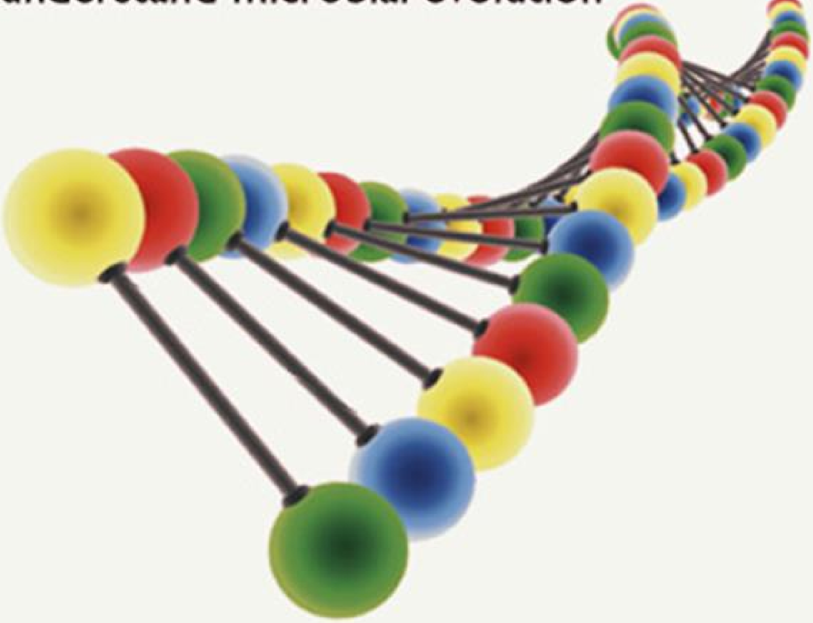
Scientifica
2010

MARCO FONDI

Bioinformatics of genome evolution

From ancestral to modern metabolism Phylogenomics and comparative genomics to understand microbial evolution



PREMIO TESI DOTTORATO
FIRENZE UNIVERSITY PRESS – UNIVERSITÀ DEGLI STUDI DI FIRENZE

FIRENZE
UNIVERSITY
PRESS

PREMIO FIRENZE UNIVERSITY PRESS
TESI DI DOTTORATO

– 21 –

COLLANA PREMIO TESI DI DOTTORATO
Commissione giudicatrice, anno 2010

Marco Fondi

Bioinformatics of genome evolution: from ancestral to modern metabolism

Phylogenomics and comparative genomics
to understand microbial evolution

Firenze University Press
2011

Bioinformatics of genome evolution : from ancestral to modern metabolism : phylogenomics and comparative genomics to understand microbial evolution / Marco Fondi. – Firenze : Firenze University Press, 2011.
(Premio FUP. Tesi di dottorato ; 21)

http://digital.casalini.it/9788866550457

ISBN 978-88-6655-043-3 (print)
ISBN 978-88-6655-045-7 (online)

Immagine di copertina: © Kazakov Alexey | Dreamstime.com

Università degli Studi di Firenze
Firenze University Press
Borgo Albizi, 28
50122 Firenze, Italy
http://www.fupress.com/

Printed in Italy

Contents

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

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 earlyheterotrophic 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 progresses 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

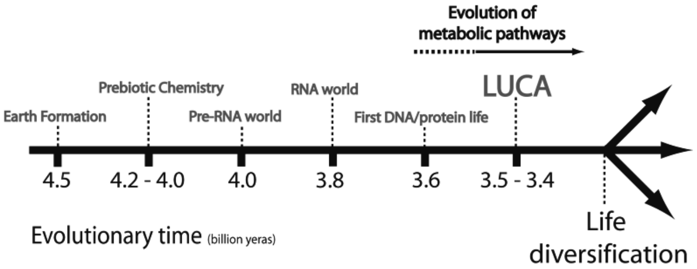


Figure 1.1: Evolutionary time line from the origin of Earth to the diversification of life.

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

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).

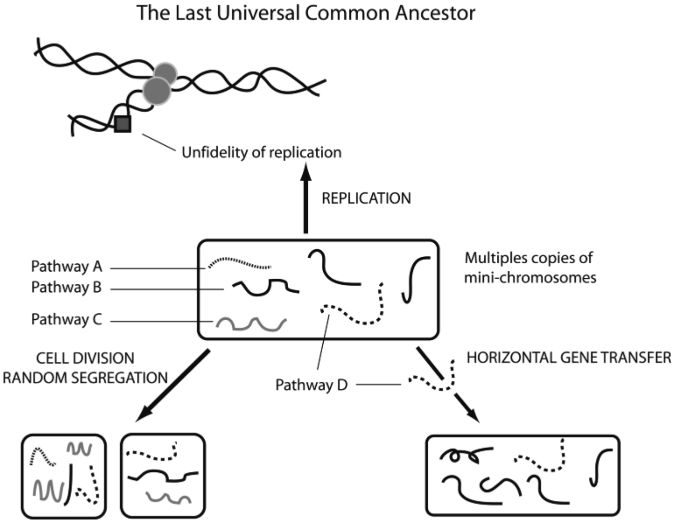


Figure 1.2: Representation of a possible progenotes community

This community evolved into a smallernumber 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 lineagesthat 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).

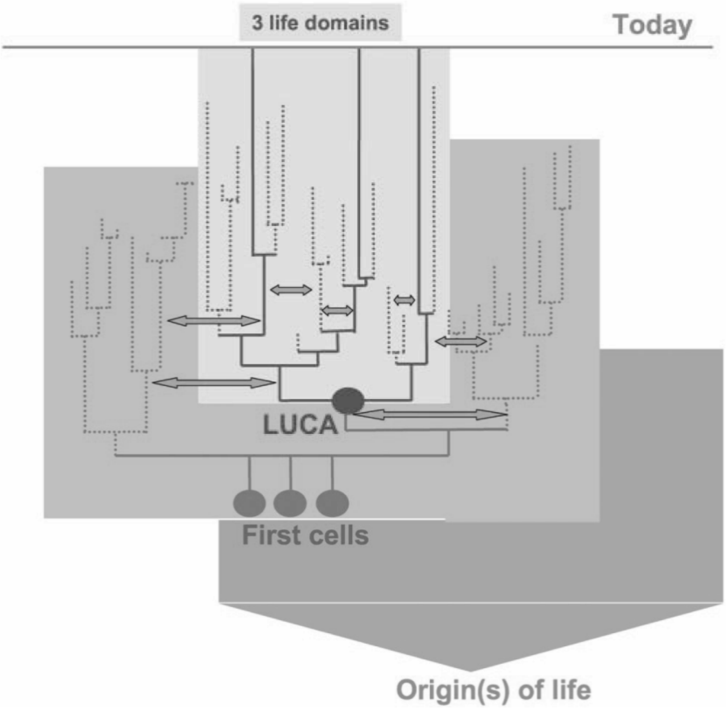


Figure 1.3: LUCA was the last bottleneck in a long series of ancestors to the three present- day cellular domains: Archaea, Bacteria, and Eukarya (from (Forterre & Gribaldo, 2007)).

According to this view, contemporarygenomes 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, stillshared 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 relativecontent 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 preexisting 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 start-

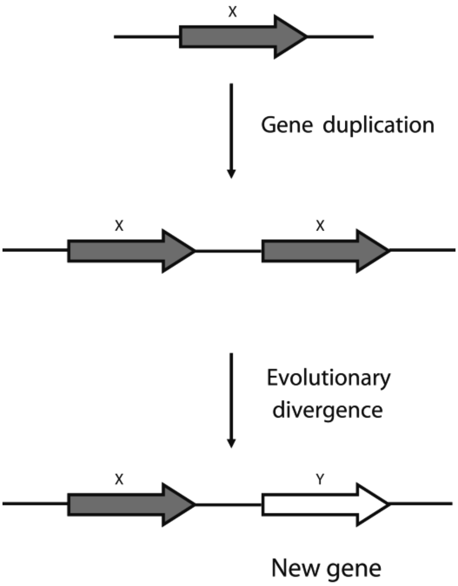


Figure 1.4: Gene duplication.

er 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 105/103, 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).

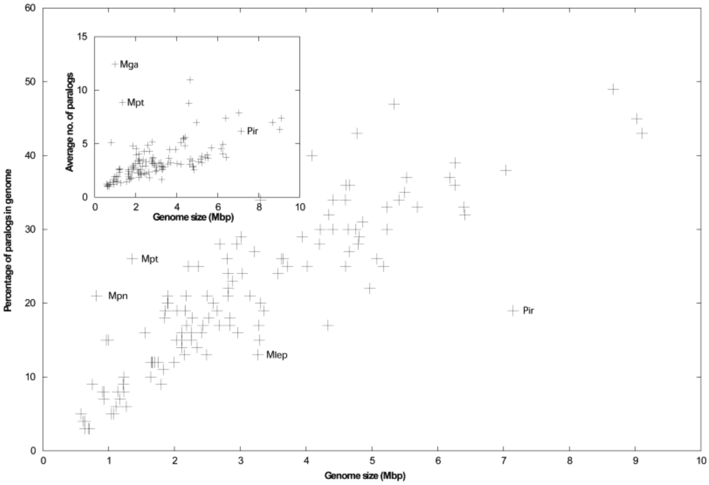


Figure 1.5: Relationship between percentage of genes belonging to paralogous families plotted versus genome size in 127 bacterial genomes

Thus, therate 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.

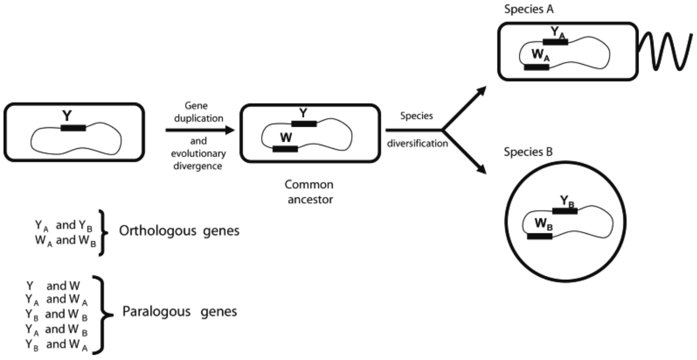


Figure 1.6: Orthologous and paralogous genes.

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).

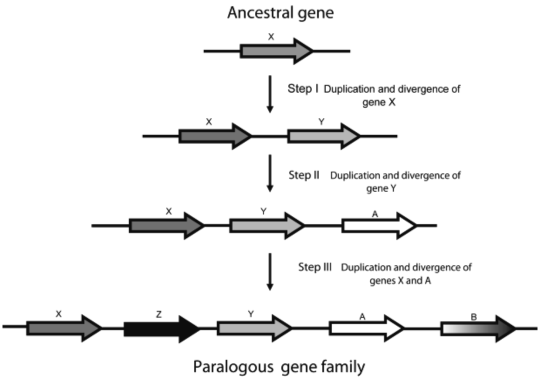

Ancestral gene
X
Step I Duplication and divergence of gene X
X
Y
Step II Duplication and divergence of gene Y
X
Y
A
Step III Duplication and divergence of genes X and A
X
Z
Y
A
B
Paralogous gene family

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 surmised, in fact, that, after a gene duplicates, one of the two copies becomes dispensable and can undergo several types of mutational events, mainly substitu-

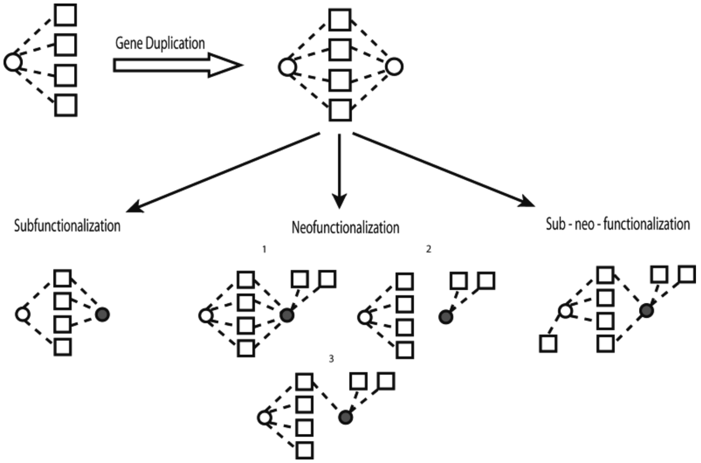


Figure 1.8: Evolutionary models of functional divergence between duplicate genes. Genes and the function(s) they code are represented with circles and squares, respectively. Dotted lines link genes with their functions.

tions, 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).

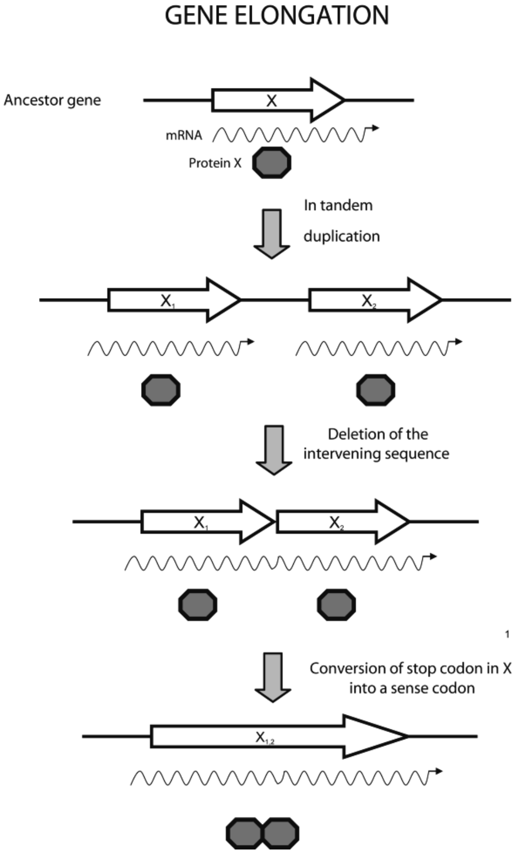


Figure 1.9: Representation of a gene elongation event.

This type of duplication has 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, previousextensive analyses (Reizer & Saier, 1997) have illustrated the modular assembly and de- sign 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 centers 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 colocalized 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 coordinate 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 define 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 elucida- tion 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 evolutionaryforces 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 thecatabolism 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,

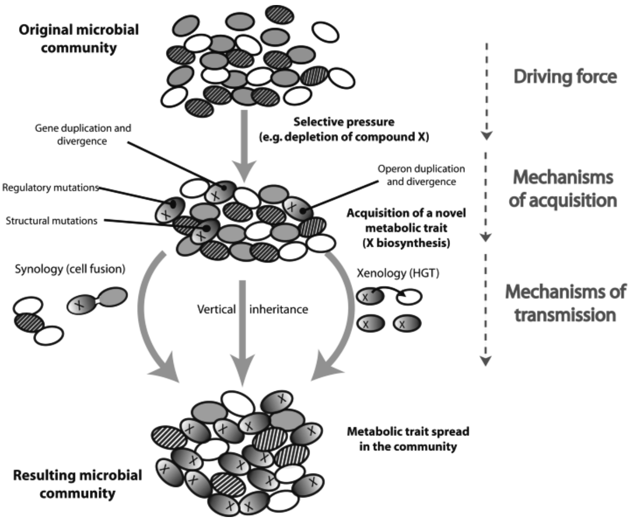

Original microbial community
Driving force
Selective pressure (e.g. depletion of compound X)
Gene duplication and divergence
Regulatory mutations
Structural mutations
Operon duplication and divergence
Mechanisms of acquisition
Acquisition of a novel metabolic trait (X biosynthesis)
Synology (cell fusion)
Xenology (HGT)
Vertical inheritance
Mechanisms of transmission
Metabolic trait spread in the community
Resulting microbial community

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))

because it rendered the primordial cells less dependent on theexternal 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 inter- connected (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 question, 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.

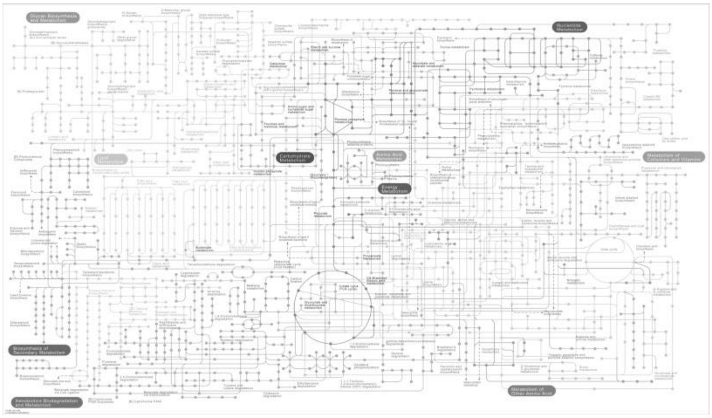

Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Lipid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolism
Metabolism of Other Amino Acid
Xenobiotics Biodegradation and Metabolism

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

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 essentialfor 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

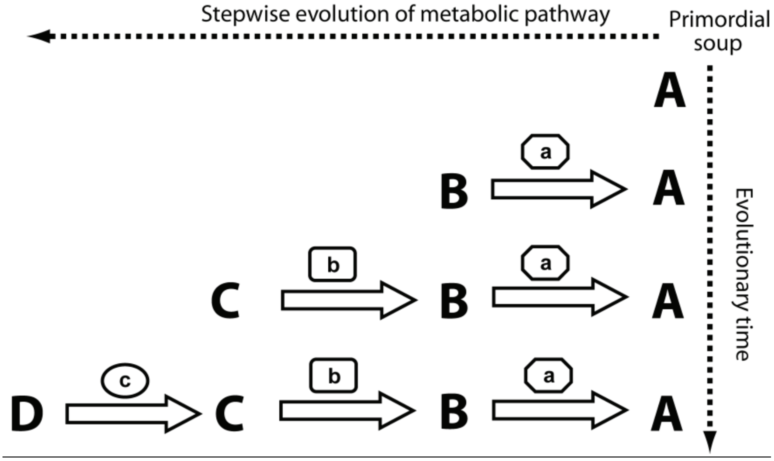


Figure 1.12: Schematic representation of the Horowitz hypothesis on the origin and evolution of metabolic pathways (from (Fondi et al., 2009)).

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 ofmultiple 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 cellulargenomes. 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 heterotrophic 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 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

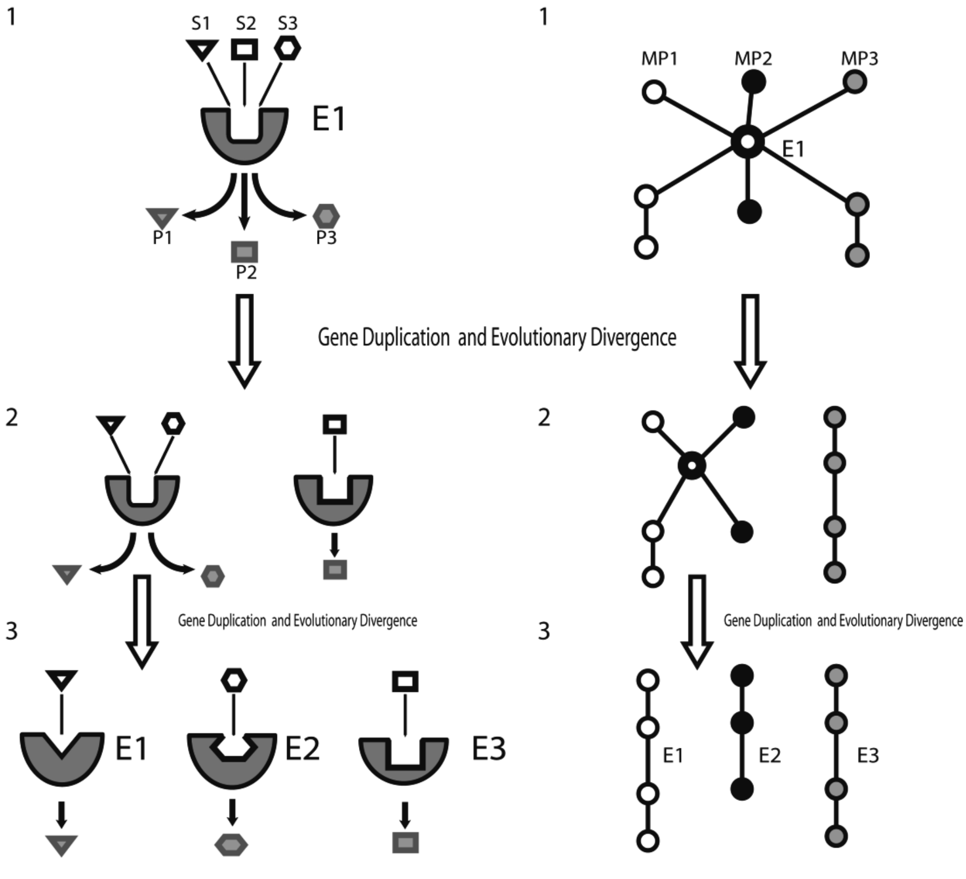

1
S1
S2
S3
E1
P1
P2
P3
1
MP1
MP2
MP3
E1
Gene Duplication and Evolutionary Divergence
2
2
Gene Duplication and Evolutionary Divergence
3
Gene Duplication and Evolutionary Divergence
3
E1
E2
E3
E1
E2
E3

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

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 Miller (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. tryptophan (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 chal- lenging 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 related 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 transcription translation 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).

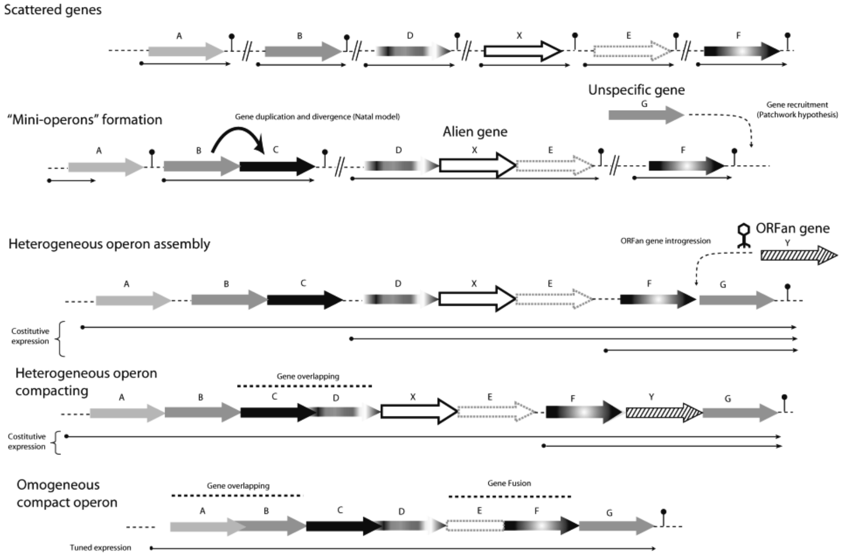


Figure 1.14: Schematic representation of the ”piece-wise” model for operon assembly (from (Fondi et al., 2009)).

Accordingto 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).

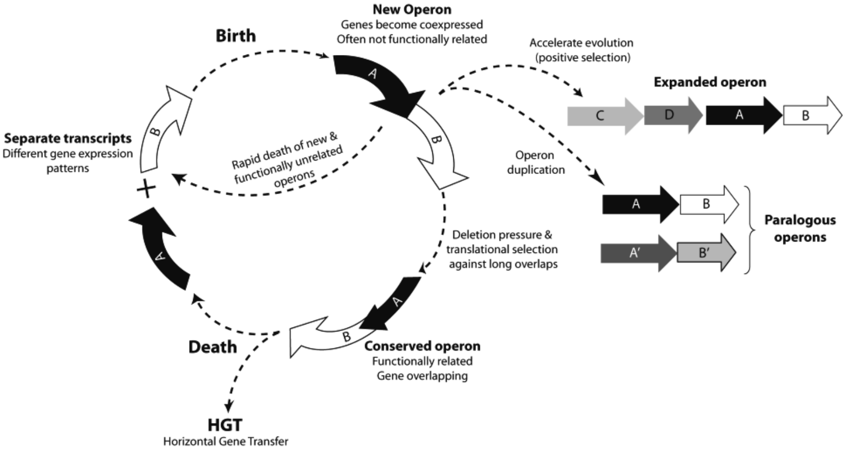


Figure 1.15: The life cycle of operon (from (Fondi et al., 2009), modified from (Price et al., 2006)).

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.

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 bioinformatic tools allowed the storage and the interpretation of several sources of information (gene structure and organization, gene regulation, protein-protein 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 phylgenomics 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 specific databases (including KEGG, GeneOntology, RefSeq).

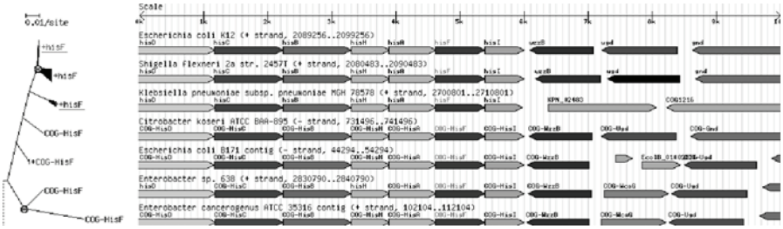

0.01/site
+hisF
+hisF
+hisF
COG-HisF
+COG-HisF
COG-HisF
COG-HisF
Scale
0k
1k
2k
3k
4k
5k
6k
7k
8k
9k
Escherichia coli K12 (+ strand, 2089256..2099256)
hisD hisC hisB hisH hisA hisF hisI wzzB ugd gnd
Shigella flexneri 2a str. 2457T (+ strand, 2080483..2090483)
hisD hisC hisB hisH hisA hisF hisI wzzB ugd gnd
Klebsiella pneumoniae subsp. pneumoniae MGH 78578 (+ strand, 2700801..2710801)
hisD hisC hisB hisH hisA hisF hisI KPN_02483 COG1216
Citrobacter koseri ATCC BAA-895 (- strand, 731496..741496)
COG-HisD COG-HisC COG-HisB COG-HisH COG-HisA COG-HisF COG-HisI COG-WzzB COG-Ugd COG-Gnd
Escherichia coli B171 contig (- strand, 44294..54294)
COG-HisD COG-HisC COG-HisB COG-HisH COG-HisA COG-HisF COG-HisI COG-WzzB COG-Ugd
Enterobacter sp. 638 (+ strand, 2830790..2840790)
hisD COG-HisC COG-HisB hisH COG-HisA COG-HisF COG-HisI COG-WzzB COG-WcaG COG-Ugd
Enterobacter cancerogenus ATCC 35316 contig (+ strand, 102104..112104)
COG-HisD COG-HisC COG-HisB COG-HisH COG-HisA COG-HisF COG-HisI COG-WzzB COG-WcaG COG-Ugd

Figure 1.16: Output of MicrobesOnLine webserver when probed with ”HisF” text search.

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 1.16). 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) (Figure 1.17).

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

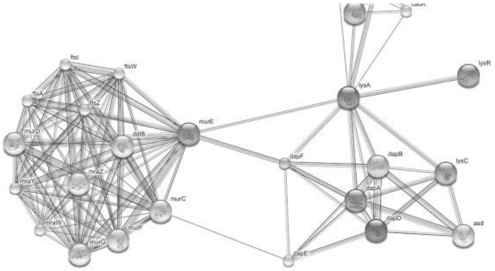


Figure 1.17: Output of String webserver when probed with ”LysA” text search.

shared 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 orthologidentification 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 retrieval 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), MUS- CLE (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 (adopt- ing 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 implementingthis 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 analyzed, 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 inter- action, 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 1.18 some examples of major biological systems that recently have been analyzed taking advantage of graph theory are proposed . These large-scale analyses arebeginning 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 poorlyconnected ones (Figure 1.19). Recently, scale-free behaviors have been found in manybiological networks, including nervous systems (Watts & Strogatz, 1998), metabolic net- works (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 knock- out 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).

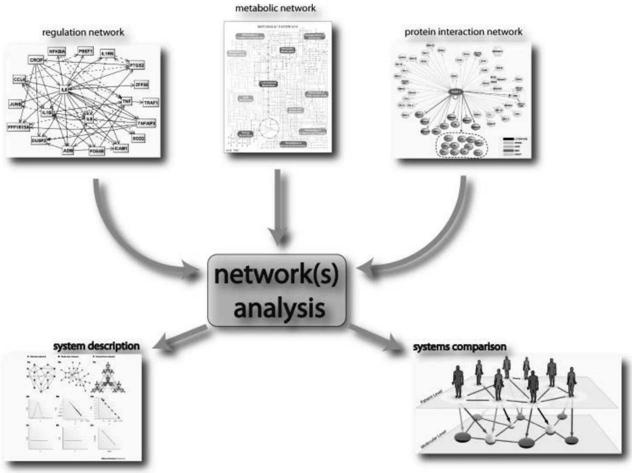


Figure 1.18: The importance of data information in studying complex systems.

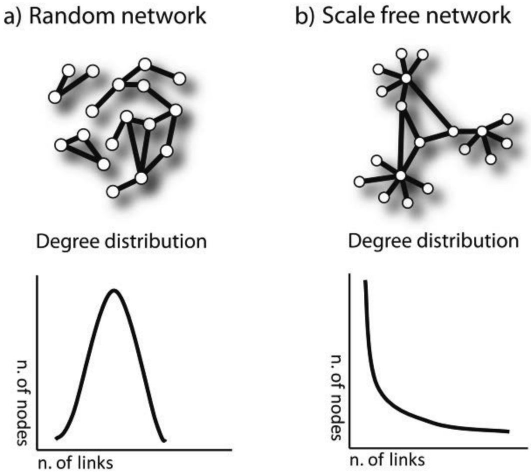


Figure 1.19: Degree distributions of (a) random and (b) scale free networks.

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 struc- tural 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., Se- gurens, 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 com- mon 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 Bioin- formatics, 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. evo- lutionary pathway leading to three isofunctional enzymes. J Biol Chem, 261, 1052–7.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Per- fetto, 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: implica- tions 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 xeno- biotic: 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–

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 struc- tural 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: AnIntroduc- tion. 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 estab- lishment, 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 biosyn- thetic 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. Bioinfor- matics, 15, 759–62.

Fondi, M., Brilli, M., Emiliani, G., Paffetti, D. & Fani, R. (2007). The pri- mordial 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 pre- diction. 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., Schim- mack, 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 innova- tion 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 maris- mortui: 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 oper- ons: does transcriptional connectivity matter? Genetica, 124, 145–66.

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

Hegeman, G.D. & Rosenberg, S.L. (1970). The evolution of bacterial enzyme sys- tems. 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 pro- teins, 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 phyloge- netic 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 func- tion 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 Mi- crobiol, 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., Pa- ley, 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., Rahi- man, 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 en- zymes 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 orthologgroups 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 ar- chaea 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., Bir- ney, 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 oppor- tunism in the design of metabolic pathways during evolution. J Mol Evol, 43, 293–303.

48

Miller, S.L. (1953). Production of amino acids under possible primitive earth condi- tions. 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 path- ways 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 diver- gence 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 comprehen- sive 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 inter-

pretation 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 dis- placement 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 cellularlife: 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). Oper- ondb: 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 neigh- borhoods 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 homol- ogy 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 molec- ular 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 align- ment. 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 evolu- tionary 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 func- tion 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 Micro- biol, 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 path- ways, 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 wefocused 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), flavo-

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

noids (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.

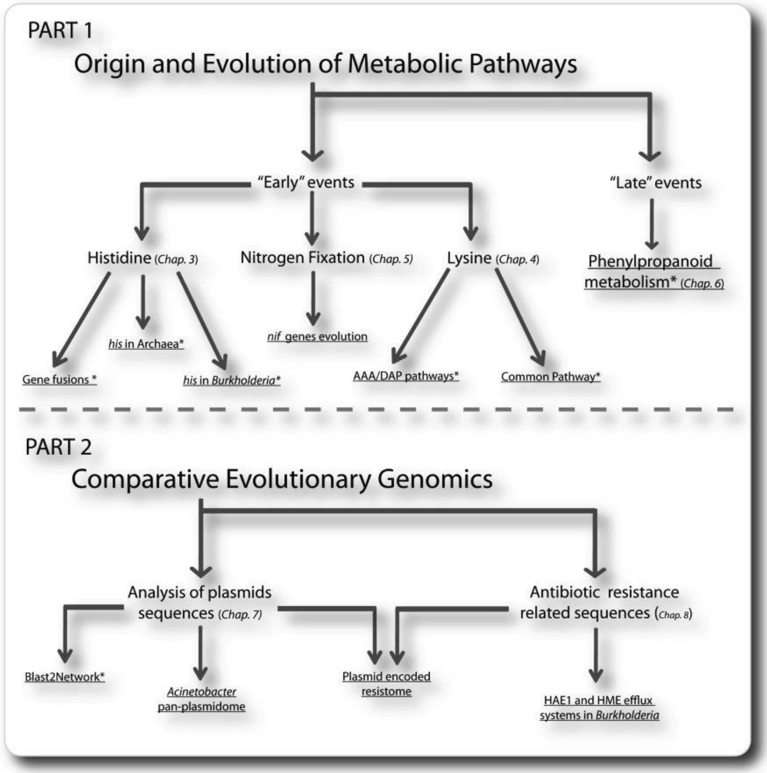

PART 1
Origin and Evolution of Metabolic Pathways
"Early" events
"Late" events
Histidine (Chap. 3)
Nitrogen Fixation (Chap. 5)
Lysine (Chap. 4)
Phenylpropanoid metabolism* (Chap. 6)
Gene fusions *
his in Archaea*
his in Burkholderia*
nif genes evolution
AAA/DAP pathways*
Common Pathway*
PART 2
Comparative Evolutionary Genomics
Analysis of plasmids sequences (Chap. 7)
Antibiotic resistance related sequences (Chap. 8)
Blast2Network*
Acinetobacter pan-plasmidome
Plasmid encoded resistome
HAE1 and HME efflux systems in Burkholderia

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

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 has 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.

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 intoAICAR, which is then recycled into the de novo purine biosynthetic pathway, and imidazole glycerol phosphate (IGP), which in turn

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

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 ?nal 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 organization of his genes has been fo-

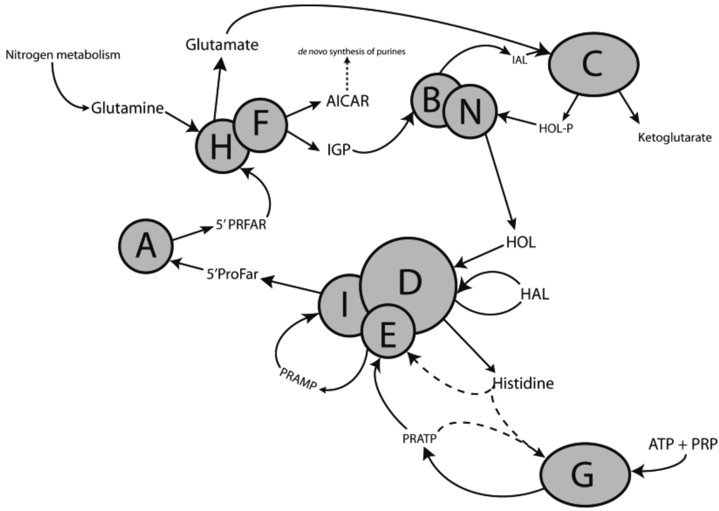


Figure 3.1: The histidine biosynthetic pathway.

cused 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 di?erent 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 influenced 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.).

3.1The 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 (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. In conclusion, although gene duplication is probably the most widely known mechanism responsible for the origin and evolution of metabolic path- ways 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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

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 *Burkholderiahis* 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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

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 different 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 speci?c 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 pathwayorigin 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 these 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 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,

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

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 the work presented in work was to trace the evolutionary pathway leading to this scenario in the extant proteobacteria.

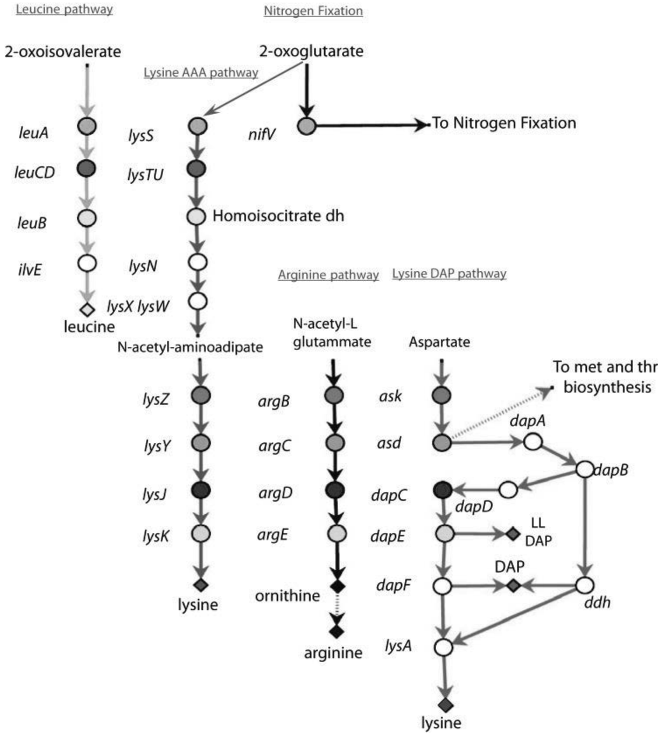


Figure 4.1: The extant lysine, leucine, and arginine biosynthetic routes. Evolutionary relationship between lysine, leucine, and arginine biosynthetic genes. Genes sharing the same color and the same level are homologs. Genes coloured in white have no homolog in the above mentioned metabolic routes.

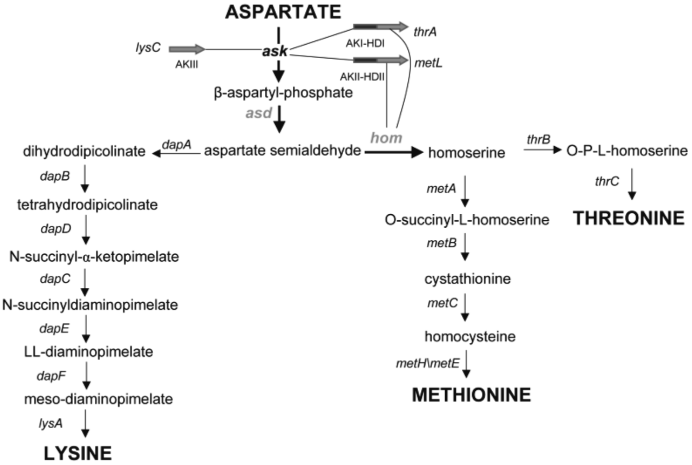


Figure 4.2: The lysine biosynthetic pathway. Genes marked in red (*ask, asd,* and *hom*) constitute the Common Pathway.

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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor’s web-site (www.fupress.com).

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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

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 mapping 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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

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), pro- tection 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.

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

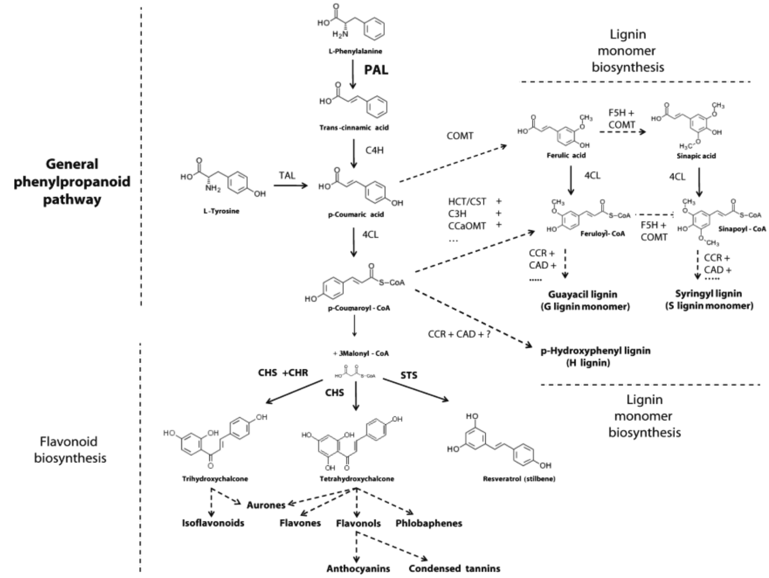


Figure 6.1: The Plant phenylpropanoid metabolism.

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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor’s web-site (www.fupress.com).

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 harboured bymembersoftheEnterobacteriaceaefamilyofγ-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* genus (Chapter 8).
3. Analyze the horizontal flow of plasmids encoded resistome (i.e. genes involved in conferring resistance to antibiotics, Chapter 9).

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

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 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. 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 describing and visualizing 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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

Chapter 8
Exploring plasmids evolutionary dynamics: the Acinetobacter pan-plasmidome

Prokaryotic plasmids have a dual importance in the microbial world: first they have a great impact on the metabolic functions of the host cell, providing additional traits that can be accumulated in the cell without altering the gene content of the bacterial chromosome. Additionally and/or alternatively, from a genome perspective, plasmids can provide a basis for genomic rearrangements via homologous recombination and so they can facilitate the loss or acquisition of genes during these events, which eventually may lead to horizontal gene transfer (HGT). Given their importance for conferring adaptive traits to the host organisms, the interest in plasmid sequencing is growing and now many complete plasmid sequences are available online.

By using the newly developed Blast2Network bioinformatic tool, a comparative analysis was performed on the plasmid and chromosome sequence data available for bacteria belonging to the genus *Acinetobacter*, an ubiquitous and clinically important group of γ-proteobacteria. Data obtained showed that, although most of the plasmids lack mobilization and transfer functions, they have probably a long history of rearrangements with other plasmids and with chromosomes. Indeed, traces of transfers between different species can be disclosed.

We show that, by combining plasmid and chromosome similarity, identity based, network analysis, an evolutionary scenario can be described even 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.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

Chapter 9
The horizontal flow of plasmid encoded resistome: clues from inter-generic similarity networks analysis

By integrating sequence similarity data of plasmid-encoded antibiotic resistance determinants with those coming from a less transferred molecular marker, we constructed a network in which all the sequences that most likely underwent horizontal gene transfer (HGT) were linked together. The analysis of this network revealed that either geographical barriers or taxonomical distance can often be overcome since phylogenetically unrelated bacteria, and/or those inhabiting distinct environments, were found to share common antibiotic resistance determinants, probably as a result of (one or multiple) HGT event(s). Data obtained also revealed that bacteria viable through multiple environments (ubiquitous) are likely to give a crucial contribution to the spreading of bacterial resistance towards antimicrobial compounds. These analyses represent a first attempt to give an almost global picture of the horizontal flow of antibiotic resistance determinants at the whole bacterial community level, also underlining the power of HGT among bacteria and how this 'horizontal flow' is poorly affected by both taxonomy and physical distance. Finally, data presented may be useful in the infections control procedures, suggesting which bacterial species are more likely acting as vectors of antibiotic resistance determinants.

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

Chapter 10
Structure and Evolution of HAE1 and HME efflux systems in Burkholderia genus

The genus *Burkholderia* includes a variety of species with opportunistic human pathogenic strains, whose increasing global resistance to antibiotics has become a public health problem. In this context a major role could be played by multidrug efflux pumps belonging to Resistance Nodulation Cell-Division (RND) family, which allow bacterial cells to extrude a wide range of different substrates, including antibiotics. This study aims to i) identify *rnd* genes in the 21 available completely sequenced *Burkholderia* genomes, ii) analyze their phylogenetic distribution, iii) define the putative function(s) that RND proteins perform within the *Burkholderia* genus and iv) try tracing the evolutionary history of some of these genes in *Burkholderia*.

BLAST analysis of the 21 *Burkholderia* sequenced genomes, using experimentally characterized *ceoB* sequence (one of the RND family counterpart in the genus *Burkholderia*) as probe, allowed the assembly of a dataset comprising 254 putative RND proteins. An extensive phylogenetic analysis revealed the occurrence of several independent events of gene loss and duplication across the different lineages of the genus *Burkholderia*, leading to notable differences in the number of paralogs between different genomes. A putative substrate (antibiotics (HAE1 proteins)/heavy-metal (HME proteins)) was also assigned to the majority of these proteins. No correlation was found between the ecological niche and the lifestyle of *Burkholderia* strains and the number/type of efflux pumps they possessed, while a relation can be found with genome size and taxonomy. Remarkably, we observed that only HAE1 proteins are mainly responsible for the different number of proteins observed in strains of the same species. Data concerning both the distribution and the phylogenetic analysis of the HAE1 and HME in the *Burkholderia* genus allowed depicting a likely evolutionary model accounting for the evolution and spreading of HME and HAE1 systems in the *Burkholderia* genus.

A complete knowledge of the presence and distribution of RND proteins in *Burkholderia* species was obtained and an evolutionary model was depicted. 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.

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

For editorial purposes, the scientific article corresponding to the issue discussed in this paragraph has not been inserted in present book but is available in its open-access digital format on the editor's web-site (www.fupress.com).

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 in- creasing 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 harbour. 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.

Marco Fondi, *Bioinformatics of genome evolution : from ancestral to modern metabolism: phylogenomics and comparative genomics to understand microbial evolution* ISBN 978-88-6655-043-3 (print) ISBN 978-88-6655-045-7 (online) © 2011 Firenze University Press

PREMIO FIRENZE UNIVERSITY PRESS
TESI DI DOTTORATO

Coppi E., *Purines as Transmitter Molecules. Electrophysiological Studies on Purinergic Signalling in Different Cell Systems*, 2007

Natali I., *The Ur-Portrait.* Stephen Hero *ed il processo di creazione artistica in* A Portrait of the Artist as a Young Man, 2007

Petretto L., *Imprenditore ed Università nello start-up di impresa. Ruoli e relazioni critiche*, 2007

Mannini M., *Molecular Magnetic Materials on Solid Surfaces*, 2007

Bracardi M., *La Materia e lo Spirito. Mario Ridolfi nel paesaggio umbro*, 2007

Bemporad F., *Folding and Aggregation Studies in the Acylphosphatase-Like Family*, 2008

Buono A., *Esercito, istituzioni, territorio. Alloggiamenti militari e «case Herme» nello Stato di Milano (secoli XVI e XVII)*, 2008

Castenasi S., *La finanza di progetto tra interesse pubblico e interessi privati*, 2008

Gabbiani C., *Proteins as Possible Targets for Antitumor Metal Complexes: Biophysical Studies of their Interactions*, 2008

Colica G., *Use of Microorganisms in the Removal of Pollutants from the Wastewater*, 2008

Inzitari M., *Determinants of Mobility Disability in Older Adults: Evidence from Population-Based Epidemiologic Studies*, 2009

Di Carlo P., *I Kalasha del Hindu Kush: ricerche linguistiche e antropologiche*, 2009

Pace R., *Identità e diritti delle donne. Per una cittadinanza di genere nella formazione*, 2009

Macrì F., *Verso un nuovo diritto penale sessuale. Diritto vivente, diritto comparato e prospettive di riforma della disciplina dei reati sessuali in Italia*, 2009

Vignolini S., *Sub-Wavelength Probing and Modification of Complex Photonic Structures*, 2009

Decorosi F., *Studio di ceppi batterici per il biorisanamento di suoli contaminati da Cr(VI)*, 2009

Di Patti F., *Finite-Size Effects in Stochastic Models of Population Dynamics: Applications to Biomedicine and Biology*, 2009

Polito C., *Molecular imaging in Parkinson's disease*, 2010

Fedi M., *«Tuo lumine». L'accademia dei Risvegliati e lo spettacolo a Pistoia tra Sei e Settecento*, 2010

Orsi V., *Crisi e Rigenerazione nella valle dell'Alto Khabur (Siria). La produzione ceramica nel passaggio dal Bronzo Antico al Bronzo Medio*, 2010

Fondi M., *Bioinformatics of genome evolution: from ancestral to modern metabolism. Phylogenomics and comparative genomics to understand microbial evolution*, 2010

Marino E., *An Integrated Nonlinear Wind-Waves Model for Offshore Wind Turbines*, 2010

Romano R., *Smart Skin Envelope. Integrazione architettonica di tecnologie dinamiche e innovative per il risparmio energetico*, 2010

Finito di stampare presso
Grafiche Cappelli Srl – Osmannoro (FI)