded in the same coherent group (at 50% identity) and two of them are related also at the 100% threshold (Supplementary Material S3).

8.3.2.2 Analysis of protein dendrograms

The information stored in the adjacency matrix (see Material and Methods) and previously used to build plasmids similarity dendrograms, can also be used to identify those proteins that are commonly found together (co-occurrence) in the plasmids of the dataset. The obtained protein co-occurrence dendrograms (Figure 8.3b and Supplemental Material S3) identified three main groups at the 50% identity threshold (cp1, cp2, *mer*). In this dendrogram all the Mer-related proteins belong to the same cluster (mer cluster) comprising both proteins whose function is strictly related to mercury resistance/efflux process (their functional assignments include a cation efflux system, periplasmatic mercuric ion binding, a mercuric ion reductase and mercury ion transport) and proteins apparently not related to heavy-metal resistance and assigned with functions related either to regulation or DNA mobilization. These latter may co-occur with mercury resistance to provide accessory functions that are necessary for the transposition and the integration of the *mer* operon. In this context, it is worth noting the presence of a common core of mer genes, which is constituted by those seven proteins with distance equal to zero in the dendrogram at the 50% identity threshold (Figure 8.3b). The co/occurrence profile of these sequences suggests that their simultaneous presence within a plasmid might be essential for mercury resistance to occur. Indeed, this clade comprises proteins involved in mercuric ion binding, reduction and transport, although regulatory proteins are present (MerR). The main feature of cluster cp1 is the perfect co-occurrence of OXA-58 oxacillinase with a protein assigned as AraC binding protein. The former belongs to a well-known class of carbapenem-hydrolysing enzymes (OXA-type) [Walther-Rasmussen & Hoiby, 2006], conferring reduced susceptibility to carbapenem to the bacterial host cells, whereas the latter is a regulator of transcription that changes the way in which it binds DNA when the protein forms a complex with its monosaccharide ligand, L-arabinose [Soisson *et al.*, 1997]. Quite interestingly, the finding that the genes encoding these two proteins are found always together on the same plasmid molecule might reflect the way

the carbapenem resistance process is regulated and this, in turn, provide a good target for an experimental validation. Cluster cp2 (Figure 8.3b) mainly includes proteins whose function has not been characterized yet and, although proposing a set of good candidates for further experimental studies, is poorly informative for the purposes of this work.

8.3.3 Relationships between *Acinetobacter* plasmids and chromosomes

In order to check for the existence of genes shared between plasmids and chromosomes and to look for possible indications of past and/or recent rearrangements between them, we compared the 493 plasmid proteins at different identity thresholds with all the proteins of each of the available *Acinetobacter* genomes. It is reasonable to assume that the higher the degree of sequence identity shared by a chromosomal and a plasmid encoded protein, the greater the probability that the corresponding coding gene has been exchanged between them. For this reason, similarity (identity based) networks using each of the seven *Acinetobacter* completely sequenced genome available in NCBI GenBank (*A. baumannii* 17978, *A. baumannii* AYE, *A. baumannii* SDF, *A. baumannii* AB0057, *A. baumannii* AB307, *A. baumannii* ACICU and *A. baylyi* ADP1) and the 29 plasmids were constructed at different thresholds of identity. A preliminary analysis at 50% identity including all the 24,086 chromosomal encoded and the 493 plasmid encoded proteins was carried out, thereafter only those chromosomal proteins sharing at least one link with plasmid proteins were selected for comparisons at higher thresholds. In this way, we obtained a "mini-chromosome" for each of the seven *Acinetobacter* chromosomes comprising only those proteins sharing a link with (at least) one plasmid encoded protein. All the proteins (498) of these mini-chromosomes were then compared by B2N with the 493 plasmid encoded proteins. The identity networks obtained are shown in Figure 8.4 and Supplementary Material S4. Data obtained can be summarized as follows:

1. At the lowest threshold (50%) six plasmids (pYA1, pAC450, p4ABAYE, p1ABAYE, p1ABSDF, and pAV1) did not exhibit any link with any of the chromosomally encoded proteins. Three of them (pYA1, pAC450, p4ABAYE) did not share any link also with the other plasmids (see Figure 8.1), suggesting that these plasmids may have originated outside these *Acinetobacter* strains. The number of isolated plasmids raised to 18 at the 100% threshold. In total, at the 50% identity threshold, 359 isolated plasmid nodes were found. Most of them (245) did not retrieve any functional assignment when probing the COG database. Most of the 114 remaining sequences (74) were found to be involved in translation and DNA replication, recombination and repair processes.

2. The pKLH-family plasmids are strongly interconnected (74 links) with the *A. baumannii* AYE chromosome, a connection degree which is maintained also at higher thresholds (90% and 100%, 42 and 20 links, respectively). Such connection, even though at a lesser extent, was also disclosed with the *A. baumannii* B0057 chromosome. More in detail, at the 100% identity threshold, connections still exist between

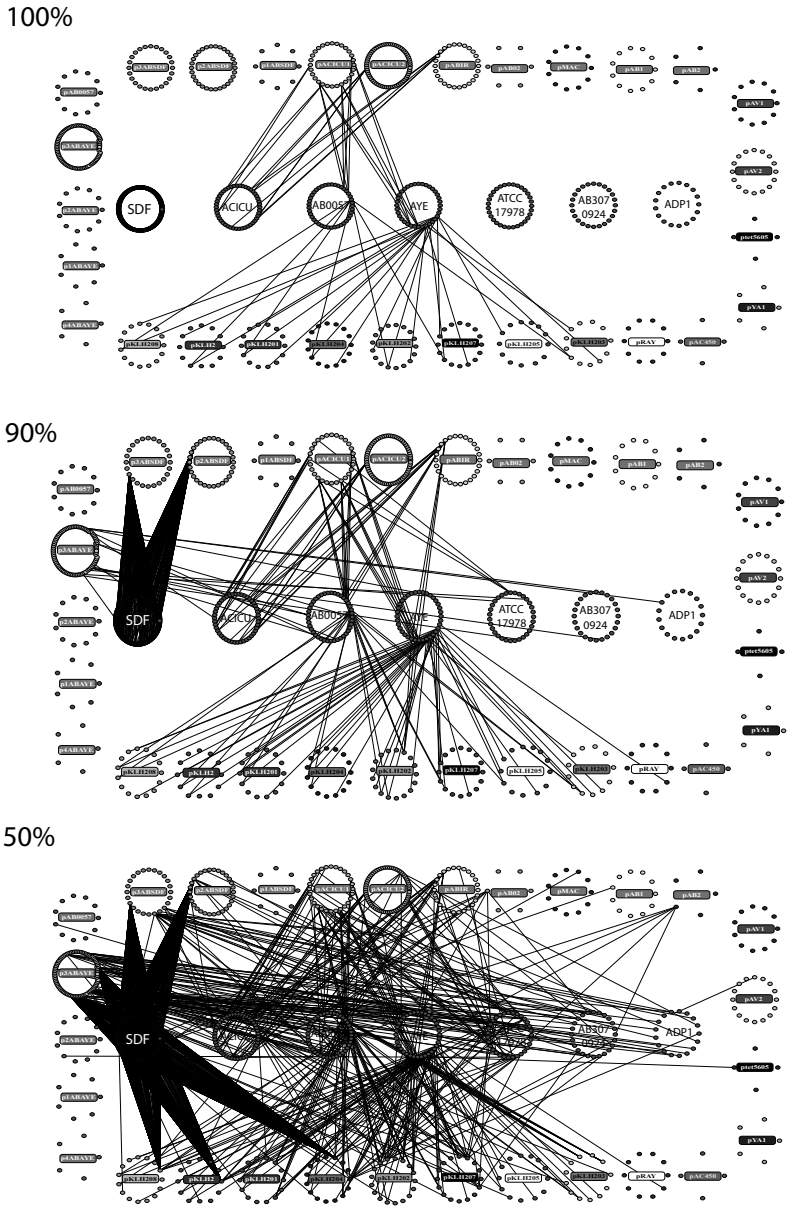


Figure 8.4: Identity relationships between the proteins of the *Acinetobacter* plasmid and mini-chromosome proteins. Mini-chromosomes (see text for the details of mini-chromosomes construction) are shown in the center and plasmids are circularly arranged. 100%, 90% and 50% identity threshold are shown. For clarity purposes, only the name of the corresponding strain is reported on minichromosomes.

plasmid encoded MerC, MerR and MerE and their counterparts on the *A. baumannii* AYE chromosome (in the case of *A. baumannii* B0057, only MerC is connected to its chromosomal counterpart). Lastly, two transposases from each of the two chromosomes are linked with plasmid encoded counterparts (on the pACICU1 plasmid). This finding strongly suggests a recent transfer of some genes between these,

either one of the two or both chromosomes and the pKLH plasmids.

3. Concerning the DNA molecules (plasmids and chromosomes) present in the same cytoplasm, it can be highlighted that three (p1ABAYE, p2ABAYE, and p4ABAYE) of the four plasmids isolated from *A. baumannii* AYE did not share any link with the corresponding chromosome, even at the lowest threshold (50%). Hence, the corresponding genes have not been exchanged between them and the host chromosome. Their origin is unclear since they do not share any link with the other six chromosomes either. However, it is worth noting that plasmid p2ABAYE has several links with other plasmids (see Figure 8.1) from different *A. baumannii* strains. The fourth and larger plasmid (p3ABAYE) showed only a limited connection degree with the corresponding chromosome. Concerning the two plasmids (pAB1 and pAB2) from strain *A. baumannii* 17978, pAB2 exhibited just one link with the corresponding chromosome, that disappeared at the 70% threshold (see Supplementary Material S4). A similar scenario can be depicted for the three plasmids isolated from *A. baumanni* SDF. No link, neither with its corresponding chromosome nor with the six other ones, was disclosed for plasmid p1ABSDF. Each of the other two plasmids were related to the corresponding chromosome via a single protein (assigned as transposases), which was connected to multiple (279) almost identical proteins located on the corresponding chromosome. The only exception is represented by plasmids isolated from *A. baumanni* ACICU, which appeared to be related to their corresponding chromosome as well as to other A. baumanni chromosomes.

4. At the chromosome level, at the 50% threshold the most interconnected one was that of *A. baumannii* SDF (296 shared proteins), while that with least proteins shared with plasmids was the *Acinetobacter baylyi* ADP1 chromosome (4 proteins) (Table 8.3). At the same identity threshold (Figure 8.4c), also the plasmids of the pKLH-family (particularly pKLH208, pKLH2, pKLH204) and p3ABAYE showed extensive links with *A. baumanni* SDF chromosome. In this case three plasmid proteins, (GI codes 14141701 from pKLH2, 30502913 from pKLH204 and 24411190 from pKLH208), all assigned as putative transposases, were responsible for most of the links. The functional assignment of proteins linked at the threshold of 50% identity between chromosomes and plasmids (Table 8.3) revealed that the large majority of the proteins shared are transposases or integrase. This is mostly due to as many as 279 out of 296 proteins in *A. baumannii* SDF and 9 out of 46 in *A. baumannii* ACICU. The other proteins, excluding 26 and 59 proteins with no functional assignment and no Pfam/COG hits respectively, were assigned to possible transcriptional regulators, membrane proteins and transporters, mercury resistance and detoxification. Interestingly, *A. baumannii* AYE and *A. baumannii* AB0057 share the highest amount of mer related proteins and integrases (4 and 2 respectively) among all the other *Acinetobacter* genomes.

Functional category	*A. baumannii ATCC17978*	*A. baumannii AYE*	*A. baumannii SDF*	*A. baumannii AB0057*	*A. baumannii AB307*	*A.baumannii ACICU*	*Acinetobacter sp. ADP1*	Total
Transposases	4	0	279	3	0	2	0	288
No Pfam hits	5	8	7	13	8	14	4	61
Uncharcterized/others	7	5	0	3	0	7	4	26
Cold-shock DNA-binding domain	2	4	1	3	4	2	3	19
Integrase	1	3	0	3	0	9	0	16
Zinc-binding dehydrogenase	1	2	1	2	2	2	2	12
Nucleoside recognition	3	1	1	1	1	1	1	9
ABC transporter	0	1	1	2	1	1	1	7
H-NS histone family	1	1	1	1	1	1	1	7
Penicillin binding protein transpeptidase domain	1	1	1	2	1	1	0	7
Mer-related protein	0	4	0	2	0	0	0	6
Haloacid dehalogenase-like hydrolase	1	1	1	1	1	1	0	6
Cation efflux family	1	1	1	1	1	1	0	6
regulatory proteins tetR-like	0	1	1	1	1	1	1	6
Catalase	1	1	0	1	1	1	0	5
Secretory lipase	1	1	0	1	1	1	0	5
AraC-like ligand binding domain	1	1	0	1	1	1	0	5
Resolvase	1	0	1	0	0	0	0	2
Proteins shared with plasmids at 50% identity	31	36	296	41	24	46	17	493

Table 8.4: Identity relationships between the proteins of the *Acinetobacter* plasmid and mini-chromosome proteins. Mini-chromosomes (see text for the details of mini-chromosomes construction) are shown in the center and plasmids are circularly arranged. 100%, 90% and 50% identity threshold are shown. For clarity purposes, only the name of the corresponding strain is reported on minichromosomes.

8.4 Discussion

The genus *Acinetobacter* comprises strains and species playing an important role in different ecological niches including soil and insects, whereas particular species have emerged as opportunistic pathogens. Several *Acinetobacter* strains recovered so far harbor plasmid molecules, some of which have been correlated to the peculiar adaptation to environmental conditions (e.g., pathogenicity, resistance to heavy-metals and antibiotics, biodegradation of hydrocarbons). In this work we have analyzed the entire set of 29 available plasmid sequences (the currently available/accessible pan-plasmidome), together with seven *Acinetobacter* fully sequenced genomes, to try to depict a possible evolutionary scenario of plasmids within a bacterial genus, and to describe the horizontal gene flow within this genus. Data obtained indicate that the 29 plasmids can be divided, by the extent of identity degree of the proteins they code for, into different groups. In particular we found that the group of pKLH-family plasmids and, to a lesser extent, those harbored by different strains of *A. baumannii*, are still interconnected at a high (90-100%) amino acid sequence identity value. This finding suggests that they might be the outcome of an evolutionary molecular history starting from ancestral plasmid backbones, which very likely underwent several and different rearrangements during the flow in different hosts, capturing and/or losing genes from their genomes. The analysis of networks constructed, excluding from the analysis those sequences responsible for heavy metal resistance and located on pKLH2-family plasmids (data not reported), reveals that pKLH2-family plasmids share a number of links higher than that exhibited with the other plasmids. Besides, some of these proteins are linked at a 100% threshold. The interconnected proteins belonging to

these pKHL2-family plasmids are involved in processes, such as DNA synthesis and DNA translocation (*cinH* like), apparently not related to mercury resistance. This finding fits with the model proposed by Kholodii et al. [2003] to explain the evolution of pKLH2 plasmids, according to which plasmids harboring mer operons are relics of an ancient plasmid that has undergone several rounds of fusions with other plasmids, followed by deletions, stabilizing the resulting mercury resistance plasmids. Hence, the occurrence of several independent recombination events might have led to the evolutionary relatedness of pHLK2, involving also the flanking regions of the mer operon. Moreover, these results reinforce data presented in other comparative studies and stating that plasmids from different and often geographically separated taxa may still share similar "core" genes [Heuer *et al.*, 2004; Jerke *et al.*, 2008; Rawlings, 2005]. Moreover, the finding that plasmid pAV2 (from *A. venetianus* VE-C3), shares some interconnections (at 100% identity) with plasmid pAB1 from a different host species (*A. baumannii*), suggests that pAV2 might be the result of recombination events that occurred between its ancestor and (at least) the ancestor of pAB1. This may be the case for several other plasmids in this study. The possibility that different plasmids may have inhabited the same host cells is emphasized by the finding that only for a few of them Inc-like proteins, causing co-existence incompatibility between plasmids, have been found (p3ABAYE, pACICU2, data not shown). From the 50% identity threshold plasmid network, some additional information can be retrieved. In fact pAV1 from *A. venetianus* VE-C3 shows several links with *A. baumannii* plasmids, while plasmids of the pKLH family (isolated from different *Acinetobacter* species) are linked with *A. baumannii* plasmids. Furthermore, the finding that plasmids sharing the same genes have been isolated from different strain/species may suggest the existence of both an intra- and interspecific flow of these molecules through horizontal gene transfer mechanisms. These data may suggest a time-scale of events, from the older to the most recent, paralleled by the increasing identity thresholds. In other words, some of the recombination events should have occurred very recently since the shared proteins exhibit a very high degree of sequence identity (up to 100%), whereas others (involving the genes coding for proteins sharing a low degree of sequence identity, i.e. 50%) should be more ancient. Based on homology relationships, a total of 46 clusters were found among the proteins identified as connectors between different plasmids. At the 50% identity threshold, 8 clusters are composed by proteins involved in recombination, while the others mostly reflect the relationships between pKLH-family plasmids, being composed by genes of the mer operon involved in mercury resistance encoded by that plasmid family. The finding that some plasmids (or their ancestors) might have 'inhabited' different cells belonging to different *Acinetobacter* species raises the question of what mechanism was responsible (i.e. transduction, conjugation and/or transformation) for their transfer between different hosts. Because most of the plasmids analyzed are relatively small molecules and do not harbor *tra* and/or *mob* genes, it is plausible that they might have been transmitted through transformation and/or transduction, the latter by uptake in a bacteriophage. In fact some bacteriophages, like P22 of *Salmonella typhimurium* have been shown to transduce plasmids in addition to chromosomal markers [Mann & Slauch, 1997] other than transconjugation. Actually, the species *A. baylyi* with ADP1 (BD413) be-

ing the most widely studied strain has been shown to be naturally competent [Barbe *et al.*, 2004a; Iwaki & Arakawa, 2006; Johnsborg *et al.*, 2007; Pontiroli *et al.*, 2009; Vaneechoutte *et al.*, 2006; Watson & Carter, 2008]. For other species, this property is largely unknown although in the literature there are numerous unfounded assumptions that natural competence is a general feature of the genus. Despite the large number of links connecting most of the plasmids in our dataset, four of them, namely pAC450, pRAY, pYA1 and p4ABAYE, did not possess any of the proteins identified in the similarity network and consequently did not show any link. With the exception of pRAY, they did not share any link with the *Acinetobacter* chromosomes either. The differences in gene content exhibited by these plasmids suggests possible evolutionary pathways that did not cross those of the other *Acinetobacter* plasmids and chromosomes analyzed. To investigate more deeply the evolutionary scenario of our plasmid dataset, we analyzed the relationships between plasmid-borne proteins and the completely sequenced genomes available. In fact, although prokaryotic plasmids have played and are still playing a key role in metabolic and genome evolution little is known about the evolutionary relationships existing between them and the chromosome(s), including the molecular rearrangements they underwent during their flow throughout the microbial community world. Data obtained in this work show the existence of extensive links between all *Acinetobacter* chromosomes and most plasmids (at 50% and even at 90% identity threshold). The finding that several connections were maintained up to the 90% identity threshold implies that the degree of divergence between the plasmid and chromosomal encoded proteins was very limited, which in turn strongly suggests that the encoding genes were exchanged (relatively) recently, independently from the possible recombination events that may have occurred between plasmids sharing the same protein coding gene. This is particularly relevant for the *A. baumanni* AYE chromosome, which seems to be the major contributor of plasmid genes, since it shares at least one link with 12 out of the 29 plasmids analyzed at a 90% identity threshold and for the *A. baumanni* SDF chromosome, which shows several links with the corresponding plasmids p2ABSDF and p3ABSDF. Even though it cannot be a priori completely excluded, the possibility that some of the plasmids might have inherited some genes from other species of the genus *Acinetobacter* or even of other genera, the degree of sequence similarity is sufficiently high to suggest evolutionary recent exchanges between those chromosomes and the plasmids. It is also interesting to note that plasmids from the same host (as pAV1 and pAV2, pAB1 and pAB2, pMAC and pAB02) show links with different *Acinetobacter* chromosomes, suggesting independent evolutionary pathways not related to the particular host in which they have been isolated. Plasmids pYA1 and pAC450, i.e. two of the three that do not share any link with the other plasmids, did not share any link with any of the chromosomes at the identity threshold of 50%, suggesting that they may have acquired/exchanged these genes from/with other bacterial chromosomes. However, the lack of knowledge on the genome sequences of their respective current hosts hampers discussion about their co-evolution with their host's chromosomes. In fact, the presence of a large pangenome for the genus *Acinetobacter* with a core genome accounting for only 50-70% of the total genome [Vallenet *et al.*, 2008], strongly limits a full evolutionary reconstruction of plasmid life histories. This gap will probably be filled in the near future, when more

sequence data from other representatives of this genus will be released. In agreement with the presence of a large mobile gene pool, transposases are the most important functional category of shared proteins, especially for *A. baumannii* SDF. This result is in line with previous findings of comparative genomics [Vallenet *et al.*, 2008] that showed that *A. baumannii* SDF is riddled with numerous relics of mobile elements, including transposons, insertion sequences and prophage elements. As expected, *A. baumannii* AYE shares the highest amount of mer related proteins and integrases among all the other *Acinetobacter* genomes. Actually *A. baumannii* AYE possesses mercury resistance genes (*mer* operon) [Fournier *et al.*, 2006], whereas all the other Acinetobacter strains are not apparently capable to detoxify mercury [Barbe *et al.*, 2004b; Smith *et al.*, 2007; Vallenet *et al.*, 2008]. These data indicate that a HGT event might have been responsible for the appearance of resistance to heavy metal in *A. baumannii* AYE. According to this idea, this strain might have acquired the whole *mer* resistance operon and integrated in its chromosome after a recombination event (possibly with a pKLH2 plasmid). This finding is in agreement with data proposed by Osborn et al. [1997] who suggested that transposition events appear to have been extensively involved in the evolution of mer determinants in Gram-negative bacteria. It is to be noticed that an event of HGT in the opposite direction, i.e. from the *Acinetobacter baumannii* AYE chromosome to one or more plasmids, cannot be a priori excluded. However, since all the other chromosomes lack the mer operon, we reckon the first scenario as the most parsimonious and the most probable. However, the reconstruction of the complete evolutionary scenario of the mer genes will be possible only when the genome sequences of strains harboring pKLH plasmids will be available. Increasing the threshold from 90% to 100% identity resulted in the elimination of all the links between most plasmids and the chromosomes. Surprisingly, also the highly interconnected *A. baumannii* SDF chromosome lost all the links between its proteins (mostly transposases) and plasmids. This latter result might be accounted for by the hypothesis of the absence of particular structural/functional constraints acting on the transposases, leading to a (relatively) rapid diversification of proteins during evolution. On the contrary *A. baumannii* AYE, B0057 and ACICU maintained most of the links with the plasmids, revealing either strong functional constraints over the sequences of the shared proteins (mainly involved in heavy-metal resistance) or, alternatively, recent HGT events.

8.5 Conclusions

Data obtained in this work reveal that the absence of mobilization and transfer functions in most of the *Acinetobacter* plasmids seems not to pose particular barriers to horizontal gene transfer (HGT) since they have probably a long history of rearrangements with other plasmids and with chromosomes. Furthermore, a phylogenetic profiling pipeline was applied to the whole body of plasmids encoded sequences, revealing interesting co-occurrences that, in turn, may help to shed some light in the functioning mechanisms of proteins involved in antibiotic resistance and mercury detoxification. In fact, in our opinion, this analysis provides promising candidates for further experimental validations in the field of antibiotic resistance and bioremediation. Lastly, we have shown that, by

combining plasmid and chromosome similarity, identity based, network analysis, we have been able to describe an evolutionary pathway also for highly mobile genetic elements that lack extensively shared genes. In particular we found that transposases and selective pressure for mercury resistance seem to have played a pivotal role in plasmid evolution in *Acinetobacter* genomes sequenced so far.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.

Bach, H. & Gutnick, D.L. (2006). Novel polysaccharide-protein-based amphipathic formulations. *Applied Microbiology and Biotechnology*, **71**, 34–38.

Barbe, V., Vallenet, D., Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., Cruveiller, S., Robert, C., Duprat, S., Wincker, P., Ornston, L.N., Weissenbach, J., Marliere, P., Cohen, G.N. & Medigue, C. (2004a). Unique features revealed by the genome sequence of *Acinetobacter* sp adp1, a versatile and naturally transformation competent bacterium. *Nucleic Acids Research*, **32**, 5766–5779.

Barbe, V., Vallenet, D., Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., Cruveiller, S., Robert, C., Duprat, S., Wincker, P., Ornston, L.N., Weissenbach, J., Marliere, P., Cohen, G.N. & Medigue, C. (2004b). Unique features revealed by the genome sequence of *Acinetobacter* sp. adp1, a versatile and naturally transformation competent bacterium. *Nucleic Acids Res*, **32**, 5766–79.

Barberio, C. & Fani, R. (1998). Biodiversity of an *Acinetobacter* population isolated from activated sludge. *Research in Microbiology*, **149**, 665–673.

Bentley, S.D. & Parkhill, J. (2004). Comparative genomic structure of prokaryotes. *Annual Review of Genetics*, **38**, 771–792.

Bergstrom, C., Lipsitch, M. & Levin, B. (2000). Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics*, **155**, 1505–1519.

Brilli, M., Mengoni, A., Fondi, M., Bazzicalupo, M., Li, P. & R, F. (2008). Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using blast2network. *BMC Bioinformatics*.

Cevallos, M.A., Cervantes-Rivera, R. & Gutierrez-Rios, R.M. (2008). The repabc plasmid family. *Plasmid*, **60**, 19–37.

Chen, T.L., Siu, L.K., Lee, Y.T., Chen, C.P., Huang, L.Y., Wu, R.C.C., Cho, W.L. & Fung, C.P. (2008). aciinetobacter baylyi as a pathogen for opportunistic infection. *Journal of Clinical Microbiology*, **46**, 2938–2944.

Davis, K.A., Moran, K.A., McAllister, C.K. & Gray, P.J. (2005). Multidrug-resistant *Acinetobacter* extremity infections in soldiers. *Emerg Infect Dis*, **11**, 1218–24.

Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid*, **42**, 73–91.

de Vries, J. & Wackernagel, W. (2002). Integration of foreign dna during natural transformation of *Acinetobacter* sp by homology-facilitated illegitimate recombination. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2094–2099.

Decorosi, F., Mengoni, A., Baldi, F. & Fani, R. (2006). Identification of alkane monoxygenase genes in *Acinetobacter* venetianus ve-c3 and analysis of mutants impaired in diesel fuel degradation. *Annals of Microbiology*, **56**, 207–214.

Dijkshoorn, H., Nierkens, V. & Nicolaou, M. (2008). Risk groups for overweight and obesity among turkish and moroccan migrants in the netherlands. *Public Health*, **122**, 625–30.

Dijkshoorn, L., Nemec, A. & Seifert, H. (2007). An increasing threat in hospitals: multidrug-resistant *Acinetobacter baumannii*. *Nat Rev Microbiol*, **5**, 939–51.

Dutta, C. & Pan, A. (2002). Horizontal gene transfer and bacterial diversity. *Journal of Biosciences*, **27**, 27–33.

Eberhard, W. (1990). Evolution in bacterial plasmids and levels of selection. *The Quarterly review of biology*, **65**, 3–22.

Fernandez-Lopez, R., Garcillon-Barcia, M., Revilla, C., Lazaro, M., Vielva, L. & F., d.l.C. (2006). Dynamics of the incw genetic backbone imply general trends in conjugative plasmid evolution. *FEMS Microbiology Reviews*, **30**, 942–966.

Fournier, P.E., Vallenet, D., Barbe, V., Audic, S., Ogata, H., Poirel, L., Richet, H., Robert, C., Mangenot, S., Abergel, C., Nordmann, P., Weissenbach, J., Raoult, D. & Claverie, J.M. (2006). Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. *PLoS Genet*, **2**, e7.

Frost, L., Leplae, R., Summers, A. & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature reviews. Microbiology*, **3**, 722–732.

Gonzalez, F.A., Bonapace, E., Belzer, I., Friedberg, I. & Heppel, L.A. (1989). Two distinct receptors for atp can be distinguished in swiss 3t6 mouse fibroblasts by their desensitization. *Biochem Biophys Res Commun*, **164**, 706–13.

Hawkey, P.M. & Munday, C.J. (2004). Multiple resistance in gram-negative bacteria. *Reviews in Medical Microbiology*, **15**, 51–61.

Heuer, H., Szczepanowski, R., Schneiker, S., Puhler, A., Top, E.M. & Schluter, A. (2004). The complete sequences of plasmids pb2 and pb3 provide evidence for a recent ancestor of the incp-1beta group without any accessory genes. *Microbiology*, **150**, 3591–9.

Iacono, M., Villa, L., Fortini, D., Bordoni, R., Imperi, F., Bonnal, R.J.P., Sicheritz-Ponten, T., De Bellis, G., Visca, P., Cassone, A. & Carattoli, A. (2008). Whole-genome pyrosequencing of an epidemic multidrug-resistant *Acinetobacter baumannii* strain belonging to the european clone ii group. *Antimicrobial Agents and Chemotherapy*, **52**, 2616–2625.

Iwaki, M. & Arakawa, Y. (2006). Transformation of *Acinetobacter* sp bd413 with dna from commercially available genetically modified potato and papaya. *Letters in Applied Microbiology*, **43**, 215–221.

Jerke, K., Nakatsu, C.H., Beasley, F. & Konopka, A. (2008). Comparative analysis of eight arthrobacter plasmids. *Plasmid*, **59**, 73–85.

Johnsborg, O., Eldholm, V. & Havarstein, L.S. (2007). Natural genetic transformation: prevalence, mechanisms and function. *Research in Microbiology*, **158**, 767–778.

Juni, E. (1972). Interspecies transformation of acinetobacter: Genetic evidence for a ubiquitous genus. *J. Bacteriol.*, **112**, 917–931.

Kholodii, G., Mindlin, S., Gorlenko, Z., Petrova, M., Hobman, J. & Nikiforov, V. (2003). Translocation of transposition-deficient (tn(d)pklh2-like) transposons in the natural environment: mechanistic insights from the study of adjacent dna sequences. *Microbiology-Sgm*, **150**, 979–992.

Khomenkov, V.G., Shevelev, A.B., Zhukov, V.G., Zagustina, N.A., Bezborodov, A.M. & Popov, V.O. (2008). Organization of metabolic pathways and molecular-genetic mechanisms of xenobiotic degradation in microorganisms: A review. *Applied Biochemistry and Microbiology*, **44**, 117–135.

Mann, B.A. & Slauch, J.M. (1997). Transduction of low-copy number plasmids by bacteriophage p22. *Genetics*, **146**, 447–56.

Marti, S., Sanchez-Cespedes, J., Blasco, M.D., Ruiz, M., Espinal, P., Alba, V., Fernandez-Cuenca, F., Pascual, A. & Vila, J. (2008). Characterization of the carbapenem-hydrolyzing oxacillinase oxa-58 in an *Acinetobacter* genospecies 3 clinical isolate. *Antimicrobial Agents and Chemotherapy*, **52**, 2955–2958.

Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. (2005). The microbial pan-genome. *Current opinion in genetics and development*, **15**, 589–594.

Morales-Jimenez, J., Zuniga, G., Villa-Tanaca, L. & Hernandez-Rodriguez, C. (2009). Bacterial community and nitrogen fixation in the red turpentine beetle, dendroctonus valens leconte (coleoptera: Curculionidae: Scolytinae). *Microb Ecol*.

Mugnier, P., Poirel, L., Pitout, M. & Nordmann, P. (2008). Carbapenem-resistant and oxa-23-producing *Acinetobacter baumannii* isolates in the united arab emirates. *Clinical Microbiology and Infection*, **14**, 879–882.

Nemec, A., Musilek, M., Maixnerova, M., De Baere, T., van der Reijden, T.J., Vaneechoutte, M. & Dijkshoorn, L. (2009). aciinetobacter beijerinckii sp. nov. and *Acinetobacter* gyllenbergii sp. nov., haemolytic organisms isolated from humans. *Int J Syst Evol Microbiol*, **59**, 118–24.

Osborn, A.M. & Boltner, D. (2002). When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid*, **48**, 202–212.

Osborn, A.M., Bruce, K.D., Strike, P. & Ritchie, D.A. (1997). Distribution, diversity and evolution of the bacterial mercury resistance (mer) operon. *FEMS Microbiol Rev*, **19**, 239–62.

Peleg, A., Seifert, H. & Paterson, D. (2008). aciinetobacter baumannii: emergence of a successful pathogen. *Clinical microbiology reviews*, **21**, 538–582.

Pontiroli, A., Rizzi, A., Simonet, P., Daffonchio, D., Vogel, T.M. & Monier, J.M. (2009). Visual evidence of horizontal gene transfer between plants and bacteria in the phytosphere of transplastomic tobacco. *Applied and Environmental Microbiology*, **75**, 3314–3322.

Rawlings, D.E. (2005). The evolution of ptf-fc2 and ptc-f14, two related plasmids of the incq-family. *Plasmid*, **53**, 137–47.

Reams, A.B. & Neidle, E.L. (2003). Genome plasticity in acinetobacter: new degradative capabilities acquired by the spontaneous amplification of large chromosomal segments. *Molecular Microbiology*, **47**, 1291–1304.

Slater, F.R., Bailey, M.J., Tett, A.J. & Turner, S.L. (2008). Progress towards understanding the fate of plasmids in bacterial communities. *Fems Microbiology Ecology*, **66**, 3–13.

Smith, M.G., Gianoulis, T.A., Pukatzki, S., Mekalanos, J.J., Ornston, L.N., Gerstein, M. & Snyder, M. (2007). New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes Dev*, **21**, 601–14.

Soisson, S.M., MacDougall-Shackleton, B., Schleif, R. & Wolberger, C. (1997). Structural basis for ligand-regulated oligomerization of arac. *Science*, **276**, 421–5.

Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol*, **333**, 863–82.

Tsuda, M., Tan, H.M., Nishi, A. & Furukawa, K. (1999). Mobile catabolic genes in bacteria. *Journal of Bioscience and Bioengineering*, **87**, 401–410.

Vallenet, D., Nordmann, P., Barbe, V., Poirel, L., Mangenot, S., Bataille, E., Dossat, C., Gas, S., Kreimeyer, A., Lenoble, P., Oztas, S., Poulain, J., Segurens, B., Robert, C., Abergel, C., Claverie, J.M., Raoult, D., Medigue, C., Weissenbach, J. & Cruveiller, S. (2008). Comparative analysis of *Acinetobacters*: three genomes for three lifestyles. *PLoS ONE*, **3**, e1805.

Vaneechoutte, M., Young, D.M., Ornston, L.N., De Baere, T., Nemec, A., Van Der Reijden, T., Carr, E., Tjernberg, I. & Dijkshoorn, L. (2006). Naturally transformable *Acinetobacter* sp strain adp1 belongs to the newly described species *Acinetobacter* baylyi. *Applied and Environmental Microbiology*, **72**, 932–936.

Walther-Rasmussen, J. & Hoiby, N. (2006). Oxa-type carbapenemases. *J Antimicrob Chemother*, **57**, 373–83.

Watson, S.K. & Carter, P.E. (2008). Environmental influences on *Acinetobacter* sp strain bd413 transformation in soil. *Biology and Fertility of Soils*, **45**, 83–92.

Young, D.M., Parke, D. & Ornston, L.N. (2005). Opportunities for genetic investigation afforded by *Acinetobacter* baylyi, a nutritionally versatile bacterial species that is highly competent for natural transformation. *Annu Rev Microbiol*, **59**, 519–51.

Zaneveld, J.R., Nemergut, D.R. & Knight, R. (2008). Are all horizontal gene transfers created equal? prospects for mechanism-based studies of hgt patterns. *Microbiology-Sgm*, **154**, 1–15.

Chapter 9

The horizontal flow of plasmid encoded resistome: clues from inter-generic similarity networks analysis

9.1 Introduction

Bacterial antibiotic resistance is nowadays a major clinical issue all over the world, revealing the power and the success of microbial evolution and adaptation [Bennett, 2008]. Although resistance has been a continuing problem since antibiotics were introduced, it is the raising of the number, diversity and range of resistant organisms that has become a huge clinical problem [Tenover, 2006]. Indeed, continuous administration of antibiotics can lead to the increase of the pathogens resistance to antimicrobial compounds [D'Costa *et al.*, 2006; Martinez, 2008; Wright, 2007]. Moreover, in recent decades there has been a dearth of new classes of discovered antibiotics [Charles & Grayson, 2004]. An integrated network of antibiotic resistance mechanisms can ensure to bacteria the protection against a plethora of chemical compounds. The different strategies adopted include (i) the presence of an enzyme inactivating the antimicrobial agent, (ii) a mutation in a gene involved in the synthesis of the target of the antimicrobial agent that reduces its binding capacity, (iii) the post-transcriptional and post-translational modification of the target of the antimicrobial agent, which reduces its binding capacity, (iv) the reduced uptake of the antimicrobial agent and, finally, (v) the active efflux of the antimicrobial agent [Fluit *et al.*, 2001]. The genetic determinants responsible for all these different antibiotic resistance mechanisms mainly reside on plasmid molecules. In fact, plasmid-encoded antibiotic resistance encompasses most, if not all, classes of antibiotics currently in clinical use and comprises resistance to many of them that are in the forefront of antibiotic therapy [Bennett, 2008]. Usually plasmids do not accommodate any of the "core" genes required by the cell for basic growth and division, but rather carry genes that may be useful periodically to enable the cell to exploit particular environmental conditions, such as the survival in the presence of a potentially lethal antibiotic [Bennett, 2008]. Furthermore,

plasmids are extremely important in microbial evolution, because they can be transferred between micro-organisms, representing natural vectors for the transfer of genes and functions they code for [Brilli *et al.*, 2008]. In this context, the plasmid-mediated transfer of genes conferring resistance to antibiotics (and/or to heavy metals), and of pathogenicity genes represents the most important effects of "bacterial sex" from both an evolutionary and ecological viewpoint [Kohiyama *et al.*, 2003]. A formidable sexual promiscuity has given bacteria a unique advantage over other organisms because it provides an awesome mechanism for ongoing adaptation and reticulate evolution-a sort of permanently and rapidly evolving communal genome [Kohiyama *et al.*, 2003]. It is clear that the gene pool, which consists of genes present on a plethora of diverse mobile genetic elements, may at the end result in the lateral transfer of genes also among phylogenetically unrelated bacteria and this is the process whereby bacteria may become multi-resistant to antibiotics [Martinez *et al.*, 2009]. The risk that antibiotic resistance poses for human health has meant that research in this area has focused, up to now, primarily on their role within clinical settings [Martinez, 2008]. On the other hand, recent studies identified an unexpected concentration of environmental antibiotic resistance in natural microbial communities [D'Costa *et al.*, 2006], leading to the supposition that natural environments may represent important reservoirs of antibiotic resistance genes. As a consequence, scientists have started exploring the presence of bacterial encoded antibiotic resistance in different ecological niches. Studies on the antibiotic resistance potential of soil microorganisms have revealed remarkable frequency of resistance to antibiotics that for decades have served as gold-standard treatments, as well as those only recently approved for human use [D'Costa *et al.*, 2006]. Moreover, it has been noticed that an important part of the dispersal and evolution of antibiotic-resistant bacteria depends on water environments [Baquero *et al.*, 2008] where bacteria inhabiting different environments can get in touch and, eventually, undergo HGT. All these lines of evidence suggest that antibiotic resistance genes in human pathogens can originate from a multitude of bacterial sources (i.e. coming from distinct environments) and that the whole microbial resistome can be considered as a single gene-pool in which bacteria can find the right "weapon-shield" that is required for their survival. In particular, some authors [Baquero *et al.*, 2008] identified four main genetic reactors where genetic exchange and recombination shape the future evolution of resistance determinants. Accordingly, the primary reactor is constituted by the human and animal microbiota, the secondary one involves the hospitals, long-term care facilities, farms, or any other place in which susceptible individuals are crowded and exposed to bacterial exchange The tertiary reactor corresponds to the wastewater and any type of biological residues originated in the secondary reactor. Lastly, the fourth reactor is represented by soil and the surface or ground water environments, where the bacterial organisms originated in the previous reactors mix and counteract with environmental organisms. Classically, antibiotic resistance evolution and spreading across microbial communities have been investigated mainly through experimental approaches [Cattoir *et al.*, 2008; D'Costa *et al.*, 2006; Dugan *et al.*, 2004; Koike *et al.*, 2007; Mackie *et al.*, 2006; Sobecky, 2002]. Nevertheless, the use of plasmids pyro-sequencing as a routine laboratory technique [Schluter *et al.*, 2008], the assembly of databases embedding detailed

information on antibiotic resistance genetic determinants [Liu & Pop, 2009; Scaria *et al.*, 2005], together with the development of bioinformatics tools enabling the visualization of sequence homology relationships through similarity networks [Brilli *et al.*, 2008] can pave the way to large scale comparative analyses adopting computational biology strategies. Therefore, the aim of this work was to integrate all these sources of information and explore the distribution of plasmids encoded antibiotic resistance within the microbial community, building similarity networks of antibiotic resistance determinants representative of nearly 1000 plasmids. Moreover, in order to describe the horizontal flow of antibiotic resistance coding genes across the microbial community and the barriers posed by taxonomical and geographical separation, we focused the attention on shared antibiotic resistance determinants that were likely transferred *via* HGT combining these data with the habitat assignment of microorganisms. Lastly, since HGT between two taxa implies a shared habitat at the time of transfer, or alternatively the presence of a vector of transmission [Hooper *et al.*, 2009], these connections provide a means for evaluating the movement of microorganisms inhabiting different ecological niches and their role in the spreading (vectors) of antibiotic resistance within the prokaryotic kingdom.

9.2 Methods

9.2.1 Dataset assembly

Bacterial plasmids sequences were retrieved from the NCBI FTP database (ftp://ftp.ncbi.nih.gov/refseq/release/plasmid/). Data concerning the phenotypes (lifestyle, habitat, etc,) and the taxonomic affiliation of each microorganism were retrieved from the GOLD database [Liolios *et al.*, 2008] at http://www.genomesonline.org/DBs/goldtable.txt. On may 2009, 122482 plasmid encoded aminoacid sequences were available; each of these sequences was used as seed in a BLAST [Altschul *et al.*, 1997] search against the ARDB (Figure 9.1) database [Liu & Pop, 2009] using the following parameters: e-value $e^{-}10$, minimum alignment length 50 amino acid (aa), that is a degree of amino acid sequence identity sufficiently high to retrieve all the proteins that should perform a function related to antibiotic resistance [Friedberg, 2006]. In this way, a set of 5030 sequences putatively associated to antibiotic resistance process was retrieved (See Supplemental Material S1 for the complete list of accession codes of the proteins used in this work). These sequences belonged to 956 different plasmids and were representative of 364 organisms corresponding to 134 different bacterial genera.

9.2.2 Network construction and links normalization

The overall adopted strategy is schematically reported in Figure 9.1. Briefly all the (5030) sequences that, on the basis of sequence similarity resulted hypothetically related to antibiotic resistance, were used in an all vs. all BLAST comparison [Altschul *et al.*, 1997], using default parameters. The obtained BLAST output was post-processed (using *in home*

developed Perl codes) and visualized [using the software Visone (http://visone.info/)] as a similarity network in which nodes represent proteins and links the identity values they share. Since we aimed at identifying those sequences that most likely underwent HGT events, we combined sequence identity relationships (the values of links in the similarity network) with that coming from likely vertically inherited molecular markers, that is the 16S rDNA coding genes of the microorganisms possessing them. In other words, a given link value between two proteins in the network was normalized by computing it as the product between its value (the identity percentage shared by the two sequences) and the distance between the 16S rRNA genes representative of the genus of the organisms possessing them (see below). Hereinafter we will refer to this normalized value as Normalized Identity Value (NIV). Thus, links connecting very similar sequences from closer microorganisms tended to have lower NIV than similar sequences retrieved in more

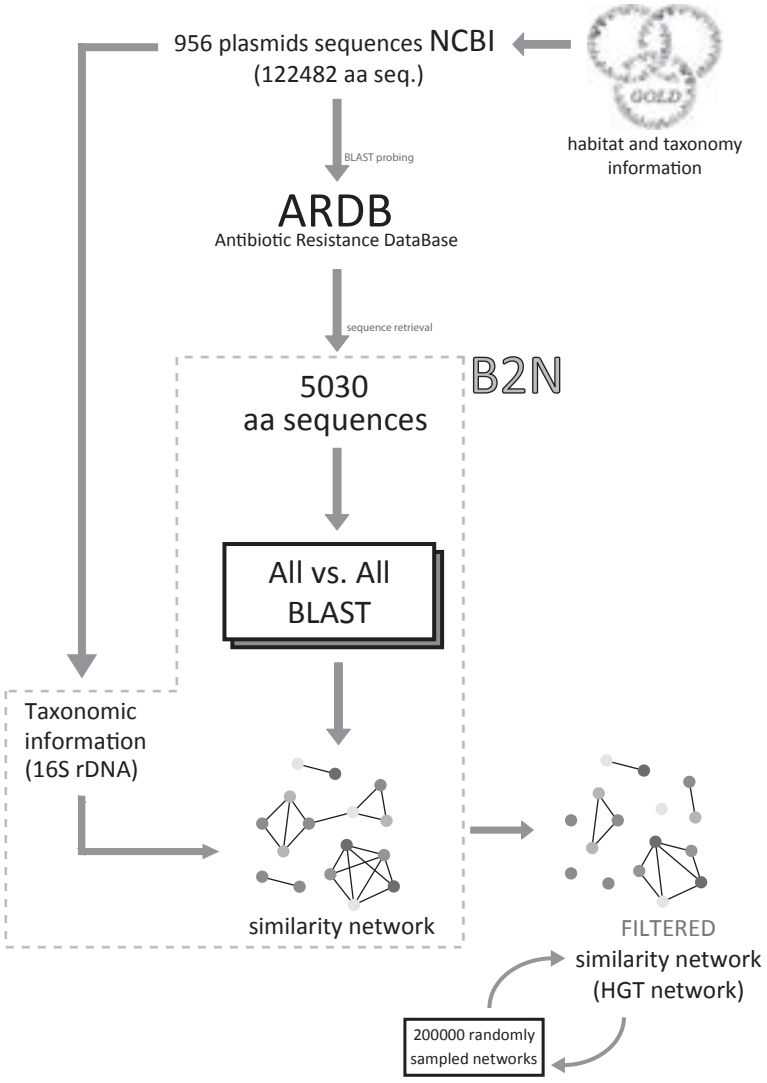


Figure 9.1: Scheme of the data analysis workflow.

distantly related organisms. Hence, in principle, the higher are the NIV the higher is the probability that the genes encoding the linked proteins were transferred between two microroganisms via HGT. Accordingly, link values were modified adopting the following strategy: i) the16S RNA coding genes of each genus represented in the dataset were retrieved from the RDB project (http://rdp.cme.msu.edu/) (for each genus the longest sequences available were chosen and a consensus sequence was built, adopting the tool implemented in the BioEdit package and with a threshold of 80% of frequency to include a nucleotide in the consensus sequence). ii) With this dataset a (square) distance matrix was built using the DNAdist option implemented in the Phylip package [Felsenstein, 1989]. The obtained matrix accounted for the distance between each genus of the dataset, according to their 16 rDNA sequences, and embedded values raging between 0.01789 (for the closest genera in the dataset, i.e. *Escherichia* and *Shigella*) and 0.3675 (the distance between *Staphylococcus* and *Bacteroides*). iii) The matrix was used as input in the B2N flowchart and each identity value of the network was then normalized by computing it as the product of its value and the distance of the genera of the corresponding linked nodes. So, as an example, two identical sequences (link value 100) retrieved from microorganisms belonging to two genera whose 16S rDNA scored 0.15 in the matrix construction, would be linked with a NIV of 15. Conversely, if the same sequences belonged to more closely related genera (e.g. with a 16S rDNA distance of 0.1) their link would acquire a NIV of 10. It is worth of notice that, since a value of 0 was assigned to each position of the matrix diagonal (that is, the distance of all the species belonging to the same genus), we were able to take into account the HGT events occurring only between species belonging to different genera, whereas intra-generic gene transfer (including both horizontal and vertical inheritance and hereinafter referred to as IGT) acquired a NIV of 0. According to the applied selection criteria (see below) these sequences were excluded from further analysis. Hence, although in the next sessions we will use the general definition of "vertical and horizontal inheritance" it must be stated clearly that, according to our normalization method, IGT remained excluded from analysis since we considered horizontally transferred genes exchanged only between species belonging to different genera. In other words, the operational taxonomic unit (OTU) of gene exchange will be considered the bacterial genus.

9.3 Results

9.3.1 Dataset construction and features

The taxonomical and the environmental distribution (retrieved probing the GOLD database) of the assembled dataset is reported in Figure 9.2, which showed that: i) the vast majority (about 80%) of the microorganisms embedded in our dataset could be assigned to a given super-habitat (Figure 9.2a). More in details, 10.2% and 14.2% were represented by microorganisms inhabiting soil and water environments, respectively (hereinafter referred to as soil and water microorganisms). Near half of the dataset was composed by bacteria commonly isolated from hosts (44.7%, hereinafter referred to as host microorganisms),

whereas 10.5% of the dataset was represented by bacteria found in multiple habitats and that we classified as "ubiquitous". As already pointed out by Hooper et al. [2009], each of these super-habitats may include different sub-habitats. In other words, bacteria categorized as inhabiting the human intestinal microflora or those living in the termite gut would both fall within the host super-habitat. Similarly, bacteria colonizing wastewaters or hydrothermal vents were both embedded in the water super-habitat. ii) The analysis of the taxonomical distribution of our dataset (Figure 9.2b) revealed that Proteobacteria account for more than 50% of the total taxonomical diversity. Within Proteobacteria, the γ subdivision is the most represented one (29.67% of the total) followed by α and β subdivisions with 12.36% and 6.04%, respectively. Firmicutes represent another important fraction (20%) of the retrieved antibiotic resistance related sequences. The other mostly represented taxonomical groups are Actinobacteria (9.34%), Spirochaetes (5.49%) and Cyanobacteria (2.74%). Taxonomical units possessing less than 2.5% of the total have been classified as "others" and, together, account for the 11.81% of the total.

9.3.2 Vertically *vs.* horizontally acquired antibiotic resistance genes

A similarity, identity-based, network in which (5030) nodes represent proteins and (259726) links the identity values shared among them was obtained using the sotware B2N as described in Materials and Methods and as represented in Figure 9.3 (the entire, Visone-formatted, network is available as Supplemental Material S3). For this network, the distribution of the identity values and of the normalized identity values (NIV), are shown in Figure 9.3a and b, respectively, whereas the total amount of links at different NIVs is reported in Figure 9.4a. The analysis of Figure 9.3a reveals that most of the sequences share a degree of sequence identity ranging between 20% and 60%, with a peak around 36-38%. Interestingly, we also detected an unexpected peak at 98-100% identity value. After the normalization of links identity values (i.e. the conversion to NIV, see Material and

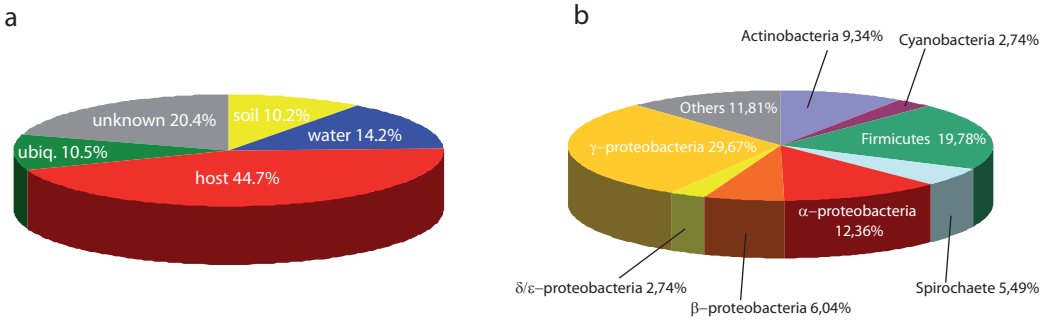


Figure 9.2: Environmental (a) and taxonomical (b) distributions of the organisms used during the analyses.

Methods) two groups of links can be identified, the former exhibiting a NIV ranging from 0 to 1.5 and the latter with a different distribution with an average NIV value around 5.0 (black and grey lines under the histogram in Figure 9.3b, respectively). The purpose of this work was to identify the horizontal flow of antibiotic resistance related genes among the bacterial community rather than those resulting from "vertical" inheritance. In a genome context, this issue can be achieved through different approaches, the most adopted being the analysis of the neighborhood of the genes shared [Hooper *et al.*, 2009] by two distinct microorganisms (indeed, the conservation of those regions would be an indication that these sequences were inherited from a common ancestor) or the bidirectional best hit (BBH) method. However, these approaches can be hardly applied when facing plasmid molecules, given their high variability both in gene content and/or organization. Thus, in order to overcome this limitation, we used a different approach and assumed that if two plasmids from taxonomically distant bacteria share a gene coding for proteins exhibiting a high degree of sequence similarity, then it is very likely that the two genes are

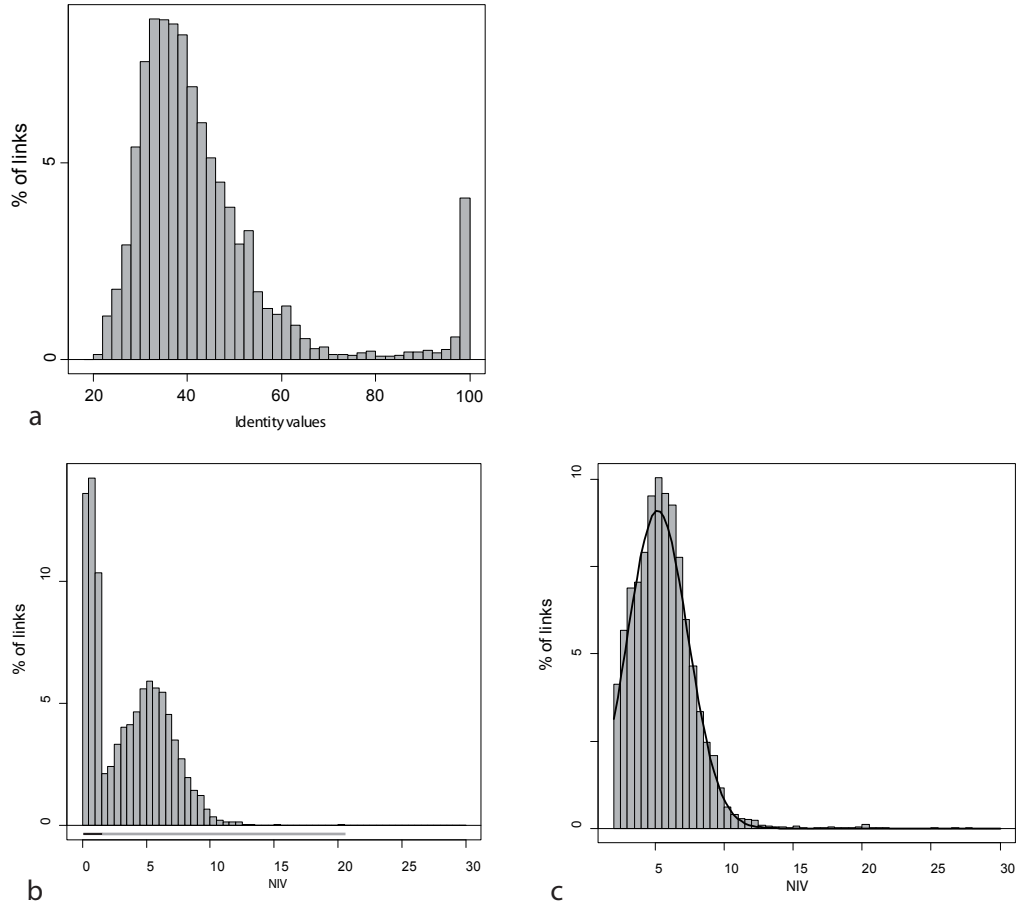


Figure 9.3: Distribution of the 259726 link values (see text for details) in the network embedding the 5030 retrieved antibiotic resistance determinants and representing identity values (a) and normalized identity values (NIV) (b).

the outcome of a HGT event rather than vertical inheritance. Basing on this assumption, we can select from the network those links connecting proteins whose coding genes were horizontally transferred by examining the NlV. Indeed, in principle the higher the NlV, the higher the probability that connected proteins are coded for by horizontally transferred genes. Hence, basing on the procedure of links values normalization, see Materials and Methods, it can be surmised that the first group of links, in which NIVs are comprised within 0 and 1.5 (black line Figure 9.3b, representing about 40% of links), mainly connected (almost) identical sequences from closely related microorganisms, that is sequences unlikely the result of HGT. Accordingly, this group of links was removed from the network and further analyses were focused on the second groups of links (Figure 9.3b, histogram bars above the grey line). In principle, these links might be the outcome of two different evolutionary forces, i.e. HGT between bacteria belonging to different genera and/or vertical inheritance. Interestingly, the distribution of their NIVs fits a normal distribution with mean of 5.6 and a standard deviation (std) of 2.19 (Figure 9.3c). This finding allows to define a 90% confidence interval (CI) as (mean+1.65*std) that, standing to the values of our distribution (Figure 9.3c), corresponds to NIV = (5.6+1.65*2.19)=9.12. This means that, within all the NIVs, the probability of observing a value greater than 9.12 is less than 0.05. Hence, it can be surmised that links with NIV greater than 9.12 are those that connect the most similar sequences from the most distantly related microorganisms, i.e. those sequences that likely underwent horizontal gene transfer (CI>90%). In order to give a support to this assumption, the relative amount of inter- and intra-phylum connections at different NIV thresholds was estimated. 9.4a. Data obtained are reported in Figure

9.4b and revealed that when no threshold (NT column) was applied to NIV, the majority of links (about 70%) is represented by intra-phylum connections (grey bars) rather than connections between organisms belonging to different phyla (black bars). Increasing the NlV threshold resulted in the decrease of the percentage of intra-phylum connections and in the increase of the inter-phylum one (Figure 9.4b). Inter-phylum connections overcame intra-phylum ones at a threshold NlV=3. The difference between inter- and intra-phylum links gradually increased until a NIV comprised between 8 and 9, corresponding to the first major gap between them (70.6% of inter-phylum genera connections vs. 29.4% of intra-phylum genus connections). Hence, it can be surmised that the links embedded in the unfiltered network (without setting any threshold value) were mainly the result of either vertical inheritance (intra-phylum) or IGT. Accordingly, the gradual increasing of the threshold values resulted in the elimination of most of the connections likely related to vertical inheritance or IGT, up to a NIV threshold of 9, where the amount of links related to vertical inheritance is probably minimized. The whole body of data presented in this section, i.e. the analysis of NIV distribution and quantification of inter- and intra-phylum connections suggests that link values higher than 9 were able to connect those nodes that very likely were shared among taxonomically unrelated organisms as the outcome of HGT events. Hence, we set the threshold to 9 and filtered the original network eliminating links with lower values. The resulting network possessed 8607 links (6532 of them involving microorganisms with a defined habitat assignation) and 5030 nodes (2231 of which connected) and was used for further analyses. In this way 50 main clusters (i.e.

with more than 5 nodes) were obtained embedding all major antibiotic resistance related functions and also some clusters embedding aspecific exporters (e.g. ABC transporters and/or MDR proteins) (Table 9.1). Three of them were chosen for a deeper analysis (see below). This choice was based on both the clinical relevance and the (partially) documented history of HGT events of the genes coding for proteins forming the three clusters [Jacobs & Chenia, 2007; Karunaratne *et al.*, 2000; Koike *et al.*, 2007; Lau *et al.*, 2008; Noble *et al.*, 1992]. They comprised TetA sequences (tetracycline efflux protein) involved in tetracycline resistance, CAT sequences (chloramphenicol acetyltransferase), involved in chloramphenicol resistance and AphA sequences (aminoglycoside 3'-phosphotransferase)

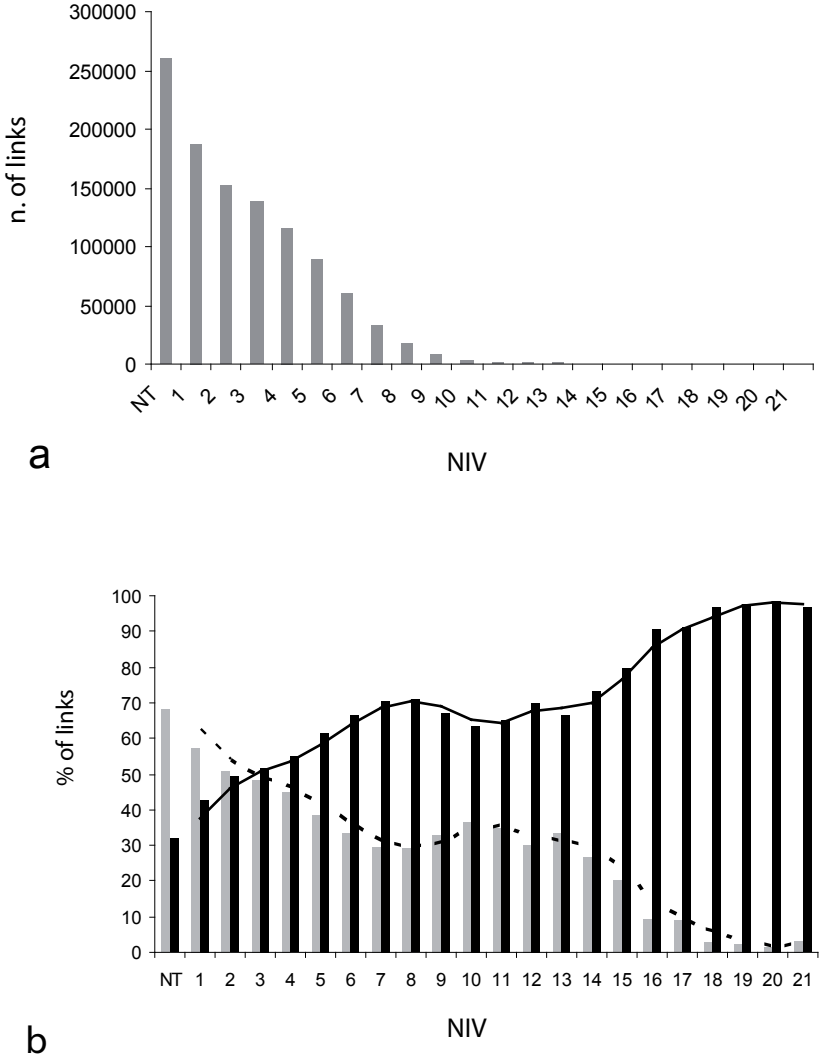


Figure 9.4: (a) Overall amount of links shown at different NIV. (b) Relative amount of inter-phylum (grey bars) and inter-phyla (black) connections at different NIV..

conferring kanamycin resistance (see below for discussion).

Cluster number	Function/Protein	n. of nodes
1	Fosfidomycin reisstance protein	5
2	NodS methyltransferase	5
3	ABC transporter subunit	5
4	Beta lactamase	5
5	RND family, MFP subunit	6
6	Streptothricin acetyltransferase	6
7	VanX	6
8	Vancomycin resistance protein	6
9	QacA efflux transporter	6
10	NreB protein	7
11	ABC related trnasporter	8
12	Acriflavin resistance protein	8
13	aminoglycoside adenyl transferase	8
14	small multidrug efflux protein	8
15	major facilitator transport	8
16	Chloramphenicol exporter	8
17	ABC transporter subunit	8
18	DTP-glucose 4,6 dehydratase	9
19	major facilitator transport	9
20	ABC transporter	9
21	macrolide phosphotransferase	10
22	Formyltethrahydrofolate deformylase	11
23	aminoglycoside acetyl transferase	11
24	MDR protein	12
25	Hypothetical protein	14
26	linomycin	14
27	Tetracycline efflux protein	15
28	aminoglycoside modifying system	15
29	Hypothetical protein	19
30	erythromycin resistance	20
31	Streptogramin resistance protein	23
32	Vancomycin resistance protein	24
33	cation Efflux system	25
34	Heavy metal efflux system	25
35	AphA	26
36	streptomycin resistance	36
37	DfrA	36
38	CAT	40
39	Spectinomycin resistance	46
40	multidrug exporter	47
41	ABC transporter	48
42	streptomycin resistance	49
43	ABC transporter	56
44	MDR protein	56
45	TetA	60
46	ABC transporter	73
47	Beta-lactamase	77
48	Dihydropteorathe synthase	80
49	MDR related protein	213
50	ABC transporter	449

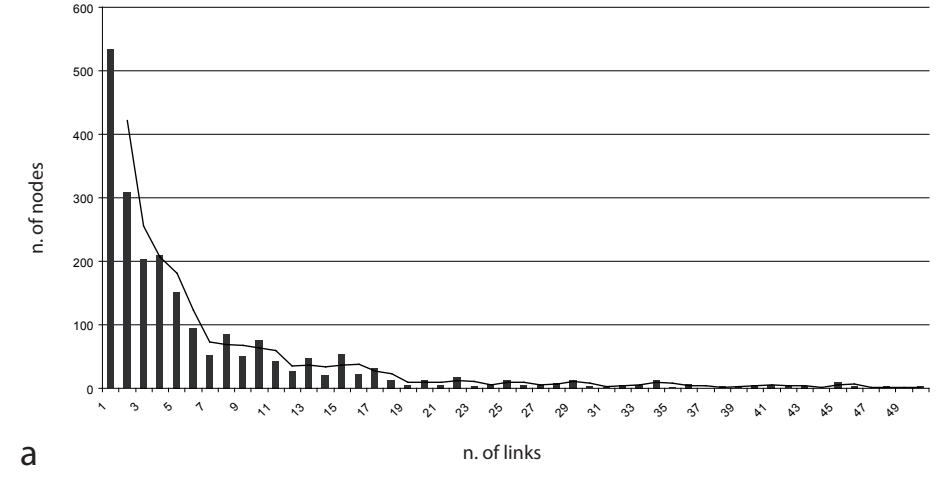
Table 9.1: All the main clusters (embedding 5 or more nodes) retrieved from the overall HGT network. The number of nodes and the functions of the embedded proteins are also reported. Grey rows indicate those clusters that have been selected for deeper analysis (see text for details).

9.3.3 Network properties

The structure of the overall network (5030 nodes and 8607 links) was firstly analyzed. Studying the topology of biological networks might allow the understanding of the structures and dynamics that have been exploited by nature [Dwight Kuo *et al.*, 2006]. A

topological feature often found in large complex networks is the so-called "scale-free" topology [Barabasi & Albert, 1999]. In networks with such a topology, the vertex connectivity [P(k)] distribution, decays as a power-law [Dwight Kuo *et al.*, 2006], that is $P(k) \approx k^{-\gamma}$, with k representing the number of connections. This indicates a non-random structure of the network and the presence of a few highly connected nodes (in this context, proteins) linking the bulk of poorly connected ones. At the light of an HGT network, such a topology might reveal an enhanced tendency of a given bacterium (the one hosting "hub" proteins) in the spreading of antibiotic resistance genes across the microbial community. Recently, scale-free behaviours have been found in many biological networks, including nervous systems [Watts & Strogatz, 1998], metabolic networks [Jeong *et al.*, 2000] protein domains [Wuchty, 2001] and horizontally transferred genes [Dagan *et al.*, 2008]. This prompted us to investigate the distribution of connections in the network built in this work and showing the spreading of antibiotic resistance determinants. To this purpose we determined the number of links (k) exhibited by each node (protein) of the network. Data obtained are reported in Figure 9.5a and revealed that the majority of nodes embedded in the obtained network possess very few connections, whereas highly connected nodes (hubs) are poorly represented. Moreover, nodes degree distribution decreases fallowing a power law (Figure 9.5a) Dividing each point of the histogram by the total number of nodes in the network provided the connectivity P(k), that is the probability that a node has k links (to reduce noise, logarithmic binning was applied). Data obtained are shown in Figure Figure 9.5b and indicated that the connectivity P(k) follows a power-law distribution with γ=3.2. This finding suggests a scale-free behaviour of the network describing the interconnections (HGT events) among antibiotic resistance determinants. In other words, few nodes (hubs) dominate the overall connectivity of the network, that is they exhibit a great number of links, whereas the majority of the nodes has only few connections. This finding suggests that bacteria represented in our network do no contribute equally to the HGT of antibiotic resistance genes. In particular, it seems that a small part of them are able to transfer antibiotic resistance to a broad variety of other microorganisms, whereas for the largest fraction of microbes a very small number of HGT events (links) was retrieved. The analysis revealed the existence of 96 highly interconnected nodes exhibiting more than 30 links (the complete list is available as Supplemental Material S2). The majority of them (53) belonged to microorganisms isolated from a host source as it might be expected from the environmental distribution of the organisms embedded in our dataset (Figure 9.3a) and mainly included *Staphylococcus* species. Hubs assigned to ubiquitous sources (19) included representatives of the *Corynebacterium genus*, although species belonging to *Pseudomonas* and *Enterococcus* genera were also found. Soil and water nodes were less represented, 11 and 12 nodes respectively. Among highly interconnected nodes assigned to a soil habitat, we retrieved *Ralstonia solanacearum* and *Bacillus cereus* species. Finally, Cyanobacteria (*Synechococcus* and *Acaryochloris* genus) and *Aeromonas* species were those possessing nodes with more than 30 links of aquatic microorganisms. Concerning the functions performed by these highly interconnected proteins we found that most of the main functional classes present in ARDB database [Liu & Pop, 2009] were represented and a clear prevalence of a

particular one in respect to the others could not be detected. This latter observation suggests that the network topology and properties are not restricted to a particular function but, instead, can be probably extended to all antibiotic resistance determinants.



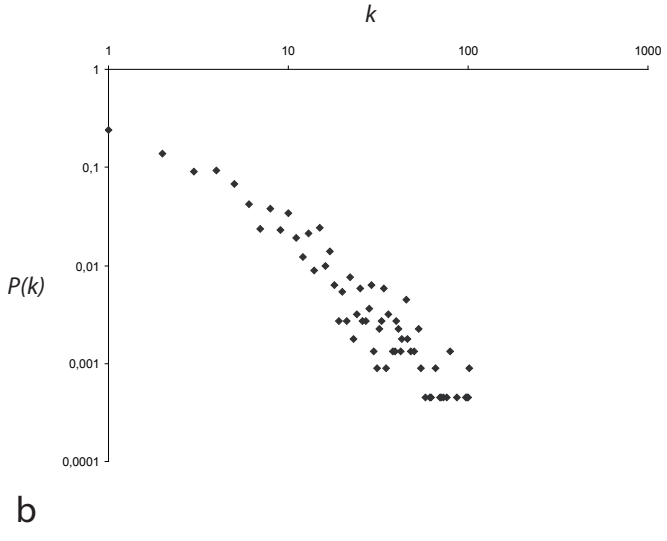


Figure 9.5: Distribution of the number of links (a) and of connectivity P(k) (b) calculated for the filtered network (see text for details). To reduce noise, in (b) logarithmic binning was applied.

9.3.4 Analysis of TetA, CAT, and AphA clusers

Within the whole network, three clusters were chosen for a further analysis, i.e. those embedding TetA, CAT and AphA sequences (Figure 9.6) and accounting for tetracycline, chloramphenicol and kanamycin resistances, respectively. The analysis of these networks revealed both the reliability of the implemented approach and supported previous speculations, based on experimental evidences, on the role of HGT in the spreading of bacterial resistance to these antimicrobial compounds. In fact, it is known that extensive transfer of tetracycline resistance genes has occurred between different species of bacteria that colonize the human body either permanently as members of the normal microflora or transiently as pathogens, because these bacteria frequently interact within the same host [Speer *et al.*, 1992]. Accordingly, several experimental approaches have suggested that HGT might have played a pivotal role in spreading of tetracycline resistance among bacteria belonging to the species embedded in the TetA cluster (Figure 9.6). In details, horizontal gene transfer, especially plasmid mediated, has been recently invoked as an important mechanism for acquisition and dissemination of tetracycline resistance in *Laribacter hongkongensis* [Lau *et al.*, 2008]. Furthermore, three antibiotic resistant strains were identified during a systematic susceptibility screening of *C. glutamicum* and all R-determinants were localized on large plasmids suggesting that the resistances were acquired by horizontal gene transfer [Tauch *et al.*, 2002]. Lastly, the increasing incidence of multidrug resistance amongst *Aeromonas* spp. isolates, which are both fish pathogens and emerging opportunistic human pathogens has been attributed to the horizontal transfer of mobile genetic elements [Jacobs & Chenia, 2007]. Our analysis agrees with these experimental data and further illustrates how geographical and taxonomical distances can be overcome since bacteria that are phylogenetically unrelated (e.g. *Corynebacterium* representatives and Enterobacteria) and/or that live in distinct habitat (e.g. aquatic and host bacteria) can share the same antibiotic resistance determinants, probably following (one or more) HGT(s). Similar considerations can be drawn also in the case of cloramphenicol resistance, a well known problem especially in host bacteria such as representatives of *Staphylococcus* [Bhakta *et al.*, 2003] *Neisseria* [Galimand *et al.*, 1998], *Enterococcus* [Gould *et al.*, 2004] and *Salmonella* [Karunaratne *et al.*, 2000] genera. Moreover, recent investigations have revealed that mariculture waters are inhabited by abundant chloramphenicol-resistant bacteria, mainly represented by *Vibrio* spp. [Dang *et al.*, 2009]. Notably, the presence of links among representatives of all these species (Figure 9.6, CAT cluster), regardless their habitat and/or taxonomical affiliation, strongly suggests that the occurrence of such antibiotic resistance determinants in representatives of these species might be due to HGT events. The third cluster (Figure 9.6, AphA cluster) embedded nodes representative of the product of the gene aphA, encoding aminoglycoside phosphotransferase and mediating resistance to kanamycin. Interestingly, even in this case, bacteria inhabiting different environments and belonging to different taxonomical divisions were found connected (Figure 9.4), thus revealing an intricate horizontal evolutionary pathway of kanamycin resistance. Noteworthy, this finding is partially supported by several independent lines of evidence. The analysis of the genetic organization of pTP10 from *Corynebacterium striatum* M82B, for example, suggested that

the plasmid is composed of eight DNA segments (embedding also kanamycin resistance determinants), the boundaries of which are represented by transposons and insertion sequences [Tauch *et al.*, 2000]. Hence, it has been suggested that, the mosaic structure of pTP10 might represent the evolutionary consolidation into a single plasmid molecule of antimicrobial resistances from microorganisms found in different habitats [Tauch *et al.*,

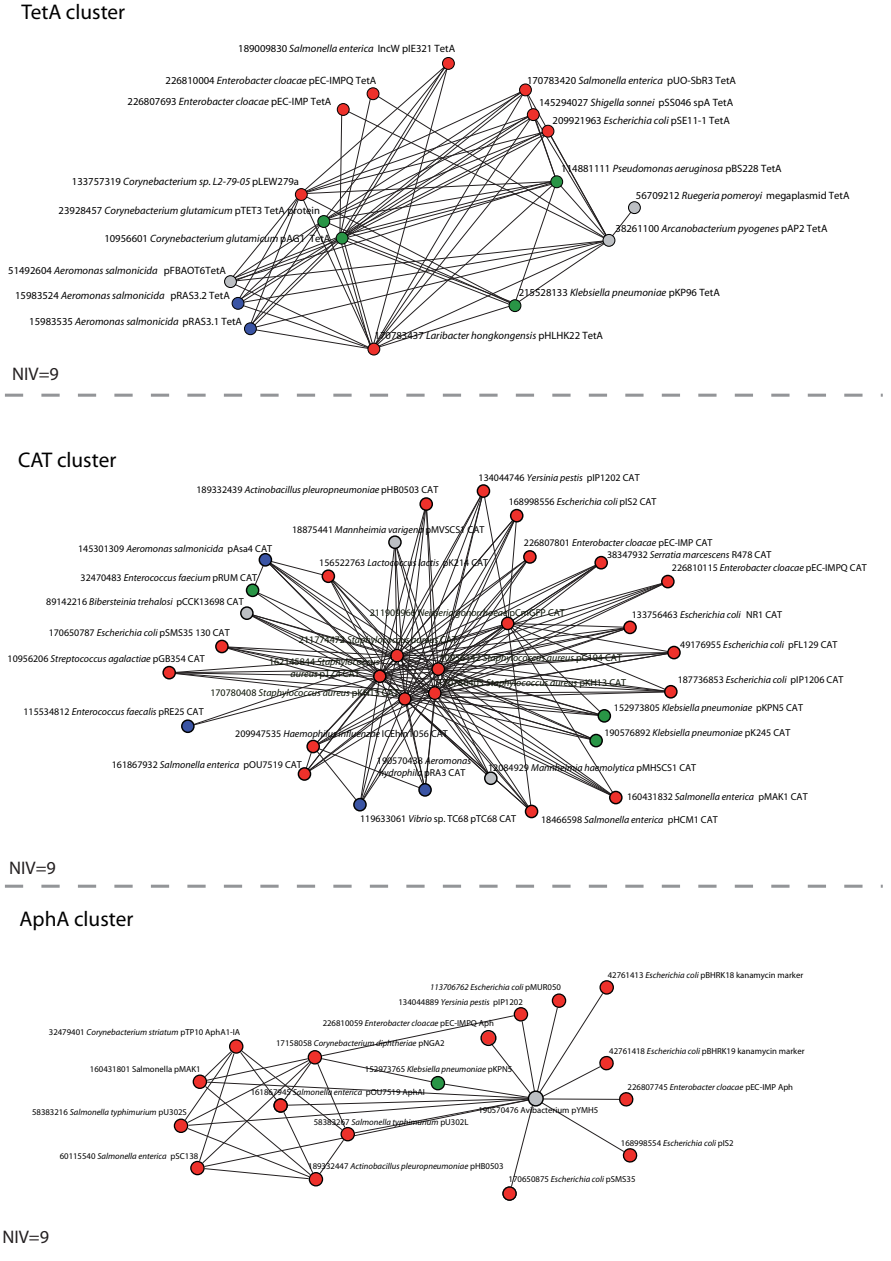


Figure 9.6: TetA, CAT and AphA clusters. Nodes were coloured according to the habitat assigned to their source organisms: yellow-soil, red-host, blue-water and green-ubiquitous. Grey nodes belong to organisms lacking habitat assignation in GOLD database. For clarity purposes, in most cases redundant belonging to the same species are not shown.

2000]. Similarly, *aphA* appears to be widespread in other organisms like representatives of *Avibacterium* [Hsu *et al.*, 2007], *Escherichia* [Gow *et al.*, 2008] and *Salmonella* [Chen *et al.*, 2007] genera. Interestingly, *aph* gene has been recently described for the first time native *Actinobacillus pleuropneumoniae* plasmids [Kang *et al.*, 2009]. Finally, *Klebsiella plasmids* have been shown to encode kanamycin, neomycin, tobramycin, trimethoprim, and sulfonamides resistance, although the plasmid backbones appear to be unrelated, suggesting that translocation of a multiple-drug-resistance-determining region as well as horizontal transfer may have occurred [Partridge & Hall, 2003]. Our analysis supports these findings also suggesting some possible donor and/or acceptor of kanamycin resistance that these plasmids might have encountered during their evolutionary history and, more in general, reveals that the presence of kanamycin resistance on a number of plasmids might be due to the recombination of different plasmids.

9.3.5 Cross-habitat interconnections

9.3.5.1 Links from and to host microorganisms

As previously pointed out, natural environments may represent reservoirs of antibiotic resistance genes for opportunistic and obligate pathogenic bacteria (mainly embedded in the "host" category). For this reason, the amount of antibiotic resistance gene exchange within host bacteria and between host bacteria and those inhabiting different ecological niches was evaluated. The statistical significance of these connections was evaluated by repeating the same calculation for 200000 random networks obtained by re-shuffling existing connections and counting the fraction of times the random network had a number of intra- or inter-habitat connections greater than the observed one. This gives a *p-value* accounting for the statistical significance of the number of intra- or inter-habitat connections of the original network. Interestingly, we found that connections linking host organisms with organisms assigned to multiple environments (ubiquitous) were statistically enriched ($p\text{-}value<10^{-4}$) (Table 9.2). Among ubiquitous microorganisms included in the dataset, *Klebsiella pneumoniae* possessed a high number of connections with host bacteria (385) accounted for by 61 different nodes. Remarkably, these nodes are linked to microorganisms belonging to most of the taxonomical units represented in the dataset and including Enterobacteria, *Staphylococcus genus*, and plant hosts (*Agrobacterium* and *Rhizobium* genera). Similarly, we observed a statistically significant enrichment ($p\text{-}value<10^{-4}$) evaluating the connections within host microorganisms. Among the 2373 connections retrieved, we found major animal symbionts and pathogens (such as *E. coli*, *S. aureus* and *Actinobacillus pleuropneumoniae*) although plant hosts were also retrieved (i.e. *Rhizobium leguminosarum* and *Agrobacteirum vitis*). Conversely, we found that soil and water bacteria, (626 and 821 links, respectively) are significantly ($p\text{-}value<10^{-4}$ and $p\text{-}value= 0.001$, respectively) less connected to host bacteria than what expected if connections were distributed randomly in the whole network (Table 9.2). Soil microorganisms majorly connected to host bacteria all belong to α-proteobacteria. Nodes belonging to *Ralstonia* genus are the most represented (being involved in 242 connections) although *Sinorhizobium* species are also present. Water bacteria found in connection with host nodes mainly include *Acaryochloris*

Habitat	N. of links	*p-vaue*
host-host	2373	$<10^{-4}$
host-ubiquitous	1509	$<10^{-4}$
water-ubiquitous	264	0.04
water-water	150	$<10^{-4}$
host-water	821	0.001
soil-host	626	$<10^{-4}$
soil-water	272	0.002
soil-ubiquitous	216	$<10^{-4}$
soil-soil	172	*n.s.*
ubiquitous-ubiquitous	129	$<10^{-4}$
TOTAL	6532	

Table 9.2: Total amount and statistical significance of inter and intra habitat connections. Grey rows indicate enriched connections in respect to expected values from random network construction. Conversely white rows indicate connections less represented than expected by chance. n.s. stands for "statistically not significant", i.e. *p-value*> 0.05.

marina MBIC11017, *Polaromonas* and *Aeromonas* species.

9.3.5.2 Other cross-habitat interconnections

Other cross-habitat interconnections accounted for 752 connections of the whole network, and included:

1. 272 co-occurrences among microorganisms inhabiting soil and water environments were found. Probably not surprisingly, most of the organisms responsible for these connections are related to soil and agriculture and include *Paracoccus denitrificans* PD1222, *Sinorhizobium* species and *Ralstonia solanacearum* GMI1000 (all belonging to the α subdivision of proteobacteria). Aquatic bacteria responsible for these connections belonged to cyanobacterial species such as *Acaryochloris marina* MBIC11017 and *Synechococcuss* species. The amount of connections retrieved between soil and water microorganisms is significantly lower (*p-value*$= 0.002$) than what expected by chance.

2. 264 and 216 links connected ubiquitous with water and soil microorganisms respectively. Interestingly, water-ubiquitous connections were significatively enriched (*p-value*$= 0.04$) whereas the number of links between soil and ubiquitous was lower than what expected by chance (*p-value*$< 10^{-4}$). In both cases, among ubiquitous organisms, *Burkholderia phymatum* STM815 and *Klebsiella* nodes were found. Obtained inter- and intra-habitat connections, together with their corresponding statistical support, are reported in Figure 9.7.

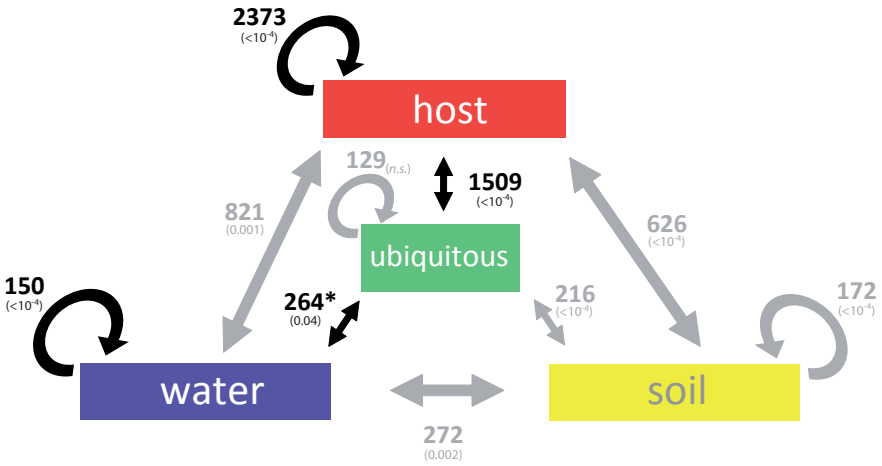


Figure 9.7: Schematic representation of retrieved inter- and intra-habitat links (represented by arrows). Black and grey rows indicate over- and under- represented values (in comparison with 200000 randomly sampled networks), respectively. *p-value* are reported in parentheses. n.s. stands for "statistically not significant", i.e. *p-value*>0.05.

.

9.4 Discussion

The horizontal spreading of antibiotic resistance genes across the whole microbial community represents a crucial point in the attempt of controlling and avoiding the emergence of large bacterial infections and multi-resistant strains. Classical experimental approaches for the detection of antibiotic genes flows across bacterial populations are able to depict a detailed scenario only for a limited number of species, habitats and/or genes. To our knowledge, the one proposed in this work, is the first attempt to give a complete picture of the antibiotic resistance spreading in the whole microbial community, considering most of the antibiotic resistance molecular strategies and also inter- and intra-habitat exchanges across the whole bacterial community present in public databases. By combining sequence identity data of the proteins involved in antibiotic resistance and the one coming from more reliable molecular markers (16S rDNA sequences) we were able to construct a network embedding 5030 nodes (proteins) and 8607 links (normalized sequence identity values) representing the (putatively) horizontally transferred antibiotic resistance determinants of our dataset.

In recent years, the analysis of the topology of (biological) networks has provided useful hints in the understanding of global properties of several different systems (e.g. metabolic networks, protein-protein interaction) and has led some authors to claim that networks may represent a means of reconstructing microbial genome evolution (also accounting for the incorporation of foreign genes) [Dagan *et al.*, 2008]. The analysis of the distribution of the connections in the network of horizontally transferred antibiotic resistance genes revealed that the majority of the nodes possessed a few connections, whereas a very small part of them had a greater (>30) number of connections. Calculating the probability that

a node has k links (P(k)) revealed that that the connectivity follows a power-law distribution (Figure 9.5b). This finding suggests that the network describing the flow of genes encoding antibiotic resistance proteins is a scale free network, in which a small number of nodes account for the majority of the connections (hubs). From a biological point of view, this behaviour might be accounted for by the presence of a small number of microorganisms capable of interacting with a wide spectrum of microbes and thus revealing their likely pivotal role in the sharing of antibiotic resistance throughout the bacterial world. In our opinion this finding might provide useful insights during infection control procedures, since focusing our attentions on particular species (those possessing hub nodes) may prevent antibiotic resistance dissemination within certain habitats. Interestingly, in fact, it is known [Albert *et al.*, 2000] that scale-free networks, despite showing robustness towards random errors (e.g. the random removal of nodes) are extremely vulnerable to attacks, that is the selection and removal of a few nodes that play a vital role in maintaining the network's connectivity.

Our analysis revealed that one of these "vectors" might be represented by *Staphylococcus aureus* species, suggesting that this group of microorganisms might play a key role in the spreading of bacterial resistance. After host microorganisms, bacteria viable through different habitats (ubiquitous) were those possessing the highest number of hubs. Among these, we found species belonging to the *Corynebacterium* (*glutamicum*, *striatum* and sp. L2-79-05) and *Enterococcus* (*faecium*) genera. It is worth of notice that *Corynebacterium* representatives are found in a broad variety of habitats such as soil, plants and food products [Yassin *et al.*, 2003]. Similarly *Enteroccoccus* genus is a widely distributed microbial group, with representatives isolated from fruits and vegetable foods, water and soil, and clinical samples [Abriouel *et al.*, 2008]. Hence, bacteria belonging to these species might represent optimal vectors for the spreading of antibiotic resistance through geographically separated bacterial communities, playing a fundamental role in the mobilization of the whole bacterial antibiotic resistance gene pool. Moreover, it can be surmised that organisms viable through several habitats might be versatile enough to adapt to fluctuating conditions as well as diverse environments, thus having the opportunity to interact with more different microbial communities, as proposed also by Hooper et al. [2009]. By counting the number of connections between microorganisms inhabiting different habitats (inter- and intra.connections) and comparing them with those expected from 200000 randomly constructed networks, we found that strains commonly found in diverse environmental settings (i.e. soil, water and ubiquitous) contribute at a different extent to horizontal transfer of antibiotic resistance determinants with host bacteria (the category also embedding most of the known pathogens). In fact, we observed a statistically supported enrichment of links between ubiquitous and host bacteria. Interestingly, among them we found *Klebsiella pneumoniae*, an important pathogen both in the community and the hospital setting with widespread multi-drug resistance (MDR) [Keynan & Rubinstein, 2007]. Both these features (MDR and ubiquity) might be the cause of the connections established with host microorganisms. Furthermore, the fact that they are capable of connecting to a wide group of phylogenetically unrelated bacteria revealed that, not only geographical, but also taxonomical barriers can be "easily"

overcome in the spreading of antimicrobial determinants. Soil and water microorganisms apparently contribute to a lesser extent in direct gene exchange between host bacteria since the number of observed connections is lower than what expected by chance. Among soil microorganisms, genera belonging to the α subdivision of Proteobacteria (particularly *Sinorhizobium* and *Ralstonia*) were retrieved. This finding is in agreement with previous data on microbial interaction networks from Hooper et al. [2009]. However, these authors that the interactions between these groups of microbes are likely to occur outside of host organisms, for instance in agricultural manure or waste-water. Nevertheless, some *Ralstonia* representatives were recently recovered in respiratory cultures taken from infected nosocomial patients [Jhung *et al.*, 2007], while *Sinorhizobium* representatives were found in strict association with both hunts (*Tetraponera*) [van Borm *et al.*, 2002] and nematodes *Caenorhabditis elegans*) [Horiuchi *et al.*, 2005] extending their current view as simple soil and plant-associated bacteria. We propose that these associations, although sporadic and perhaps transient, may represent the right setting for interacting with host microorganisms and, possibly, exchanging genetic information. This finding also suggests that some of the current habitat assignments might be redefined and updated and should not include only the environment in which a given microorganism has been firstly isolated since it may not correspond to the one in which that microbe is usually found. Although to a less extent than what expected by chance (Table 9.2), some nodes belonging to aquatic microorganisms and found connected to host bacteria included representatives of *Acaryochloris marina* MBIC11017 and *Aeromonas hydrophila* and *A. salmonicida*. Interestingly, *A. marina* was isolated as a minor symbiont from a colonial ascidian [Kuhl *et al.*, 2005] revealing the capability of this bacterium to interact with other organisms and partially explaining the presence of interconnections established with host source bacteria. *Aeromonas* species were assigned to water environment in the GOLD database; however, the debate of whether this bacterium is capable of living outside its host (fishes) is not yet been solved [Deere *et al.*, 1996; Rose *et al.*, 1990; Sakai, 1986]. Interestingly, *Aeromonas* from aquaculture water systems (fish, eel farming) are particularly resistant to antibiotics [Penders & Stobberingh, 2008], and frequently contain plasmids and integrons with genes for multiple antibiotic resistance [Jacobs & Chenia, 2007]. Our analysis revealed that HGT events with host bacteria might have played a key role in the evolution of antimicrobial resistance in these strains (and vice versa). Consistent with these data is the finding that estuarine water-borne *Aeromonas* strains carry almost as frequently as Enterobacteriaceae class 1 integron platforms carrying antibiotic-resistance genes [Henriques *et al.*, 2006].

The whole body of data reported in this work has shown the importance of environmental bacteria in the evolution and in the spreading of the resistome. As suggested by Baquero et al. [2008], several different genetic reactors may exist in nature (see Introduction). According to our analysis two of them seem to play a major role in the whole HGT network, namely host and water environments (Figure 9.7). The problem of HGT between host microorganisms is a well known one. In fact, the emergence of resistant strains in patients (representing a primary genetic reactor for the sharing and the re-assembling of the resistome) is a major clinical issue. Moreover, it has been suggested [Baquero *et al.*, 2008]

that, in water, bacteria from different origins (human, animal, environmental) are able to mix, and resistance evolves as a consequence of promiscuous exchange and shuffling of genes, genetic platforms, and genetic vectors. The resulting genetic innovations might be mobilized and transferred between these different environmental settings by ubiquitous microbes, as suggested by the statistically supported enrichment of host-ubiquitous and water-ubiquitous links we observed within our dataset (Table 9.2 and Figure 9.7). Lastly, results obtained partially overlap with those previously obtained adopting a similar approach and using chromosomal transposases as molecular markers [Hooper *et al.*, 2009] and with those coming from experimental identification of antibiotic resistance determinants in both clinical and environmental settings.

9.5 Conclusions

Despite probably not exhaustive (the whole body of sequenced genomes might be biased towards certain genera and/or habitats) the analyses presented in this work represent a first attempt to give an almost complete picture of the horizontal flow of antibiotic resistance determinants at the whole bacterial community system level. Data obtained revealed that, although antibiotic resistance has mainly received clinically-centered attentions, environmental bacteria, and especially those viable through multiple environments (ubiquitous) and those inhabiting water environments, are likely to play (and to have played) a pivotal role in the emergence of bacterial resistance towards antimicrobial compounds. In fact, these bacteria are those that most of all are involved in the mobilization of the resistome gene pool and those apparently capable of crossing both geographical and taxonomical barriers. Furthermore, our data provide other support to the previous idea [Hooper *et al.*, 2009] according to which some current environmental annotations might be too restrictive, failing to represent the actual extent of microbes' habitat range. Finally, by suggesting which bacterial species seems to act as vectors of antibiotic resistance determinants, in our opinion the approach implemented in this work might provide useful material during infections control procedures.

References

Abriouel, H., Omar, N.B., Molinos, A.C., Lopez, R.L., Grande, M.J., Martinez-Viedma, P., Ortega, E., Canamero, M.M. & Galvez, A. (2008). Comparative analysis of genetic diversity and incidence of virulence factors and antibiotic resistance among enterococcal populations from raw fruit and vegetable foods, water and soil, and clinical samples. *Int J Food Microbiol*, **123**, 38–49.

Albert, R., Jeong, H. & Barabasi, A.L. (2000). Error and attack tolerance of complex networks. *Nature*, **406**, 378–82.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.

Baquero, F., Martinez, J.L. & Canton, R. (2008). Antibiotics and antibiotic resistance in water environments. *Curr Opin Biotechnol*, **19**, 260–5.

Barabasi, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–12.

Bennett, P.M. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *Br J Pharmacol*, **153 Suppl 1**, S347–57.

Bhakta, M., Arora, S. & Bal, M. (2003). Intraspecies transfer of a chloramphenicol-resistance plasmid of staphylococcal origin. *Indian J Med Res*, **117**, 146–51.

Brilli, M., Mengoni, A., Fondi, M., Bazzicalupo, M., Lio, P. & Fani, R. (2008). Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using blast2network. *BMC Bioinformatics*, **9**, 551.

Cattoir, V., Poirel, L., Aubert, C., Soussy, C.J. & Nordmann, P. (2008). Unexpected occurrence of plasmid-mediated quinolone resistance determinants in environmental aeromonas spp. *Emerg Infect Dis*, **14**, 231–7.

Charles, P.G. & Grayson, M.L. (2004). The dearth of new antibiotic development: why we should be worried and what we can do about it. *Med J Aust*, **181**, 549–53.

Chen, C.Y., Nace, G.W., Solow, B. & Fratamico, P. (2007). Complete nucleotide sequences of 84.5- and 3.2-kb plasmids in the multi-antibiotic resistant salmonella enterica serovar typhimurium u302 strain g8430. *Plasmid*, **57**, 29–43.

Dagan, T., Artzy-Randrup, Y. & Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A*, **105**, 10039–44.

Dang, H., Zhao, J., Song, L., Chen, M. & Chang, Y. (2009). Molecular characterizations of chloramphenicol- and oxytetracycline-resistant bacteria and resistance genes in mariculture waters of china. *Mar Pollut Bull*, **58**, 987–94.

D'Costa, V.M., McGrann, K.M., Hughes, D.W. & Wright, G.D. (2006). Sampling the antibiotic resistome. *Science*, **311**, 374–7.

Deere, D., Porter, J., Pickup, R.W. & Edwards, C. (1996). Survival of cells and dna of aeromonas salmonicida released into aquatic microcosms. *J Appl Bacteriol*, **81**, 309–18.

Dugan, J., Rockey, D.D., Jones, L. & Andersen, A.A. (2004). Tetracycline resistance in chlamydia suis mediated by genomic islands inserted into the chlamydial inv-like gene. *Antimicrob Agents Chemother*, **48**, 3989–95.

Dwight Kuo, P., Banzhaf, W. & Leier, A. (2006). Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems*, **85**, 177–200.

Felsenstein, J. (1989). Mathematics vs. evolution: Mathematical evolutionary theory. *Science*, **246**, 941–942.

Fluit, A.C., Visser, M.R. & Schmitz, F.J. (2001). Molecular detection of antimicrobial resistance. *Clin Microbiol Rev*, **14**, 836–71, table of contents.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Brief Bioinform*, **7**, 225–42.

Galimand, M., Gerbaud, G., Guibourdenche, M., Riou, J.Y. & Courvalin, P. (1998). High-level chloramphenicol resistance in neisseria meningitidis. *N Engl J Med*, **339**, 868–74.

Gould, C.V., Fishman, N.O., Nachamkin, I. & Lautenbach, E. (2004). Chloramphenicol resistance in vancomycin-resistant enterococcal bacteremia: impact of prior fluoroquinolone use? *Infect Control Hosp Epidemiol*, **25**, 138–45.

Gow, S.P., Waldner, C.L., Harel, J. & Boerlin, P. (2008). Associations between antimicrobial resistance genes in fecal generic escherichia coli isolates from cow-calf herds in western canada. *Appl Environ Microbiol*, **74**, 3658–66.

Henriques, I.S., Fonseca, F., Alves, A., Saavedra, M.J. & Correia, A. (2006). Occurrence and diversity of integrons and beta-lactamase genes among ampicillin-resistant isolates from estuarine waters. *Res Microbiol*, **157**, 938–47.

Hooper, S.D., Mavromatis, K. & Kyrpides, N.C. (2009). Microbial co-habitation and lateral gene transfer: what transposases can tell us. *Genome Biol*, **10**, R45.

Horiuchi, J., Prithiviraj, B., Bais, H.P., Kimball, B.A. & Vivanco, J.M. (2005). Soil nematodes mediate positive interactions between legume plants and rhizobium bacteria. *Planta*, **222**, 848–57.

Hsu, Y.M., Shieh, H.K., Chen, W.H., Sun, T.Y. & Shiang, J.H. (2007). Antimicrobial susceptibility, plasmid profiles and haemocin activities of avibacterium paragallinarum strains. *Vet Microbiol*, **124**, 209–18.

Jacobs, L. & Chenia, H.Y. (2007). Characterization of integrons and tetracycline resistance determinants in aeromonas spp. isolated from south african aquaculture systems. *Int J Food Microbiol*, **114**, 295–306.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–4.

Jhung, M.A., Sunenshine, R.H., Noble-Wang, J., Coffin, S.E., St John, K., Lewis, F.M., Jensen, B., Peterson, A., LiPuma, J., Arduino, M.J., Holzmann-Pazgal, G., Atkins, J.T. & Srinivasan, A. (2007). A national outbreak of ralstonia mannitolilytica associated with use of a contaminated oxygen-delivery device among pediatric patients. *Pediatrics*, **119**, 1061–8.

Kang, M., Zhou, R., Liu, L., Langford, P.R. & Chen, H. (2009). Analysis of an actinobacillus pleuropneumoniae multi-resistance plasmid, phb0503. *Plasmid*, **61**, 135–9.

Karunaratne, G.K., Wickremesinghe, R.S. & Perera, K.C. (2000). Salmonella typhi and chloramphenicol resistance. *Ceylon Med J*, **45**, 136–7.

Keynan, Y. & Rubinstein, E. (2007). The changing face of klebsiella pneumoniae infections in the community. *Int J Antimicrob Agents*, **30**, 385–9.

Kohiyama, M., Hiraga, S., Matic, I. & Radman, M. (2003). Bacterial sex: playing voyeurs 50 years later. *Science*, **301**, 802–3.

Koike, S., Krapac, I.G., Oliver, H.D., Yannarell, A.C., Chee-Sanford, J.C., Aminov, R.I. & Mackie, R.I. (2007). Monitoring and source tracking of tetracycline resistance genes in lagoons and groundwater adjacent to swine production facilities over a 3-year period. *Appl Environ Microbiol*, **73**, 4813–23.

Kuhl, M., Chen, M., Ralph, P.J., Schreiber, U. & Larkum, A.W. (2005). Ecology: a niche for cyanobacteria containing chlorophyll d. *Nature*, **433**, 820.

Lau, S.K., Wong, G.K., Li, M.W., Woo, P.C. & Yuen, K.Y. (2008). Distribution and molecular characterization of tetracycline resistance in laribacter hongkongensis. *J Antimicrob Chemother*, **61**, 488–97.

Liolios, K., Mavromatis, K., Tavernarakis, N. & Kyrpides, N.C. (2008). The genomes on line database (gold) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, **36**, D475–9.

Liu, B. & Pop, M. (2009). Ardb–antibiotic resistance genes database. *Nucleic Acids Res*, **37**, D443–7.

Mackie, R.I., Koike, S., Krapac, I., Chee-Sanford, J., Maxwell, S. & Aminov, R.I. (2006). Tetracycline residues and tetracycline resistance genes in groundwater impacted by swine production facilities. *Anim Biotechnol*, **17**, 157–76.

Martinez, J.L. (2008). Antibiotics and antibiotic resistance genes in natural environments. *Science*, **321**, 365–7.

Martinez, J.L., Fajardo, A., Garmendia, L., Hernandez, A., Linares, J.F., Martinez-Solano, L. & Sanchez, M.B. (2009). A global view of antibiotic resistance. *FEMS Microbiol Rev*, **33**, 44–65.

Noble, W.C., Virani, Z. & Cree, R.G. (1992). Co-transfer of vancomycin and other resistance genes from enterococcus faecalis nctc 12201 to staphylococcus aureus. *FEMS Microbiol Lett*, **72**, 195–8.

Partridge, S.R. & Hall, R.M. (2003). In34, a complex in5 family class 1 integron containing orf513 and dfra10. *Antimicrob Agents Chemother*, **47**, 342–9.

Penders, J. & Stobberingh, E.E. (2008). Antibiotic resistance of motile aeromonads in indoor catfish and eel farms in the southern part of the netherlands. *Int J Antimicrob Agents*, **31**, 261–5.

Rose, A.S., Ellis, A.E. & Munro, A.L. (1990). Evidence against dormancy in the bacterial fish pathogen aeromonas salmonicida subsp. salmonicida. *FEMS Microbiol Lett*, **56**, 105–7.

Sakai, D.K. (1986). Electrostatic mechanism of survival of virulent aeromonas salmonicida strains in river water. *Appl Environ Microbiol*, **51**, 1343–9.

Scaria, J., Chandramouli, U. & Verma, S.K. (2005). Antibiotic resistance genes online (argo): a database on vancomycin and beta-lactam resistance genes. *Bioinformation*, **1**, 5–7.

Schluter, A., Krause, L., Szczepanowski, R., Goesmann, A. & Puhler, A. (2008). Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. *J Biotechnol*, **136**, 65–76.

Sobecky, P.A. (2002). Approaches to investigating the ecology of plasmids in marine bacterial communities. *Plasmid*, **48**, 213–21.

Speer, B.S., Shoemaker, N.B. & Salyers, A.A. (1992). Bacterial resistance to tetracycline: mechanisms, transfer, and clinical significance. *Clin Microbiol Rev*, **5**, 387–99.

Tauch, A., Krieft, S., Kalinowski, J. & Puhler, A. (2000). The 51,409-bp r-plasmid ptp10 from the multiresistant clinical isolate corynebacterium striatum m82b is composed of dna segments initially identified in soil bacteria and in plant, animal, and human pathogens. *Mol Gen Genet*, **263**, 1–11.

Tauch, A., Gotker, S., Puhler, A., Kalinowski, J. & Thierbach, G. (2002). The 27.8-kb r-plasmid ptet3 from corynebacterium glutamicum encodes the aminoglycoside adenyltransferase gene cassette aada9 and the regulated tetracycline efflux system tet 33 flanked by active copies of the widespread insertion sequence is6100. *Plasmid*, **48**, 117–29.

Tenover, F.C. (2006). Mechanisms of antimicrobial resistance in bacteria. *Am J Infect Control*, **34**, S3–10; discussion S64–73.

van Borm, S., Buschinger, A., Boomsma, J.J. & Billen, J. (2002). Tetraponera ants have gut symbionts related to nitrogen-fixing root-nodule bacteria. *Proc Biol Sci*, **269**, 2023–7.

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–2.

Wright, G.D. (2007). The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol*, **5**, 175–86.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Mol Biol Evol*, **18**, 1694–702.

Yassin, A.F., Kroppenstedt, R.M. & Ludwig, W. (2003). Corynebacterium glaucum sp. nov. *Int J Syst Evol Microbiol*, **53**, 705–9.

Chapter 10

Structure and Evolution of HAE1 and HME efflux systems in *Burkholderia* genus

10.1 Introduction

The genus *Burkholderia* is an interesting and complex bacterial taxonomic unit that includes a variety of species inhabiting different ecological niches ([Compant *et al.*, 2008] and references therein). In recent years a growing number of *Burkholderia* strains and species have been reported as plant-associated Bacteria. Indeed, *Burkholderia* spp. can be free-living in the rhizosphere as well as epiphytic and endophytic, including obligate endosymbionts and phytopathogens. Several strains are known to enhance disease resistance in plants, contribute to better water management, and improve nitrogen fixation and overall host adaptation to environmental stresses ([Compant *et al.*, 2008] and references therein). On the other side, some species/isolates can be opportunistic or obligate pathogens causing human, animal or plant disease. Interaction between *Burkholderia* species and humans or animals are traditionally known for *B. mallei* and *B. pseudomallei*, that are the aetiological agent of glanders and melioidosis, respectively [Coenye & Vandamme, 2003]. Lastly, several *Burkholderia* species have been demonstrated to be opportunistic pathogens in humans. Although they are not considered pathogens for the normal human population, some are serious threats for specific patient groups. These species include *B. gladioli*, *B. fungorum* and all *B. cepacia* complex (BCC) bacteria [Coenye & Vandamme, 2003]. The BCC is a group of genetically distinct but phenotypically similar bacteria that up to now comprises seventeen closely related bacterial species [Compant *et al.*, 2008; Vanlaere *et al.*, 2008, 2009], and they are important opportunistic pathogens that infect the airways of cystic fibrosis (CF) patients [Govan *et al.*, 2007]. *Burkholderia* human infections are usually treated with antibiotics in order to improve disease control and patient survival. The increasing bacterial resistance to these molecules has become a public health problem. In this context, it seems more and more evident that the intrinsic resistance of many bacteria to antibiotics depends on the constitutive or inducible expression of active efflux systems [Nikaido, 2001; Ryan *et al.*, 2001].

This is particularly true for multidrug efflux pumps allowing bacterial cells to extrude a wide range of different substrates, including antibiotics. In contrast with other bacterial genes, encoding antibiotic resistance, acquired by horizontal gene transfer (HGT) [Martinez *et al.*, 2009], genes coding for multidrug efflux pumps are mainly harboured by the chromosome(s) of living organisms. In addition, these genes are highly conserved and their expression is tightly regulated [Martinez *et al.*, 2009]. Taken together, these characteristics suggest that the main function of these systems is likely not conferring resistance to antibiotics (used in therapy) and that they might play other roles relevant to the behaviour of bacteria in their natural ecosystems. Among the potential roles, it has been demonstrated that efflux pumps are important for detoxification processes of intracellular metabolites, bacterial virulence in both animal and plant hosts, cell homeostasis and intercellular signal trafficking [Martinez *et al.*, 2009]. This class of proteins includes a very interesting group, referred to as the RND (Resistance-Nodulation-Cell Division) superfamily, that is found ubiquitously in all the three cells domains (Bacteria, Archaea and Eucarya), being mainly involved in drug resistance of Gram-negative bacteria [Paulsen *et al.*, 1996; Poole, 2007]. Functionally characterized members of this superfamily fall into eight different families: three of them are largely restricted to Gram-negative bacteria; the other five families have a diverse phylogenetic distribution (Figure 10.1). The three families peculiar of Gram-negative Bacteria have a different substrate specificity, with one catalyzing the export of heavy metals [Heavy Metal Efflux (HME)], one responsible for the export of multiple drugs [Hydrophobe/Amphiphile Efflux-1 (HAE-1)], and the last one likely catalyzing the export of lipooligosaccharides concerned with plant nodulation related to symbiotic nitrogen fixation [putative Nodulation Factor Exporter (NFE)] [Saier & Paulsen, 2001] (Figure 10.1). In Gram-negative bacteria RND transporters act

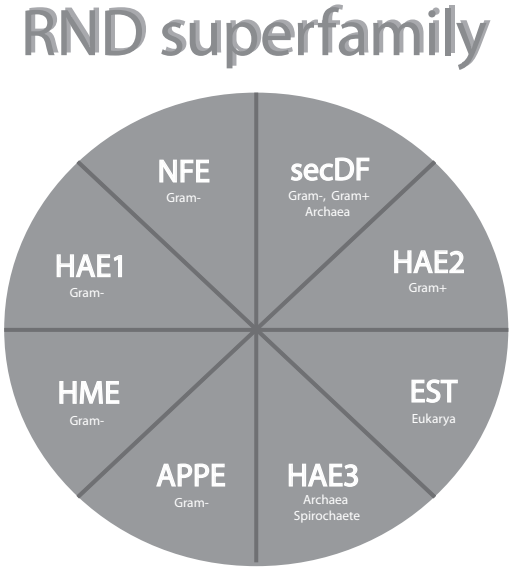


Figure 10.1: Schematic representation of the RND superfamily.

as a complex that can bind various structurally unrelated substrates from the periplasm

and/or from the cytoplasm and extrude them out directly into the external media using proton-motive force. This enzymatic complex is composed of a RND protein, located in the cytoplasmatic membrane, a periplasmic-located membrane adaptor protein, belonging to the membrane fusion protein family (MFP), and an outer-membrane channel protein (OMP) [Saier & Paulsen, 2001]. Typically, the encoding genes are organized in an operon and the MFP and the RND are usually cotranscribed [Poole *et al.*, 1993], whereas in some systems and/or species, the OMP is not linked to the other genes [Aires *et al.*, 1999; Ma *et al.*, 1993]. Most of the RND superfamily transport systems consists of a polypeptide chain 700-1300 amino acid residues long. These proteins possess a single transmembrane spanner (TMS) at their N-terminus followed by a large extracytoplasmic domain, six additional TMSs, a second large extracytoplasmatic domain, and five final C-terminal TMSs. Most RND permeases consist of a single polypeptide chain [Saier & Paulsen, 2001]. The first half of RND family proteins is homologous to the second one, suggesting that the coding gene is the outcome of an intragenic tandem duplication event of an ancestral gene (i.e. a gene elongation event [Fani, 2009]) that occurred in the primordial system prior to the divergence of the family members [Saier *et al.*, 1994]. The crystal structure of two tripartite efflux pump components, i.e. the Escherichia coli AcrA-AcrB-TolC [Higgins *et al.*, 2004; Koronakis *et al.*, 2000; Murakami *et al.*, 2002] and the *Pseudomonas aeruginosa* MexA-MexB-OprM [Akama *et al.*, 2004a,b; Sennhauser *et al.*, 2009] has been determined, whose analysis led to the proposal of a mechanism of drug transport based on the transition trough three different conformations [Murakami *et al.*, 2006; Seeger *et al.*, 2006]. Although RND proteins have been analyzed in microorganisms belonging to other genera, very little is known in the genus *Burkholderia*, that is known to exhibit multiple antibiotic resistance [Dance *et al.*, 1989; Mahenthiralingam *et al.*, 2005; Thibault *et al.*, 2004]. Indeed, members of RND superfamily have been described for only two species: *B. cenocepacia* and *B. pseudomallei*. In the *B. cenocepacia* J2315 genome, 16 genes encoding putative RND efflux pumps were discovered [Guglierame *et al.*, 2006; Holden *et al.*, 2009]. On the basis of sequence similarity, two of them have been shown to be associated with drug resistance: i) BCAM2550, *ceoB* (ORF10), a component of a system responsible for chloramphenicol, trimethoprim and ciprofloxacin resistance [Burns *et al.*, 1996; Nair *et al.*, 2004]; and ii) BCAS0765 (ORF2) that is associated with resistance to three antibiotics (fluoroquinolones, tetraphenylphosphonium, and streptomycin) as well as to ethidium bromide [Guglierame *et al.*, 2006]. In *B. pseudomallei* K96243 at least 10 operons that may code for RND efflux pump components were disclosed [Kumar *et al.*, 2008]. Although differently annotated, these pumps are conserved in other *B. pseudomallei* strains [Kumar *et al.*, 2008]. Three of these systems have been characterized from a functionally viewpoint: AmrAB-OprA, BpeAB-OprB and BpeEF-OprC. AmrAB-OprA and BpeAB-OprB are pumps that extrude aminoglycoside and macrolide [Chan *et al.*, 2004; Moore *et al.*, 1999], while BpeEF-OprC was shown to efflux trimethoprim and chloramphenicol in a surrogate *P. aeruginosa* strain [Kumar *et al.*, 2006]. Interestingly, the secretion of acyl-homoserine lactones, involved in quorum-sensing systems of *B. pseudomallei*, is dependent absolutely on the function of the BpeAB-OprB [Chan & Chua, 2005; Chan *et al.*, 2007]. Hence, given the clinical/ecological importance of these microorganisms, and the impor-

tance of RND proteins in antibiotic resistance of Gram-negative bacteria, a large-scale bioinformatic analysis was performed aiming to provide a deeper understanding of RND proteins structure/function in *Burkholderia* genus. In particular the aims of this work were: 1) to analyze the phylogenetic distribution of CeoB-like pumps in the *Burkholderia* genus; 2) to define the function(s) they perform within the *Burkholderia* genus and 3) to try tracing the evolutionary history of these genes in *Burkholderia*.

10.2 Methods

10.2.1 Sequence retrieval

Amino acid sequences from the 21 completely sequenced genomes of strains belonging to the genus *Burkholderia*, available on 1st May 2009, were retrieved from GenBank database (Table 10.1). BLAST [Altschul *et al.*, 1997] probing of database was performed with the BLASTP option of this program using default parameters. Only those sequences retrieved at an E-value below the 0.05 threshold were taken into account. 16S rRNA gene nucleotide sequences were retrieved from Ribosomal Database Project (rdp.cme.msu.edu).

10.2.2 Sequence alignment

The ClustalW [Thompson *et al.*, 1994] program in the BioEdit [Hall, 1999] package and the Muscle program [Edgar, 2004] were used to perform pairwise and multiple amino acid sequence alignments. Alignments were manually checked and mis-aligned regions were removed.

10.2.3 Phylogenetic analysis

Neighbor-Joining (NJ) phylogenetic trees were obtained with Mega 4 software [Tamura *et al.*, 2007], complete deletion option and 1000 bootstraps replicates. Maximum Likelihood phylogenetics trees were constructed using Phyml [Guindon & Gascuel, 2003], with a WAG model of amino acid substitution, including a gamma function with 6 categories to take into account differences in evolutionary rates at sites. Statistical support at nodes was obtained by non-parametric bootstrapping on 1000 re-sampled datasets by using Phyml [Guindon & Gascuel, 2003].

10.2.4 Hydropathy plot

Hydropathy plots were obtained on Protscale website (www.expasy.ch/tools/protscale.html) [Gasteiger, 2005] using Kyte and Doolittle scale [Kyte & Doolittle, 1982].

10.2.5 Residues conservation

Analyisis of conservation of amino acid residues was performed using the Weblogo application using default parameters [Crooks *et al.*, 2004].

10.3 Results and Discussion

10.3.1 Analysis of the amino acid sequences of the 16 CeoB-like proteins of *B. cenocepacia* J2315

The existence of 16 CeoB-like coding genes in the genome of *B. cenocepacia* J2315 was previously reported [Guglierame *et al.*, 2006; Holden *et al.*, 2009]. However, a deep analysis of these 16 proteins was not carried out until now. To this purpose, each sequence was firstly scanned for the presence of the four highly conserved motifs shared by RND proteins [Paulsen *et al.*, 1996; Saier *et al.*, 1994], whose consensus sequences are shown in Table 10.1 [Putman *et al.*, 2000]. The analysis of the 16 *B. cenocepacia* J2315 CeoB-like amino acid sequences revealed the existence of the motifs in each of them (see below). In order to assess the conservation of RND proteins structure of each of the 16 sequences, a hydropathy analysis, using the Kyte and Doolittle hydropathicity scale [Kyte & Doolittle, 1982] on ProtScale website [Gasteiger, 2005] (see Material and Methods), was carried out. One of the plots obtained is reported in Figure 10.2; the entire set of 16 plots is presented as Additional File 1. The analysis of each plot and a comparison with the experimentally determined secondary structure of the *E. coli* AcrB and *P. aeruginosa* MexB (not shown), allowed to identify all the 12 TMS and two large loops, that are characteristic of RND proteins.

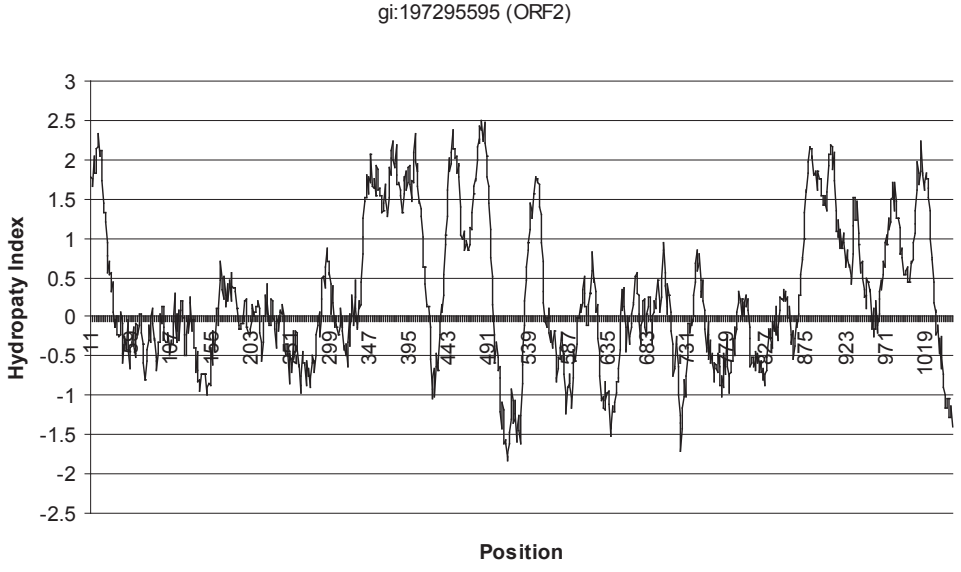


Figure 10.2: Hydropathy plot [39] of *B. cenocepacia* J2315 protein gi:197295595 (ORF2). X axis, position on amino acid sequence; y-axis, hydropathy index.

10.3.2 Organization and phylogenetic analysis of *rnd* genes in *B. cenocepacia* J2315

The analysis of the organization of the *B. cenocepacia* J2315 16 *rnd* genes (Figure 10.3) revealed that in most cases the three genes are organized in a putative operon with three different gene arrays: i) in the first one, shared by ORFs 1-4, 6-10 and 13, the *ceoB* gene is located in between two genes encoding MFP and OMP; ii) in the second array, shared by ORF11 and 12, the *ceoB* gene is located downstream from the other two genes; iii) lastly, in the third one, *ceoB* is located upstream of the other two genes. Besides, in one

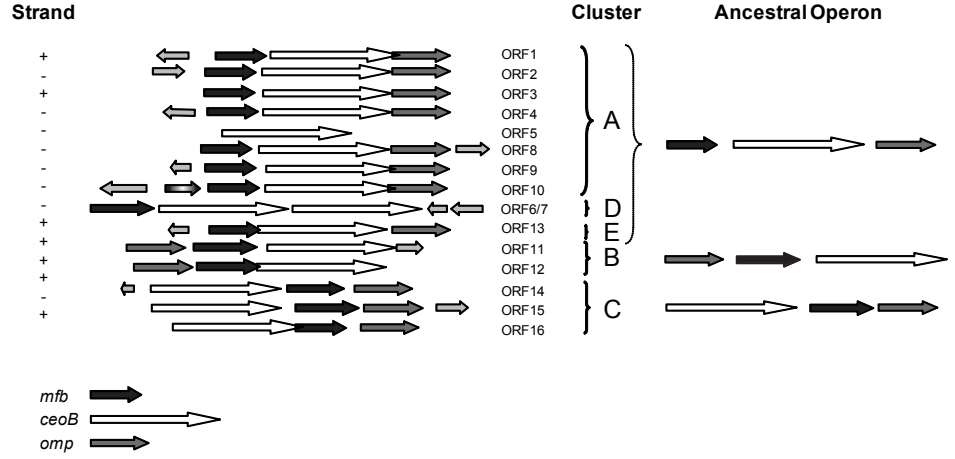


Figure 10.3: Schematic representations of the organization of the 16 gene clusters encoding CeoB-like efflux pumps in *B. cenocepacia* J2315 genome. The organization of the genes identified in *B. cenocepacia* J2315 genome was retrieved from NCBI website. Putative regulatory genes are depicted as light grey arrows. *llpe* gene present only in CeoB operon (ORF10) is depicted as orange arrows.

case (ORF5), the *ceoB*-like sequence is not embedded in a cluster including the other two genes; in another case (ORF6-7), two *ceoB*-like redundant copies were tandemly arranged. In order to analyse the phylogenetic relationships among the 16 CeoB-like proteins their amino acid sequences were aligned using the program ClustalW [Thompson *et al.*, 1994] and the multialignments obtained were used to construct the phylogenetic tree shown in Figure 10.4a. The topology of the tree, which is supported in most cases by very high bootstrap values, revealed that the 16 sequences can be split into five clusters (A, B, C, D, and E). It is worth noting that the overall different gene organization of CeoB, MFP and OMP coding genes corresponds to the subdivisions in the phylogenetic tree in Figure 10.4a. A similar phylogenetic analysis was also performed using the aminoacid sequence of MFP and OMP proteins encoded by genes embedding each operon. Data obtained revealed that the five clusters (A, B, C, D, and E) are easily recognized in the MFP tree (even though the branching order is different) (Figure 10.4b). This is in agreement

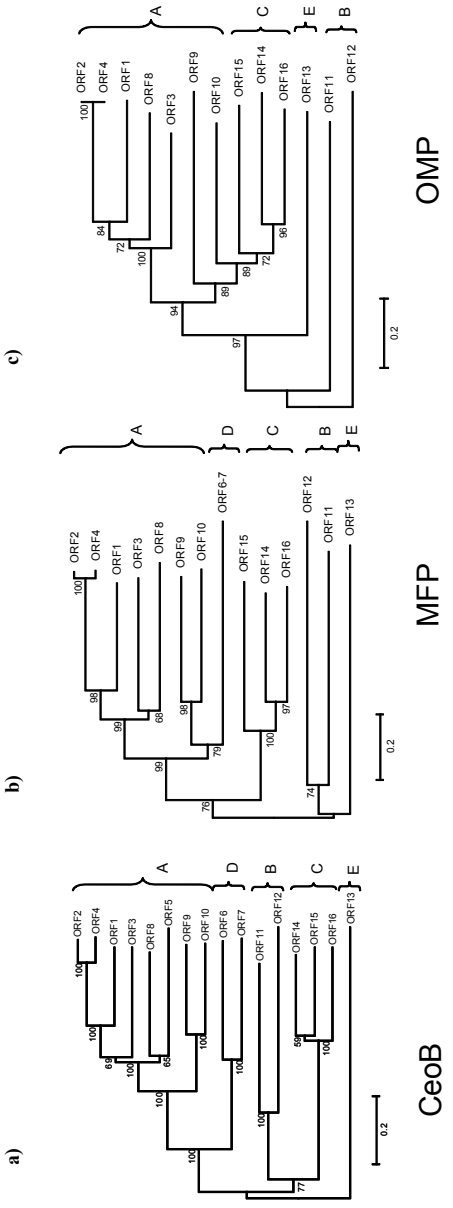


Figure 10.4: *B. cenocepacia* J2315 CeoB-like sequences phylogenetic tree (a), their MFP associated proteins (b) and OMP proteins (c).

with the notion that the two proteins have to interact and often belong to the same transcriptional unit. The topology of OMP tree is slightly different from the other two, in that some of the sequences of cluster A are intermixed with those of cluster C (Figure 10.4c).

10.3.3 Identification and distribution of *ceoB*-like genes in the genus *Burkholderia*

In order to check the distribution of the CeoB-like proteins in the entire genus *Burkholderia*, the *B. cenocepacia* J2315 CeoB amino acid sequence (gi:206564391) was used as a query to probe the 21 completely sequenced genomes of strains belonging to *Burkholderia* genus available at NCBI database (as on 1/05/2009), using default parameters. In this way, a total of 254 sequences homologous to *B. cenocepacia* J2315 CeoB were retrieved. Each sequence was analyzed for the presence of the four highly conserved motifs shared by RND proteins [Paulsen *et al.*, 1996; Saier *et al.*, 1994]. Data obtained (not reported) revealed the existence of the four motifs in all the 254 *Bukholderia* sequences, supporting the idea that they actually are members of RND superfamily. The relative frequency of each amino acid in each position was checked using the WebLogo application (see Material and Methods) (Figure 10.5). In some cases this frequency differed from the consensus

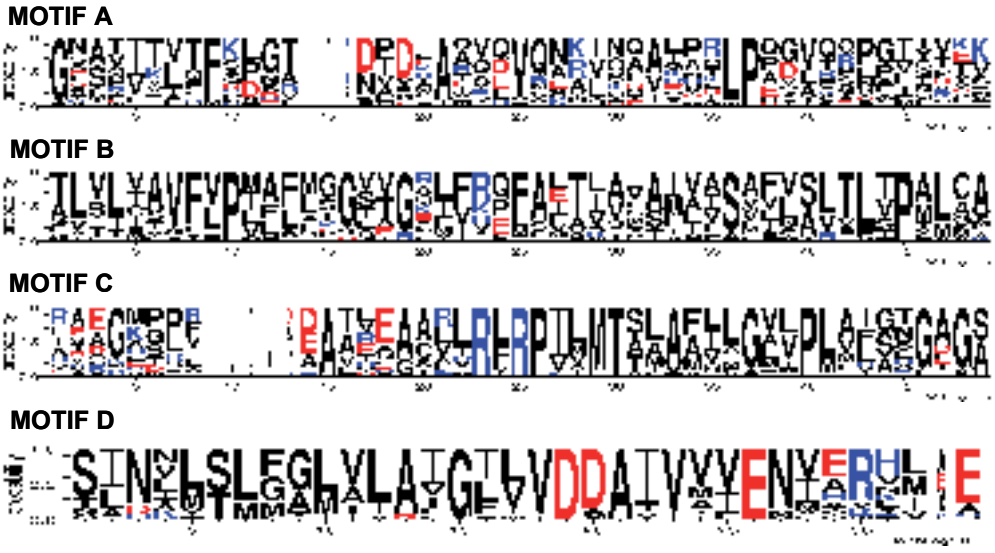


Figure 10.5: WebLogo representation of the four highly conserved motifs shared by *Burkholderia RND* proteins. Amino acids with a positive charge are represented in blue; amino acids with a negative charge are represented in red; amino acids without any charge are represented in black.

sequence(s) previously suggested [Putman *et al.*, 2000]. This is due to the fact that our dataset includes a larger number of sequences in respect to the previous ones [Paulsen *et al.*, 1996; Saier *et al.*, 1994; Tseng *et al.*, 1999]. Hence, we suggested new possible consensus motifs for these sequences (Table 10.1). Figure 10.5 and Tab.10.1 show that Motif A and Motif D represent the less and the most conserved ones, respectively. This is in agreement with the notion that motif A is located on the first periplasmatic loop; many

studies demonstrated that periplasmatic regions of RND proteins are involved in substrate recognition [Eda *et al.*, 2003; Elkins & Nikaido, 2002; Franke *et al.*, 2003; Mao *et al.*, 2002; Middlemiss & Poole, 2004; Tikhonova *et al.*, 2002]. Thus, a higher sequence variability of motif A among various proteins is consistent with the possible recognition of different substrates. On the other hand, part of motif D coincides with TMS4, which is involved in proton translocation [Goldberg *et al.*, 1999; Guan & Nakae, 2001], a function common to all RND proteins. Thus, in principle, this region should exhibit a high degree of sequence conservation. Accordingly, it resulted highly conserved also among proteins transporting different substrates. As shown in Table 10.2, a highly variable number of CeoB-like proteins, ranging from 6 (in *B. mallei* NCTC10247, NCTC10229 and SAVP1) to 18 (in *B. cenocepacia* HI2424 and MC0-3) was found. The 254 *Burkholderia* amino acid sequences retrieved were then aligned using the Muscle program (see Material and Methods) and the multialignments obtained used to construct a phylogenetic tree, schematically reported in Figure 10.6 (see the entire tree in Additional File 2). The analysis of phylogenetic tree revealed that the majority of sequences form clusters including one of the *B. cenocepacia* J2315 sequences (Figure 10.6, black triangles), while other sequences form clusters that do not comprise any *B. cenocepacia* J2315 sequence. In particular, two of these clusters contain sequences from only *B. mallei*, *B. pseudomallei* and *B. thailandiensis* (highlighted with red triangles in Figure 10.6). However, in the whole phylogenetic tree, the sequences can be easily subdivided into the five clusters (A, B, C, D, and E in Figure 10.8) corresponding to the previously identified ones (Figure 10.4a) (although embedding a variable number of sequences).

10.3.4 Functional assignment of the 254 Burkholderia CeoB-like sequences

A preliminary analysis, performed by aligning all the 254 *Burkholderia* sequences with the sequences representative of the five RND families identified in Gram-negative Bacteria and

MOTIF		CONSENSUS SEQUENCES
A	Old	G x s x v T v x F x x g t D x x x A q v q V q n k L q x A x p x L P x x V q x q g x x v x k
	Proposed	G x a x i t x t F x x g t d x d x A x x x V q x x x x x a x x x L P x x v x x p x x x x x x
B	Old	a l v l s a V F l P m a f f g G x t G x i y r q f s i T x v s A m a l S v x v a l t l t P A l c A
	Proposed	t l v l x a V F v P x a f x x G x x G l f v x f A x t x q x q x x x S a x x s l t L t P a L c a
C	Old	x x x G k x l x e A x x x a a x x R L R P I L M T s L a f i l G v l P l a i a t G x A G a
	Proposed	x x x G x x p x x A x x e A a x l R l R P I l M T x l A x x l G x x P L a x x x G x a G s
D	Old	S i N t l T l f g l v l a i G L l v D D A I V v V E N v e R v l a e
	Proposed	s i N x l s L f g l v L A i G i l V D D A I V v v E N v e R h l a E

Table 10.1: Consensus sequences of RND proteins according to Putman at al. [2000]. Table reported the previously individuated and the new proposed consensus sequence. X indicates any amino acid, capital letters show amino acids most frequently observed in a given position in more than 70% of the transport proteins, and lowercase letters represent amino acid occurring in at least than 40% of RND amino acid sequences.

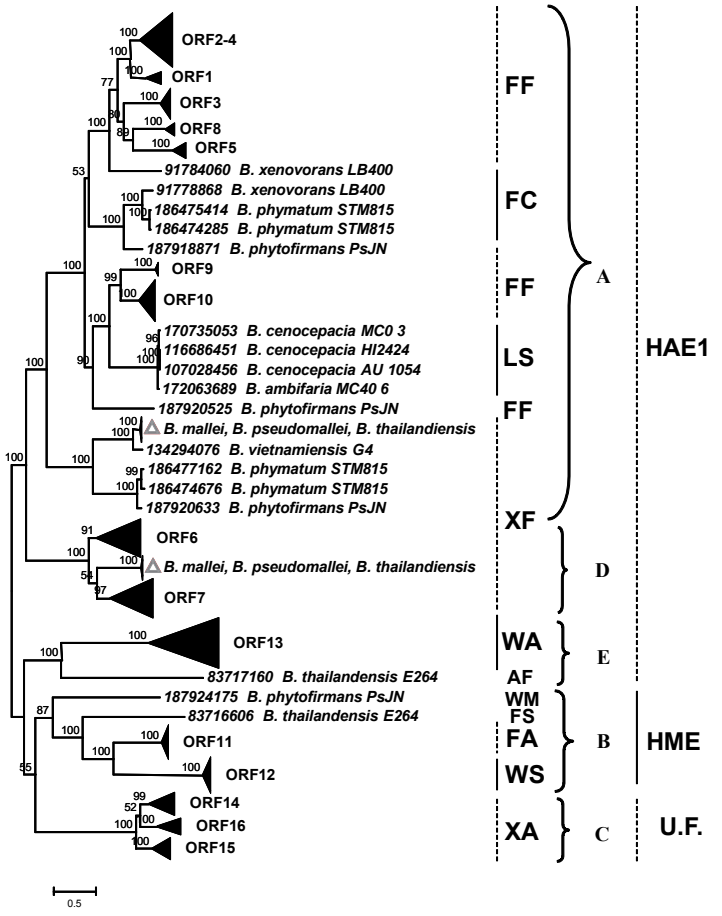


Figure 10.6: Schematic representation of phylogenetic tree constructed using the 254 *Burkholderia* CeoB-like sequence. Sequences that present the same residues in positions corresponding to positions 4-5 of E. coli AcrB are highlited.

experimentally characterized retrieved from Transport Classification Database (TCDB, www.tcdb.org), revealed that most of these sequences could be unambiguously assigned to only two RND families: HAE1 and HME. Indeed, only sequences belonging to these two families shared a significant degree of similarity with the *Burkholderia* sequences, whereas those representative of the other three families (NFE, APPE, SEC DF) resulted highly divergent from the *Burkholderia* ones and could not be reliably aligned (data not shown). To confirm this preliminary assignment and try to determine the substrate of each pump, three different analyses were performed: i) comparison of the 254 CeoB-like sequences with the amino acid sequence of HAE1 and HME experimentally characterized proteins, belonging to other microorganisms; ii) analysis of highly conserved amino acid residues, essential for proton translocation; iii) analysis of residues involved in substrate recognition.

10.3.4.1 Comparison with HAE1 and HME experimentally characterized proteins belonging to other microorganisms

A set of 62 sequences representative of HAE1 and HME families was retrieved from both TCDB and literature (all proteins and their relative substrate are reported in Additional File 3) and aligned with the 254 *Burkholderia* sequences. The multialignment was used to build the phylogenetic tree reported in Additional File 4, and a phylogenetic tree including a subset of these is shown in Figure 10.7, where the five major clusters (A, B, C, D, E) of Figure 10.4 and Figure 10.6 were easily recognized. Three of these clusters (A, D and E, red branches) included characterized proteins belonging to HAE1 family that are known to be involved in antibiotic(s) resistance. Another cluster (in blue) comprised HME proteins (Cluster B in Figure 10.7), involved in heavy metal efflux. Lastly, none of the characterized proteins showed similarity with those grouped in cluster C (pink branches). Indeed, no function could be assigned to these proteins, although they appear to be closer to HME than HAE1 sequences. The analysis of substrate specificity of each characterized protein revealed that HME proteins, transporting different metals, form two distinct clusters. The first one, contained the protein gi:206562298 from *B. cenocepacia* J2315 (ORF 12), which transports monovalent cations (Cu^{+} and Ag^{+}), the other one, included the sequence gi:206562573 from *B. cenocepacia* J2315 (ORF11), transporting divalent cations (Zn^{2+}, Co^{2+}, Cd^{2+} and Ni^{2+}).

10.3.4.2 Analysis of highly conserved amino acid residues essential for proton translocation

It has been proposed that some charged residues in TMS 4 and TMS 10 sequences are essential for proton translocation and pumping function of RND proteins [Goldberg *et al.*, 1999; Guan & Nakae, 2001; Takatsuka & Nikaido, 2006]. Some of these residues are highly conserved in all RND proteins, while others are characteristic of HAE1 and HME families.

10.3.4.3 Residues common to proteins belonging to HAE1 and HME families

The multiple alignment of the amino acid sequences of 39 proteins, belonging to HAE1 and HME families, revealed that the motif G403XXXD407XXXXXXE414 (position referred to *P. aeruginosa* MexB) in TMS 4 is highly conserved in both HAE1 and HME [Guan & Nakae, 2001]. This suggests that these residues may play an important role in proton translocation, a feature shared by all the representatives of the families [Guan & Nakae, 2001]. We checked for the presence of such residues in the 254 *Burkholderia* sequences and all of them were found in each sequence (yellow residues in Figure 10.8a). This finding confirms that these residues are very likely essential for the role performed by these proteins.

10.3.4.4 Residues specific of proteins belonging to the HAE1 family

Five residues were conserved in the HAE1 family [Guan & Nakae, 2001; Takatsuka & Nikaido, 2006]. Two of them, D407-D408 (position referred to *E. coli* AcrB), are located

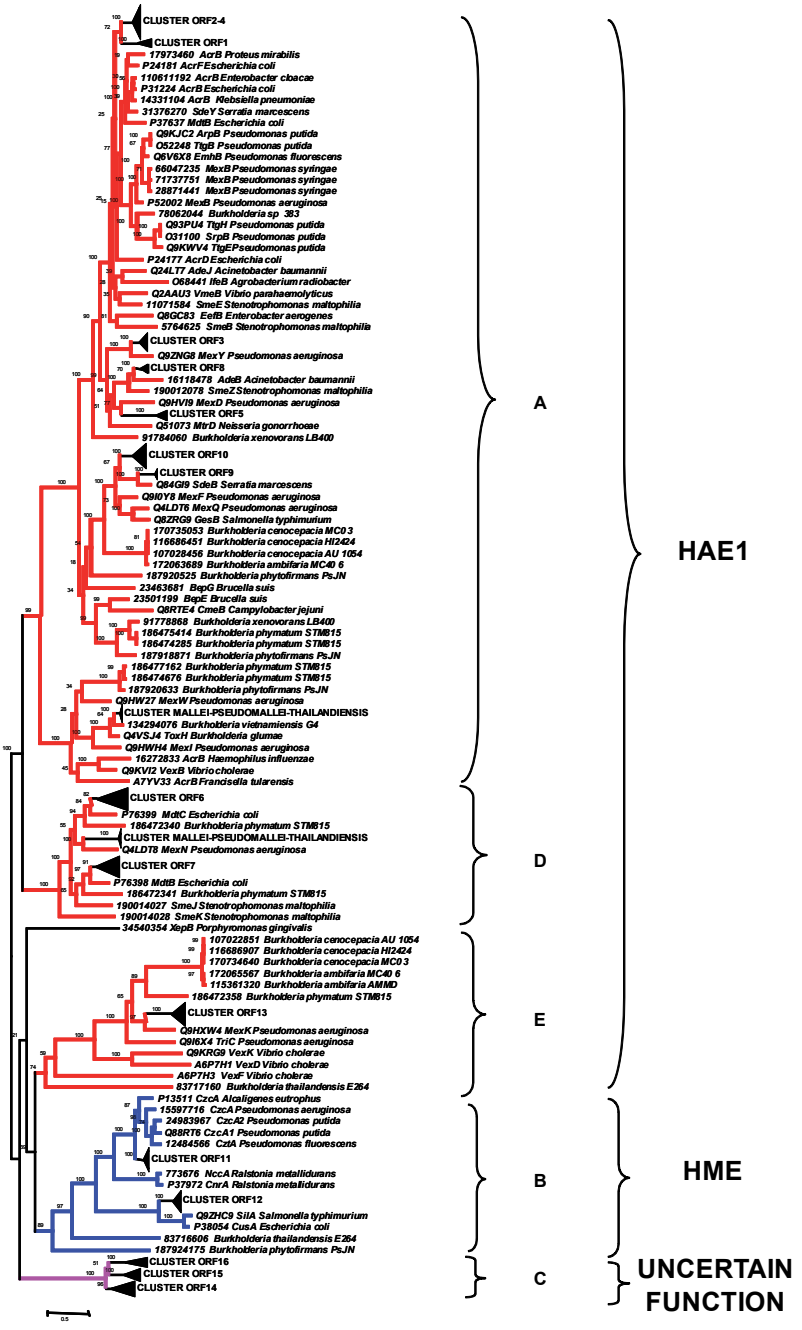


Figure 10.7: Schematic representation of the phylogenetic tree constructed using the 254 *Burkholderia* CeoB-like sequences plus sequences of characterized proteins.

in TMS 4; the other three residues, K940, R971 and T978 (position referred to *E. coli* AcrB), are within or close to TMS 10. As shown in Figure 10.8b, R971 and T978 are conserved in all sequences, suggesting that they may play an important role for both HAE1 and HME proteins. D407 and D408 are conserved, with the exception of four sequences, in all the proteins that, on the basis of phylogenetic tree, were assigned to the

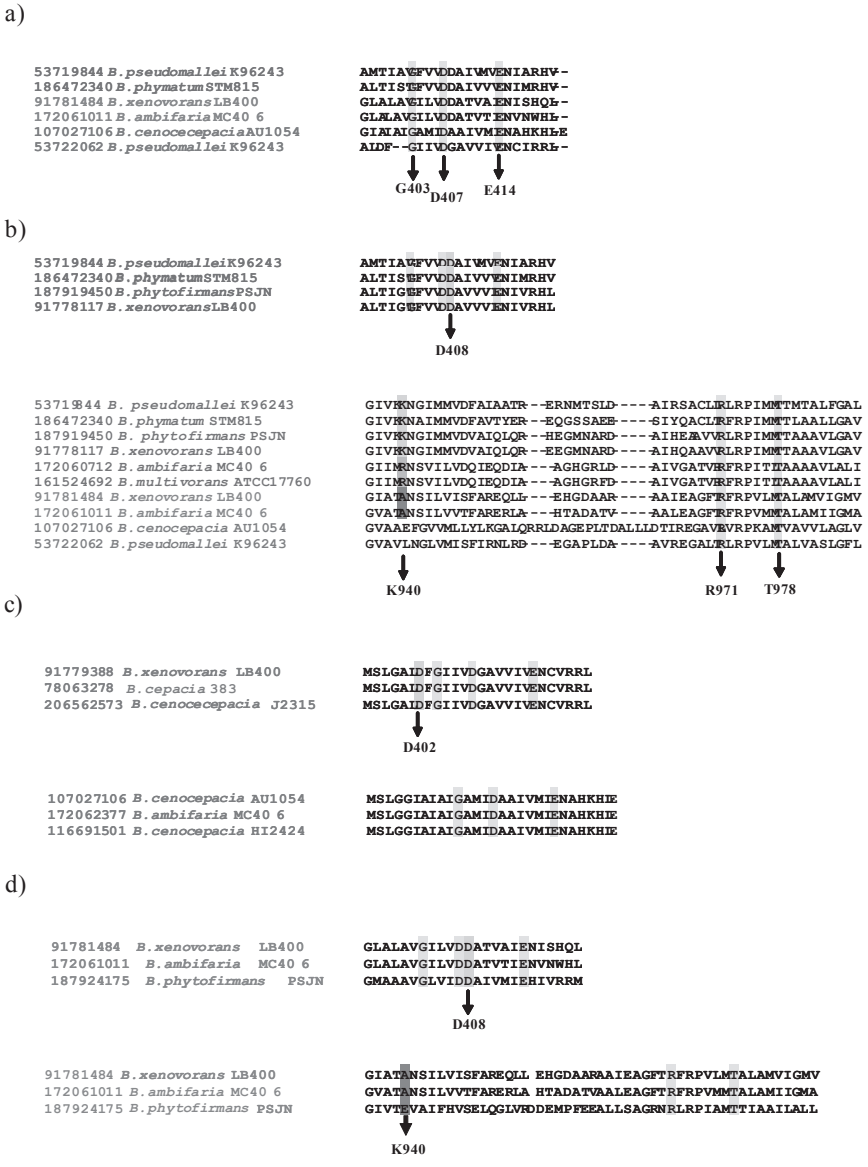


Figure 10.8: Essential residues for proton translocation in RND proteins. Only some representative proteins for each category were reported. Residues conserved among different proteins are highlighted.

HAE1 family. K940 is conserved in all putative HAE1 sequences, with the exception of one cluster (containing ORF 13 from *B.cenocepacia* J2315), where Lysine is replaced by Arginine. However, mutation study in *E. coli* [Guan & Nakae, 2001] suggested that in this particular position the side-chain length is not important, and a positive charge is simply required.

10.3.4.5 Residues specific to proteins belonging to the HME family

HME proteins, that transport divalent cations, e.g. *Ralstonia metallidurans* CzcA, possess an aspartic acid residue at position 402 in TMS 4 (position referred to CzcA of *R. metallidurans)*, in addition to previously identified D407 (that in this kind of proteins is located at position 408). HME proteins that transport monovalent cations, e.g. *E. coli* CusA, present only D408 and miss D402, which may be explained by a $1H^+/Ag^+$ ratio of transport by this system in contrast with a ratio of 2 H^+/1 Zn^{2+} for CzcA-like proteins [Goldberg *et al.*, 1999]. Figure 10.8c shows that proteins previously identified as divalent cation transporters, harbour both aspartic acid residues, whereas proteins identified as monovalent cations transporters, present only one aspartic acid (D408).

10.3.4.6 Residues specific to proteins belonging to Cluster C

Lastly, the sequences with unknown function, present the same residues of HAE1 proteins in TMS 4, but miss K940 that is conserved in all HAE1 proteins (Figure 10.8d). Thus, the analysis of functional residues of RND proteins confirms that sequences identified in *Burkholderia* spp. are RND proteins, and this is in agreement with phylogenetic analysis data. Indeed, putative HAE1 and HME proteins present residues characteristic of each family, and proteins with uncertain function confirm their apparent ambiguous collocation.

10.3.4.7 Analysis of residues involved in substrate recognition

The analysis of the amino acid sequences of *E. coli* AcrB (HAE1) and its homologs allowed to identify conserved residues at their N-terminus, including two phenylalanines (positions 4-5 of *E. coli* AcrB) exposed to the cytoplasm [Das *et al.*, 2007]. Since phenylalanine residues located elsewhere in the protein sequence have been postulated to be involved in ligand binding, Some authors suggested that these conserved residues might be involved in cytoplasmatic substrate recognition [Das *et al.*, 2007]. The analysis of the residues located at these positions in the 254 *Burkholderia* sequences revealed that different clusters exhibited different residues (Figure 10.6a):

1. a large cluster of proteins (Cluster A) shows (with the exceptions of some sequences) two phenylalanines (FF) at both positions;

2. another cluster (Cluster D), previously identified as HAE1, presents a hydrophobic amino acid at the first position and a phenylalanine at the second one (XF); item the third HAE1 cluster (Cluster E) exhibits a tryptophan and an alanine (WA);

3. a putative HME cluster for divalent cations (ORF11) (Cluster B) presents a tryptophan and a serine (WS);

4. a putative HME cluster for monovalent cations (ORF12) (Cluster B) possesses a phenylalanine and an alanine (FA);

5. the sequence cluster with uncertain function presents (Cluster C) a hydrophobic amino acid at the first position and an alanine at the second position (XA).

Hence, the whole body of data presented strongly suggest that the 254 *Burkholderia* sequences are representative of HAE1 and HME families. In particular, HAE1 proteins can be split into three different groups that transport different substrates. HME proteins are divided into two different clusters, one for monovalent and one for divalent cation export, respectively. The third protein cluster can not be assigned to any of the two families.

10.3.5 Interrelationships between number and/or type of CeoB-like proteins and genome dimension, lifestyle, pathogenicity and taxonomic position

The number of CeoB-like proteins of each *Burkholderia* strain was correlated to the genome dimension, the lifestyle, the pathogenicity and the taxonomic position in order to assess the presence of some (eventual) interrelationships. Data of genome dimensions were retrieved from NCBI website (www.ncbi.nlm.nih.gov/genomes/lproks.cgi), while lifestyle and pathogenicity information were obtained from GOLD Genomes on line database (www.genomesonline.org) (Table 10.2). Three different categories were considered for

Species	Strain	Habitat	Patho genicity	Genome size (Mpb)	Number of															
					Chromosomes	Plasmids	CeoB-like Proteins	HAE1 (A)	HAE1 (D)	HAE1 (E)	tot HAE1	ORF 11 (B)	ORF 12 (B)	tot MET	ORF 14 (C)	ORF 15 (C)	ORF 16 (C)	tot UNC	NC
B. ambifaria	AMMD	E	NP	7.57	3	1	11	3	2	2	7		1	1	1	1	1	3	
	MC40-6	E/H	P	7.6	3	1	13	4	2	3	9		1	1	1	1	1	3	
B. cepacia	383	E/H	P	8.69	3		13	8	2	1	11	1		1		1		1	
B. cenocepacia	HI2424	E/H	P	7.76	3	1	18	9	2	2	13	1	1	2	1	1	1	3	
	AU 1054	H	P	7.28	3		17	8	2	2	12	1	1	2	1	1	1	3	
	J2315	H	P	8.07	3	1	16	8	2	1	11	1	1	2	1	1	1	3	
	MC0-3	H	P	7.9	3		18	9	2	2	13	1	1	2	1	1	1	3	
B. mallei	NCTC 10247	H	P	5.9	2		6	3	1		4	1	1	2					
	NCTC 10229	H	P	5.8	2		6	3	1		4	1	1	2					
	SAVP 1	H	P	5.2	2		6	3	1		4	1	1	2					
	ATCC 23344	E/H	P	5.83	2		7	3	1	1	5	1	1	2					
B. pseudomallei	1106a	E	P	7.1	2		10	4	3	1	8	1	1	2					
	668	E	P	7	2		10	4	3	1	8	1	1	2					
	1710b	E	NP	7.31	2		10	4	3	1	8	1	1	2					
	K96243	E	P	7.3	2		10	4	3	1	8	1	1	2					
B. thailandensis	E264	E	NP	6.72	2		11	4	3	1	8	1		1					2
B. vietnamiensis	G4	E	P	8.4	3	5	11	5	2	1	8		1	1	1	1		2	
B. xenovorans	LB400	E	NP	9.8	3		17	6	4	1	11	1		1	2	1	2	5	
B. multivorans	ATCC 17616	H	P	6.99	3	1	12	5	2	1	8	1	2	3		1		1	
B. phymatum	STM815	E/H	NP	8.7	2	2	16	7	4	3	14				1	1		2	
B. phytofirmans	PsJN	E/H	NP	8.22	2	1	16	7	4	1	12				2	1		3	1

Table 10.2: Genome, genome size, habitat, pathogenicity and number RND proteins of each type present in each family of the 21 *Burkholderia* analysed genomes.

lifestyle: strains that live predominantly either in environment (water, soil, rhizosphere etc.), or into a host (plants, animals, humans) and strains that can be found in both environment and host. The average number of proteins in each category is very similar, and standard deviation is very high (2.38 for the first category, 5.53 for the second one and 3.87 for the third one) (Additional File 5). Thus, no apparent relationship between bacterial lifestyle and RND protein number was detected. The same result was obtained

considering each type of CeoB-like proteins (HAE1, HME and uncertain function) (Additional File 5). The relationship with pathogenicity (strains pathogens for plants, animals or humans) was also analysed. Also in this case, no apparent relationship exists with protein number (Additional File 5). The same result was obtained considering each type of CeoB-like proteins (HAE1, HME and uncertain function) (Additional File 5). In spite

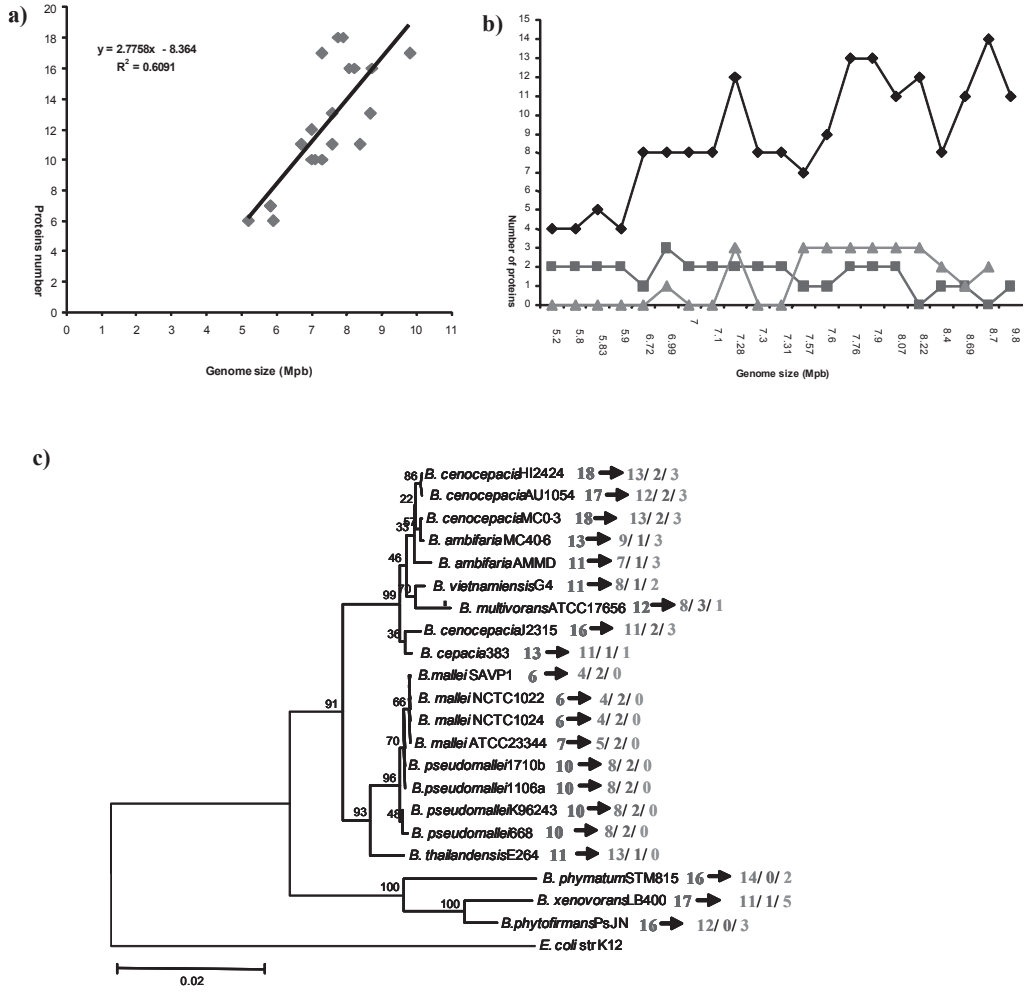


Figure 10.9: Relationship among number of CeoB-like proteins and genome size (a). Relationship among number of CeoB-like proteins for each type and genome size (b) and taxonomy (c).

of the fact that previous studies suggested that the number of multidrug efflux pumps is proportional to the genome size of a given organism [Ren & Paulsen, 2005], data reported in Figure 10.9a revealed that in *Burkholderia* genomes a strict correlation between the two parameters does not exist (R^2=0.6091). However, when the CeoB-like proteins were split into the three categories, the analysis revealed that the number of HME proteins

(blue line in Figure 10.9b, R=0.3787) and of proteins with uncertain function (pink line in Figure 10.9b, R=0.4579), is relatively constant, while the number of HAE1 proteins (red line in Figure 10.9b, R=0.6323) increases in strains with a larger genome (Figure 10.9b). In order to assess relationship with taxonomy, the phylogenetic tree reported in Figure 10.9c was constructed using the16S rRNA gene sequences of each strain; in this tree, the number of proteins for each strain is also reported. A relationship between number of proteins and taxonomy can be found. Indeed strains of the same species and/or strains of related species, possess an identical or very similar number of RND proteins. Thus, the distribution of CeoB-like proteins belonging to the three identified categories [antibiotic transport (HAE1), heavy metal transport (HME) and uncertain function], coded for by each of the 21 *Burkholderia* genomes, was also analyzed. Data obtained are summarized in Table 10.2 and Figure 10.9c showing that: i) proteins with uncertain function are not present in *B. mallei*, *B. pseuomallei* and *B. thailandiensis* strains; ii) proteins belonging to HME family are not present in *B. phymatum* and *B. phytofirmans*; iii) when different strains belonging to the same species possess a different number of RND proteins, this is due to the different number of HAE1 proteins, while proteins with uncertain function and HME maintain the same number in all strains of the same species; iv) HAE1 proteins are the most abundant in all analyzed strains.

10.3.6 Evolution of *rnd* encoding genes in *Burkholderia* genus

The analysis of the distribution of HME and HAE1 like coding sequences in the genus *Burkholderia* revealed a high variability in the copy number among the different species (Table 10.2). Interestingly, all the species branching at the root of the *Burkholderia* reference tree (as assessed by 16S rRNA coding sequences), possess a high number of HME/HAE1-like coding sequences (16 in *B. phymatum* and *B. phytofirmans*, 17 in *B. xenovorans*). Conversely, *B. mallei*, *B. pseudomallei* and *B. thailandiensis* strains possess a lower HAE1/HME copy number, ranging from 6 to 11 in *B. mallei* and *B. thailandiensis* species, respectively. Lastly, the species belonging to the BCC complex and embedded in the upper monophyletic cluster of Figure 10.9c, possess a number of HME/HAE1 copies ranging from 11, in *B. vietnamiensis* G4 and *B. ambifaria* AMMD, to 18 in *B. cenocepacia* representatives. In addition to these data, the phylogenetic tree constructed with all the 254 retrieved sequences of the *Burkholderia* genus (Figure 10.7 and Additional File 4) revealed that the *Burkholderia* species are distributed all over the tree and that the monophyly of the main *Burkholderia* clade (according to the reference phylogeny of Figure 10.9c) is overall respected, suggesting that the *ceoB*-like sequences did not undergo massive HGT events between different *Burkholderia* species or, if this occurred, it happened in the ancestor of *Burkholderia*. However, the possibility that some of these genes might have been exchanged between strains belonging to the same or different *Burkholderia* species and/or between different DNA molecules within the same cytoplasm cannot be a priori excluded. Indeed, it is known that bacteria belonging to this genus harbour two-three different chromosomes and some of them are among the largest genome-sized and most versatile bacteria known. Besides, these genomes harbour a relevant number of genes coding for transposases, integrases, and resolvases, suggesting that they might

frequently undergo DNA rearrangements that, in turn, might alter their gene structure and/or organization [[Papaleo *et al.*, 2009] and references therein]. In addition to this, some of the 21 *Burkholderia* strains harbour one or more large plasmids, which possess genes coding for genetic mobile elements. These elements are responsible for the flow of genes between plasmids and chromosomes inhabiting the same cytoplasm. At the same time, plasmids may also permit the spreading of metabolic traits between cells of the same or different species. Indeed, it has been recognized that HGT is one of the major forces driving the evolution of genes and genomes [Poole, 2009; Vaneechoutte & Fani, 2009]. The analysis of the genomic localization of the 254 *ceoB*-like genes revealed that six of them are located in four different large plasmid molecules harboured by three different strains (Table 10.3). Three of these genes (two from *B. multivorans* and one from *B. vietnamiensis*) felt in Cluster B (corresponding to HME proteins). In the case of *B. multivorans* one of the two sequences has a paralog in the chromosome, thus opening the possibility that the two copies are the result of an internal rearrangement; however, the degree of sequence identity between them is identical (96-97%) to that they shared with the *B. vietnamiensis* sequences. So, it cannot be excluded the possibility that the plasmid-borne genes might have been exchanged between the two strains through plasmid-mediated HGT event(s) occurring recently during evolution. A preliminary comparative analysis of the sequences of these two plasmids revealed that very likely they could have exchanged some regions between each other (Maida et al, manuscript in preparation). The other three sequences are harboured by two *B. phymatum* plasmids. Two of them (both from plasmid pBPHY01) code for proteins (Table 10.3) falling in the group of sequences with uncertain function (Additional File 2) and they do not have any counterpart in the host chromosomes; thus, it is possible that these two sequences might have been moved from the chromosome to pBPHY01. The third one belong to the HAE1 family and the closest paralog in the chromosome share a degree of sequence identity of 94%. The whole body of data suggests a likely evolutionary model accounting for the evolution of the HME and HAE1 systems in *Burkholderia* genus. According to this model, the ancestor of all the extant *Burkholderia* already possessed a high number of HME/HAE1-like gene copies. Although it is not possible to infer the exact copy number of CeoB coding genes in the genome of

Strain	Protein	Plasmid	
		Name	Dimension (bp)
B. multivorans ATTC17616	gi:161506614	pBMUL01	167422
	gi:161506504		
B. phymatum STM815	gi:186471278	pBPHY01	1904893
	gi:186471940		
	gi:186474676	pBPHY02	595102
B. vietnamiensis G4	gi:134287672	pBVIE02	2656616

Table 10.3: *Burkholderia* plasmids harbouring *ceoB*-like genes.

the *Burkholderia* ancestor, it is possible that this number might have been close to the one exhibited by the species embedded in the cluster at the root of the *Burkholderia* reference tree in Figure 10.9c. The high degree of sequence similarity shared by these different copies strongly suggests that they belong to a paralogous gene family, originated from an ancestral *ceoB*-like sequence that underwent many duplication events and existing long before the appearance of the ancestor of *Burkholderia*. On the basis of the available data, it is not possible to infer whether this ancestor gene was organized in operon with an OMP and/or MFP coding gene. However, the finding that most of the *ceoB*-like genes are operonically organized and that (at least) three different operon structures in *B. cenocepacia* J2315 genome, might suggest the existence of three different operon organizations in the genome of the *Burkholderia* ancestor. The possible number of each operon is still unknown and their study is beyond the scope of this work. If this idea is correct, then, starting from this ancestral gene pool, multiple events of gene duplication and gene loss would have led to the copy number patterns of the extant *Burkholderia* representatives. Accordingly, those species possessing the lowest number of HME/HAE1 related sequences (*B. mallei* and *B. pseudomallei* strains) are those for which massive genome reduction (and consequently gene loss) has been documented [Moore *et al.*, 2004]. Regarding the function of ancestral HME/HAE1-like proteins, it is not possible, standing to data presented in this work, to infer whether they were already specialized in recognizing a specific substrate or not. However, it can be mentioned the hypothesis that these ancestral efflux pumps might have been able to recognize different substrates, hence exhibiting low substrate specificity. This is in agreement with the notion that some of the efflux pumps are able to interact with different substrates. Duplication events, followed by evolutionary divergence might have concurred in refining their substrate specificity, allowing them to selectively extrude out of the cell a given chemical compound (antibiotics or heavy-metals). This idea represents a further validation (and an extension) of the "patchwork" hypothesis, originally proposed by Jensen [Jensen, 1976], to explain the origin and evolution of enzymes involved in metabolic pathways (see also [Fani, 2009; Fondi *et al.*, 2009]).

10.4 Conclusion

In this work we have performed a comprehensive comparative analysis of the HME and HAE1 efflux systems in *Burkholderia* genus. A total of 254 coding sequences were retrieved from the available *Burkholderia* genomes and analyzed at different levels, adopting different bioinformatic tools. A deep phylogenetic analysis, in which experimentally characterized sequences were also included, permitted to assign a putative function (i.e. antibiotic resistance, heavy metal efflux) to (up to now) uncharacterized *Burkholderia* sequences. Furthermore, the analysis of conserved residues involved in different functions (substrate recognition, proton translocation) of HME and HAE1 sequences allowed refining peculiar motives previously identified on the basis of a smaller protein dataset. Given the high variability in the number of HAE1 and HME coding sequences found in extant *Burkholderia* species, we tried to correlate both the number and the types (i.e. the transported substrate) with the different characteristics observed in the *Burkholderia* strains

(pathogenic lifestyle, genome size, colonized habitat). However, no apparent correlation emerged, suggesting that other forces might be responsible in determining the types and the copy number of HME/HAE1 sequences in the *Burkholderia* genus. Remarkably, we observed that only HAE1 proteins are mainly responsible for the different number of proteins observed in strains of the same species. By assuming that the physiological role of these proteins is resistance to one or more antibiotics, this finding, in turn, may suggest that the acquisition of antibiotic resistance might be the main selective pressure driving the expansion of this protein family. On the other hand, these proteins might play other roles relevant to the behaviour of Bacteria in their natural ecosystems, so other selective pressure might drive the evolution of this protein family. Data concerning both the distribution and the phylogenetic analysis of the HAE1 and HME in the genus *Burkholderia* allowed depicting a likely evolutionary model accounting for the evolution and spreading of HME and HAE1 systems in *Burkholderia* genus. The occurrence of several species-specific duplication and gene and/or operon loss events finally led to the extant pattern of copy number/type observed in modern-day *Burkholderia*. Lastly, the whole data presented in this work may serve as a basis for future experimental tests, focused especially on HAE1 proteins, aimed at the identification of novel targets in antimicrobial therapy against *Burkholderia* species.

References

Aires, J.R., Kohler, T., Nikaido, H. & Plesiat, P. (1999). Involvement of an active efflux system in the natural resistance of pseudomonas aeruginosa to aminoglycosides. *Antimicrob Agents Chemother*, **43**, 2624–8.

Akama, H., Kanemaki, M., Yoshimura, M., Tsukihara, T., Kashiwagi, T., Yoneyama, H., Narita, S., Nakagawa, A. & Nakae, T. (2004a). Crystal structure of the drug discharge outer membrane protein, oprm, of pseudomonas aeruginosa: dual modes of membrane anchoring and occluded cavity end. *J Biol Chem*, **279**, 52816–9.

Akama, H., Matsuura, T., Kashiwagi, S., Yoneyama, H., Narita, S., Tsukihara, T., Nakagawa, A. & Nakae, T. (2004b). Crystal structure of the membrane fusion protein, mexa, of the multidrug transporter in pseudomonas aeruginosa. *J Biol Chem*, **279**, 25939–42.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.

Burns, J.L., Wadsworth, C.D., Barry, J.J. & Goodall, C.P. (1996). Nucleotide sequence analysis of a gene from *Burkholderia* (pseudomonas) *cepacia* encoding an outer membrane lipoprotein involved in multiple antibiotic resistance. *Antimicrob Agents Chemother*, **40**, 307–13.

Chan, Y.Y. & Chua, K.L. (2005). The *Burkholderia pseudomallei* bpeab-oprb efflux pump: expression and impact on quorum sensing and virulence. *J Bacteriol*, **187**, 4707–19.

Chan, Y.Y., Tan, T.M., Ong, Y.M. & Chua, K.L. (2004). Bpeab-oprb, a multidrug efflux pump in *Burkholderia pseudomallei*. *Antimicrob Agents Chemother*, **48**, 1128–35.

Chan, Y.Y., Bian, H.S., Tan, T.M., Mattmann, M.E., Geske, G.D., Igarashi, J., Hatano, T., Suga, H., Blackwell, H.E. & Chua, K.L. (2007). Control of quorum sensing by a *Burkholderia pseudomallei* multidrug efflux pump. *J Bacteriol*, **189**, 4320–4.

Coenye, T. & Vandamme, P. (2003). Diversity and significance of *Burkholderia* species occupying diverse ecological niches. *Environ Microbiol*, **5**, 719–29.

Compant, S., Nowak, J., Coenye, T., Clement, C. & Ait Barka, E. (2008). Diversity and occurrence of *Burkholderia* spp. in the natural environment. *FEMS Microbiol Rev*, **32**, 607–26.

Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. (2004). Weblogo: a sequence logo generator. *Genome Res*, **14**, 1188–90.

Dance, D.A., Wuthiekanun, V., Chaowagul, W. & White, N.J. (1989). The antimicrobial susceptibility of pseudomonas *pseudomallei* . emergence of resistance in vitro and during treatment. *J Antimicrob Chemother*, **24**, 295–309.

Das, D., Xu, Q.S., Lee, J.Y., Ankoudinova, I., Huang, C., Lou, Y., DeGiovanni, A., Kim, R. & Kim, S.H. (2007). Crystal structure of the multidrug efflux transporter acrb at 3.1a resolution reveals the n-terminal region with conserved amino acids. *J Struct Biol*, **158**, 494–502.

Eda, S., Yoneyama, H. & Nakae, T. (2003). Function of the mexb efflux-transporter divided into two halves. *Biochemistry*, **42**, 7238–44.

Edgar, R.C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–7.

Elkins, C.A. & Nikaido, H. (2002). Substrate specificity of the rnd-type multidrug efflux pumps acrb and acrd of escherichia coli is determined predominantly by two large periplasmic loops. *J Bacteriol*, **184**, 6490–8.

Fani, a.F.M., R. (2009). Origin and evolution of metabolic pathways. *Physics of Life Reviews*, **6**, 23–52.

Fondi, M., Emiliani, G. & Fani, R. (2009). Origin and evolution of operons and metabolic pathways. *Res Microbiol*.

Franke, S., Grass, G., Rensing, C. & Nies, D.H. (2003). Molecular analysis of the copper-transporting efflux system cuscfba of escherichia coli. *J Bacteriol*, **185**, 3804–12.

Gasteiger, H.C.G.A.D.S.W.M.A.R.B.A., E. (2005). Protein identification and analysis tools on the expasy server. *(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press*, pp. 571–607.

Goldberg, M., Pribyl, T., Juhnke, S. & Nies, D.H. (1999). Energetics and topology of czca, a cation/proton antiporter of the resistance-nodulation-cell division protein family. *J Biol Chem*, **274**, 26065–70.

Govan, J.R., Brown, A.R. & Jones, A.M. (2007). Evolving epidemiology of pseudomonas aeruginosa and the *Burkholderia cepacia* complex in cystic fibrosis lung infection. *Future Microbiol*, **2**, 153–64.

Guan, L. & Nakae, T. (2001). Identification of essential charged residues in transmembrane segments of the multidrug transporter mexb of pseudomonas aeruginosa. *J Bacteriol*, **183**, 1734–9.

Guglierame, P., Pasca, M.R., De Rossi, E., Buroni, S., Arrigo, P., Manina, G. & Riccardi, G. (2006). Efflux pump genes of the resistance-nodulation-division family in *Burkholderia* ceno *cepacia* genome. *BMC Microbiol*, **6**, 66.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**, 696–704.

Hall, T.A. (1999). Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucl Acids Symp Ser*, **41**, 95–98.

Higgins, M.K., Bokma, E., Koronakis, E., Hughes, C. & Koronakis, V. (2004). Structure of the periplasmic component of a bacterial drug efflux pump. *Proc Natl Acad Sci U S A*, **101**, 9994–9.

Holden, M.T., Seth-Smith, H.M., Crossman, L.C., Sebaihia, M., Bentley, S.D., Cerdeno-Tarraga, A.M., Thomson, N.R., Bason, N., Quail, M.A., Sharp, S., Cherevach, I., Churcher, C., Goodhead, I., Hauser, H., Holroyd, N., Mungall, K., Scott, P., Walker, D., White, B., Rose, H., Iversen, P., Mil-Homens, D., Rocha, E.P., Fialho, A.M., Baldwin, A., Dowson, C., Barrell, B.G., Govan, J.R., Vandamme, P., Hart, C.A., Mahenthiralingam, E. & Parkhill, J. (2009). The genome of *Burkholderia* ceno *cepacia* j2315, an epidemic pathogen of cystic fibrosis patients. *J Bacteriol*, **191**, 261–77.

Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, **30**, 409–25.

Koronakis, V., Sharff, A., Koronakis, E., Luisi, B. & Hughes, C. (2000). Crystal structure of the bacterial membrane protein tolc central to multidrug efflux and protein export. *Nature*, **405**, 914–9.

Kumar, A., Chua, K.L. & Schweizer, H.P. (2006). Method for regulated expression of single-copy efflux pump genes in a surrogate pseudomonas aeruginosa strain: identification of the bpeef-oprc chloramphenicol and trimethoprim efflux pump of *Burkholderia pseudomallei* 1026b. *Antimicrob Agents Chemother*, **50**, 3460–3.

Kumar, A., Mayo, M., Trunck, L.A., Cheng, A.C., Currie, B.J. & Schweizer, H.P. (2008). Expression of resistance-nodulation-cell-division efflux pumps in commonly used *Burkholderia pseudomallei* strains and clinical isolates from northern australia. *Trans R Soc Trop Med Hyg*, **102 Suppl 1**, S145–51.

Kyte, J. & Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**, 105–32.

Ma, D., Cook, D.N., Alberti, M., Pon, N.G., Nikaido, H. & Hearst, J.E. (1993). Molecular cloning and characterization of acra and acre genes of escherichia coli. *J Bacteriol*, **175**, 6299–313.

Mahenthiralingam, E., Urban, T.A. & Goldberg, J.B. (2005). The multifarious, multireplicon *Burkholderia cepacia* complex. *Nat Rev Microbiol*, **3**, 144–56.

Mao, W., Warren, M.S., Black, D.S., Satou, T., Murata, T., Nishino, T., Gotoh, N. & Lomovskaya, O. (2002). On the mechanism of substrate specificity by resistance nodulation division (rnd)-type multidrug resistance pumps: the large periplasmic loops of mexd from pseudomonas aeruginosa are involved in substrate recognition. *Mol Microbiol*, **46**, 889–901.

Martinez, J.L., Sanchez, M.B., Martinez-Solano, L., Hernandez, A., Garmendia, L., Fajardo, A. & Alvarez-Ortega, C. (2009). Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems. *FEMS Microbiol Rev*, **33**, 430–49.

Middlemiss, J.K. & Poole, K. (2004). Differential impact of mexb mutations on substrate selectivity of the mexab-oprm multidrug efflux pump of pseudomonas aeruginosa. *J Bacteriol*, **186**, 1258–69.

Moore, R.A., DeShazer, D., Reckseidler, S., Weissman, A. & Woods, D.E. (1999). Efflux-mediated aminoglycoside and macrolide resistance in *Burkholderia pseudomallei*. *Antimicrob Agents Chemother*, **43**, 465–70.

Moore, R.A., Reckseidler-Zenteno, S., Kim, H., Nierman, W., Yu, Y., Tuanyok, A., Warawa, J., DeShazer, D. & Woods, D.E. (2004). Contribution of gene loss to the pathogenic evolution of *Burkholderia pseudomallei* and *Burkholderia* mallei. *Infect Immun*, **72**, 4172–87.

Murakami, S., Nakashima, R., Yamashita, E. & Yamaguchi, A. (2002). Crystal structure of bacterial multidrug efflux transporter acrb. *Nature*, **419**, 587–93.

Murakami, S., Nakashima, R., Yamashita, E., Matsumoto, T. & Yamaguchi, A. (2006). Crystal structures of a multidrug transporter reveal a functionally rotating mechanism. *Nature*, **443**, 173–9.

Nair, B.M., Cheung, J., K. J., Griffith, A. & Burns, J.L. (2004). Salicylate induces an antibiotic efflux pump in *Burkholderia cepacia* complex genomovar iii (b. ceno *cepacia*). *J Clin Invest*, **113**, 464–73.

Nikaido, H. (2001). Preventing drug access to targets: cell surface permeability barriers and active efflux in bacteria. *Semin Cell Dev Biol*, **12**, 215–23.

Papaleo, M.C., Russo, E., Fondi, M., Emiliani, G., Frandi, A., Brilli, M., Pastorelli, R. & Fani, R. (2009). Structural, evolutionary and genetic analysis of the histidine biosynthetic "core" in the genus *Burkholderia*. *Gene*.

Paulsen, I.T., Brown, M.H. & Skurray, R.A. (1996). Proton-dependent multidrug efflux systems. *Microbiol Rev*, **60**, 575–608.

Poole, A.M. (2009). Horizontal gene transfer and the earliest stages of the evolution of life. *Res Microbiol*.

Poole, K. (2007). Efflux pumps as antimicrobial resistance mechanisms. *Ann Med*, **39**, 162–76.

Poole, K., Krebes, K., McNally, C. & Neshat, S. (1993). Multiple antibiotic resistance in pseudomonas aeruginosa: evidence for involvement of an efflux operon. *J Bacteriol*, **175**, 7363–72.

Putman, M., van Veen, H.W. & Konings, W.N. (2000). Molecular properties of bacterial multidrug transporters. *Microbiol Mol Biol Rev*, **64**, 672–93.

Ren, Q. & Paulsen, I.T. (2005). Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLoS Comput Biol*, **1**, e27.

Ryan, B.M., Dougherty, T.J., Beaulieu, D., Chuang, J., Dougherty, B.A. & Barrett, J.F. (2001). Efflux in bacteria: what do we really know about it? *Expert Opin Investig Drugs*, **10**, 1409–22.

Saier, J., M. H. & Paulsen, I.T. (2001). Phylogeny of multidrug transporters. *Semin Cell Dev Biol*, **12**, 205–13.

Saier, J., M. H., Tam, R., Reizer, A. & Reizer, J. (1994). Two novel families of bacterial membrane proteins concerned with nodulation, cell division and transport. *Mol Microbiol*, **11**, 841–7.

Seeger, M.A., Schiefner, A., Eicher, T., Verrey, F., Diederichs, K. & Pos, K.M. (2006). Structural asymmetry of acrb trimer suggests a peristaltic pump mechanism. *Science*, **313**, 1295–8.

Sennhauser, G., Bukowska, M.A., Briand, C. & Grutter, M.G. (2009). Crystal structure of the multidrug exporter mexb from pseudomonas aeruginosa. *J Mol Biol*, **389**, 134–45.

Takatsuka, Y. & Nikaido, H. (2006). Threonine-978 in the transmembrane segment of the multidrug efflux pump acrb of escherichia coli is crucial for drug transport as a probable component of the proton relay network. *J Bacteriol*, **188**, 7284–9.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). Mega4: Molecular evolutionary genetics analysis (mega) software version 4.0. *Mol Biol Evol*, **24**, 1596–9.

Thibault, F.M., Hernandez, E., Vidal, D.R., Girardet, M. & Cavallo, J.D. (2004). Antibiotic susceptibility of 65 isolates of *Burkholderia pseudomallei* and *Burkholderia* mallei to 35 antimicrobial agents. *J Antimicrob Chemother*, **54**, 1134–8.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–80.

Tikhonova, E.B., Wang, Q. & Zgurskaya, H.I. (2002). Chimeric analysis of the multicomponent multidrug efflux transporters from gram-negative bacteria. *J Bacteriol*, **184**, 6499–507.

Tseng, T.T., Gratwick, K.S., Kollman, J., Park, D., Nies, D.H., Goffeau, A. & Saier, J., M. H. (1999). The rnd permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins. *J Mol Microbiol Biotechnol*, **1**, 107–25.

Vaneechoutte, M. & Fani, R. (2009). From the primordial soup to the latest universal common ancestor. *Res Microbiol*.

Vanlaere, E., Lipuma, J.J., Baldwin, A., Henry, D., De Brandt, E., Mahenthiralingam, E., Speert, D., Dowson, C. & Vandamme, P. (2008). *Burkholderia* latens sp. nov., *Burkholderia* diffusa sp. nov., *Burkholderia* arboris sp. nov., *Burkholderia* seminalis sp. nov. and *Burkholderia* metallica sp. nov., novel species within the *Burkholderia cepacia* complex. *Int J Syst Evol Microbiol*, **58**, 1580–90.

Vanlaere, E., Baldwin, A., Gevers, D., Henry, D., De Brandt, E., LiPuma, J.J., Mahenthiralingam, E., Speert, D.P., Dowson, C. & Vandamme, P. (2009). Taxon k, a complex within the *Burkholderia cepacia* complex, comprises at least two novel species, *Burkholderia* contaminans sp. nov. and *Burkholderia* lata sp. nov. *Int J Syst Evol Microbiol*, **59**, 102–11.

Chapter 11

Conclusions and Future Perspectives

The importance of determining the entire genome sequences from all the major domains of life was recognized more than two decades ago and was an important first step in ushering the field of genomics. With commercially available 454 pyrosequencing (followed by Illumina, SOLiD, and now Helicos), there has been an explosion of ('draft') genomes sequenced; however, these can be very poor quality genomes (due to inherent errors in the sequencing technologies, and the inability of assembly programs to fully address these errors, revealing the necessity, in the next future, to strengthen the comparative genomics approach devoted to the improvement of both the assembly and the assignment of new genome sequence data. Nevertheless, we are now leaving the so-called genomic era and we are on our way to post-genomic era. Probably in a few years the sequencing of a (small) genome will be little more than a routine laboratory tecnique and an exponentially increasing amount of completely sequenced genomes will be available in public databases. This will immediately rise the question on how to store, update and (more interestingly) interpret all the (sometimes hidden) information that genomes harbor. These issues will probably require much more effort and, consequently, the post-genomic era can be expected to last much longer than genomic one did, probably extending over several generations. Bioinformatics, that is the interdisciplinary field that blends computer science and biostatistics with biological and biomedical sciences, is expected to gain a central role in next feature and will probably play a crucial role also when planning future wet-lab experiments. Bioinformatics, indeed, has now affected several fields of biology, as the results presented in this dissertation have partially shown. In fact, the analysis of sequence data can be used in different fields, such as evolution (e.g. the assembly and evolution of metabolism), infections control (e.g. the horizontal flow of antibiotic resistance), ecology (bacterial bioremediation). Finally, it can be anticipated that the understanding of the main biological systems (including their evolutionary dynamics) that we will acquire in the next years will be strictly connected to the correct design and use of computational tools. It is through them that we will try to integrate and give a biological meaning to all the exponentially increasing amount of experimental data that will be released.