U.PORTO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

FEUP

UNIVERSIDADE DO PORTO

IBB
INSTITUTO DE BIOTECNOLOGIA E BIOENGENHARIA

FACULDADE DE ENGENHARIA/INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR

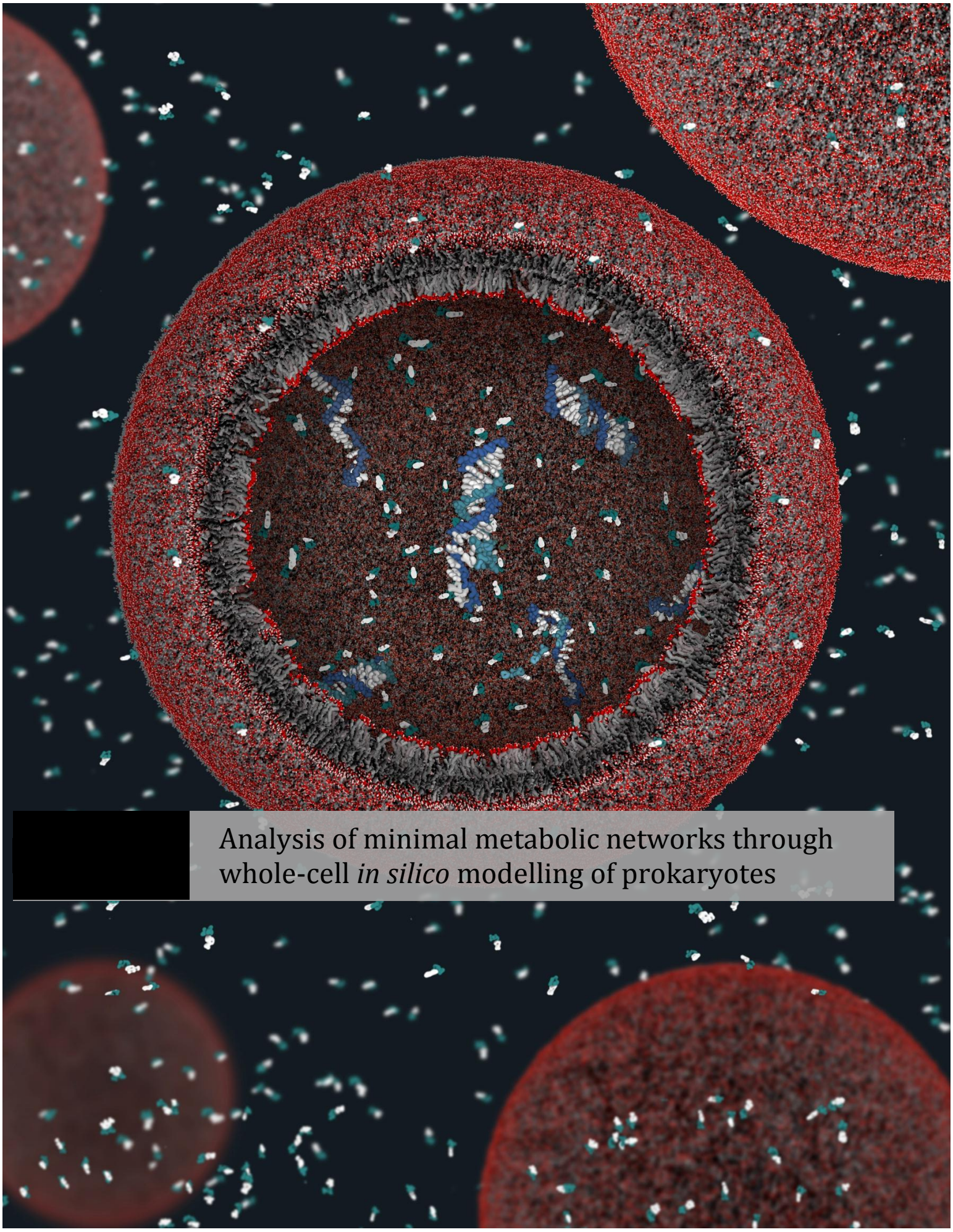
EM COLABORAÇÃO COM O INSTITUTO DE BIOTECNOLOGIA E BIOENGENHARIA/UNIVERSIDADE DO MINHO

Analysis of minimal metabolic networks through whole-cell *in silico* modelling of prokaryotes

Joana Rute Calça Xavier

Mestrado Integrado em Bioengenharia – Ramo Biotecnologia Molecular
Orientador: Isabel Rocha (IBB/UM)

Porto, Julho de 2011



Analysis of minimal metabolic networks through whole-cell *in silico* modelling of prokaryotes

Joana Xavier

"A scientist in his laboratory is not a mere technician: he is also a child confronting natural phenomena that impress him as though they were fairy tales."

Marie Curie, first woman honoured with a Nobel Prize;
Only person winning it in two different sciences until the present day.

Analysis of minimal metabolic networks through whole-cell *in silico* modelling of prokaryotes

Joana Rute Calça Xavier

Mestrado Integrado em Bioengenharia – Ramo Biotecnologia Molecular

Aprovado em provas públicas pelo Júri:

Presidente: Alexandre Quintanilha (IBMC/UP)

Vogal Externo: José Pereira-Leal (Instituto Gulbenkian de Ciência)

Orientador: Isabel Rocha (IBB/UM)

Porto, 15 de Agosto de 2011

Cover image: three-dimensional view of a model protocell approximately 100 nanometers in diameter. Credit: Janet Iwasa, Szostak Laboratory, Harvard Medical School and Massachusetts General Hospital (exploringorigins.org)

Abstract

Systems Biology has been gaining attention in the last few years (Kitano, 2002) showing new potentials in Science (Westerhoff et al, 2009), Research and Graduate Education (Ideker, 2004) and Bioindustrial applications (Blazeck & Alper, 2010) . This renewing is due to the new datasets that are being generated fast but also the novel bioinformatics potentialities that appear to deal with these. In this thesis we reviewed the state of the art towards the definition of a minimal cell, in the context of recently established systems biology tools. We also related the minimal cell concept to the concept of last universal common ancestor (LUCA), as both rely on comparative and integrative techniques of system biology. After this critical review we analysed the state of the art in whole-cell/genome-scale metabolic modelling, considering the main mathematical/computational framework to the analysis of these models – Flux Balance Analysis (FBA) (Varma & Palsson, 1993a). After, we used FBA to predict critical reactions of four different prokaryotic *in silico* models or genome-scale network reconstructions (GENREs). We used *Escherichia coli* (biotechnologically interesting and much studied), *Thermotoga maritima* (accepted as one of the oldest lineages of bacteria living today (Battistuzzi et al, 2004)), *Methanosarcina barkeri* (one archaea) and *Buchnera aphidicola* (an endosymbiont with minimal genome). We also integrated these results with another study of *Escherichia coli*'s derived minimal network independent of Carbon Sources availability (Rodrigues & Wagner, 2009). Calculating and crossing critical reactions across different species - reactions that can't be cut off the network, otherwise the biomass produced would be zero - we aimed at identifying minimal networks' components. We analysed results in the framework of the state of the art, and concluded that current top-down approaches based on parasite's minimal genomes fail to predict minimal life's essential features desired for biotechnological applications and also in the identification of LUCA's metabolic features. We also propose that a bottom-up approach, based on previous top down approaches from more datasets than the genome, is the best in future work for modelling and possibly, constructing minimal cells.

Resumo

A Biologia de Sistemas tem recolhido atenção na última década (Kitano, 2002) mostrando novos potenciais na Ciência (Westerhoff et al, 2009), investigação e ensino superior (Ideker, 2004) e aplicações Industriais (Blazeck & Alper, 2010) . Esta renovação deve-se aos novos conjuntos de informação que estão a ser gerados de forma rápida, mas também às novas potencialidades bioinformáticas que aparecem para lidar com estes. Nesta tese é revisto o estado da arte na definição de célula mínima, no contexto das ferramentas da biologia de sistemas recentemente estabelecidas. Relaciona-se o conceito de célula mínima com o de último ancestral universal comum, dado que ambos os conceitos se constroem a partir de técnicas comparativas e integrativas da biologia de sistemas. Depois desta análise crítica, analisa-se ainda o estado da arte na modelação integral de células /modelação à escala genómica, considerando a ferramenta computacional/matemática principal para análise destes modelos – Análise de balanço de fluxos (FBA, do inglês Flux Balance Analysis) (Varma & Palsson, 1993a). Seguidamente, o trabalho prático descreve o uso de FBA para prever reações críticas de quatro diferentes modelos *in silico* de procariotas, reconstruções de redes metabólicas à escala genómica (GENREs). Usou-se a *Escherichia coli* (interessante para a biotecnologia), a *Thermotoga maritima* (aceite como uma das linhagens de bactérias mais antigas que vive na atualidade (Battistuzzi et al, 2004)), a *Methanosarcina barkeri* (uma archaea) e a *Buchnera aphidicola* (genoma minimizado). Integram-se ainda os resultados com os de um outro estudo da rede mínima da *Escherichia coli* independente da fonte de carbono (Rodrigues & Wagner, 2009). Com o cálculo e intersecção das reacções críticas de diferentes espécies (reacções que não podem ser retiradas da rede, ou a biomassa produzida é zero) pretendeu-se identificar os componentes de uma rede metabólica mínima. Analisaram-se os resultados no contexto do estado da arte concluindo-se que as abordagens top-down baseadas no genoma mínimo de parasitas falham na predição das características essenciais de vida mínima desejadas para aplicações biotecnológicas e também na identificação das características metabólicas do último ancestral comum. Propõe-se uma abordagem bottom-up baseada em abordagens top-down comparativas de mais conjuntos de informação do que o genoma no trabalho futuro para a modelação e possível construção de células mínimas.

Acknowledgments

Este espaço não poderia deixar de estar em português, representando o meu enorme gosto pessoal pela transmissão, partilha e divulgação do conhecimento científico pela comunidade em geral mas em especial pelos que me rodeiam. É dedicado em especial a estes – os que comportam a minha paixão e divagações pela Ciência e Tecnologia que me deixa, por vezes, com Q.E. reduzido; a todos os que permitiram que esta dissertação e este curso chegassem a bom termo.

Em primeiro lugar, à minha família, que me acompanhou, apoiou, e ainda se interessou muito nestes cinco anos da nova aventura que foi a Bioengenharia no Porto, longe de casa, e ainda mais nesta etapa final tão complexa e intensa. Muito obrigada pela dedicação, compreensão e carinho, mãe, pai, irmã e avós, em especial.

Aos amigos mais próximos; Diogo, Ana e Vera, que para além das suas amizades inequívocas me acompanharam em longas horas de escrita das nossas dissertações; Sara, João e Carlos, Patrícia, Raquel, Rúben – pelos momentos de descontração que são tão indispensáveis a um bom funcionamento cerebral!

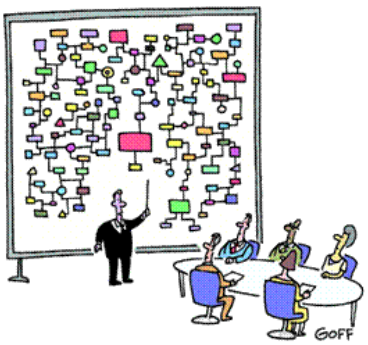
Aos colegas e amigos de Bioengenharia da UP de 2006, que me fizeram companhia no mundo desconhecido e inovador da Bioengenharia que ajudámos a criar, na FEUP e no ICBAS (e por esse país fora), como primeiros caloiros e primeiros formados. Que me acompanham a ouvir o estranhar ao nome da Bioengenharia, o seu soar como uma engenharia inferior, quando é a mais complexa que existe e que existe por si só, que descreve as únicas máquinas que o ser humano não consegue ainda montar e desmontar prontamente – as células, órgãos e organismos. Aos professores de Bioengenharia na FEUP e no ICBAS mas também a todos os investigadores que tive o prazer de ter como professores no IBMC, INEB e CIIMAR pelo conhecimento que transmitiram mas também pela inspiração; um especial obrigado à Catarina Santos (IBMC) que tanto me ensinou. Um agradecimento especial à Perpétua do Ó (INEB), à Mónica Sousa (IBMC) e à minha orientadora Isabel Rocha (IBB) que me corroboraram que as

mulheres podem ser grandes cientistas e engenheiras não deixando de ter tempo para serem mães. Ao professor Alexandre Quintanilha, um dos melhores professores e divulgador de ciência que já conheci e tive o prazer de ouvir, que muito me inspirou nestes cinco anos. Pela enorme felicidade que senti quando soube que seria o responsável por esta unidade curricular e pela honra de poder acabar o curso contando com a sua orientação.

Ao professor Manuel Simões, que pelas suas aulas dinâmicas onde podíamos contar com a presença de investigadores externos, que me deu a conhecer a professora Isabel Rocha que lidera um grupo de Biologia de Sistemas de topo, em Portugal. À professora Isabel Rocha, de novo, por me receber tão bem numa altura tão cheia da sua vida, não deixando de acreditar nas minhas capacidades. Ao grupo, SysBioPseg (IBB/CEB – Universidade do Minho, Braga) por tão bem me ter acolhido para este trabalho e por me mostrar (de uma perspetiva que no Porto ainda não existe) o que pode ser a Bioengenharia. A todos os membros, em especial à Marisa pelo apoio pessoal, Paulo pela imensa ajuda com o OptFlux e ficheiros SBML e ao Daniel que me acompanhou e aconselhou prontamente com sugestões decisivas para o rumo deste trabalho.

Finalmente e teoricamente, como de teoria mais este trabalho se trata, agradeço às primeiras formas de vida que se replicaram e a todas as que lhes seguiram, lutando pela sobrevivência, e a todas as que lhes seguiram permitindo o ser humano que eu sou. Às incríveis e simples formas de vida procariota que são simbiontes comigo, até às incríveis árvores milenares, há tanto tempo fábricas de oxigénio ainda inexplicadas; a todas as formas de vida que me apaixonam e me lembram que os grandes passos da evolução se fizeram não só à luz da pressão seletiva, mas também à luz da cooperação.

A todos os possíveis leitores, e espero que desfrutem desta tese como eu dela desfrutei.



"And that's why we need a computer."

(Imagem de www.sysbio.org)

Table of contents

Abstract vi

Resumo vii

Acknowledgments viii

Table of contents x

Abbreviations xiii

List of figures xiii

List of tables xiv

1. Introduction 1

1.1 Systems Biology and Minimal Cells 1

1.2 Motivation and Goals 2

2. State of the art 4

2.1 Minimal cell – For long a Systems Biology concept 4

2.1.1 Minimal Genomes 6

2.1.1.1 Important steps in designing a minimal genome – Theoretical and experimental approaches 7

2.1.1.2 Advances in large DNA sequences synthesis – How close are we to produce a minimal cell? 8

2.1.2 Minimal Metabolisms 9

2.1.2.1 The study of minimal metabolism *in silico* 10

2.1.3 Limitations and Prospects 12

2.2 Primordial life and metabolism – first natural minimal networks in the light of Systems Biology 14

2.2.1 – Origin of Life - Generalities 14
2.2.2 – Primordial metabolism 17
2.3 Systems Biology and Genome-Scale Metabolic Models 18
2.3.1 Building a genome-scale metabolic model 21
2.4 Flux Balance Analysis in genome-scale metabolic modelling 22
2.4.1 Flux balance analysis: Foundations of the technique 25
2.4.1.1 Model Construction/System definition 25
2.4.1.2 Mathematical representation 25
2.4.1.3 Mass Balance 25
2.4.1.4 Optimization/ Definition of the objective function 26
2.4.1.5 Flux Calculations by Linear Programming 27
2.4.1.6 Computational resources 27
2.4.2 After basic FBA: 2nd generation FBA models 28
2.4.2.1 Determining fluxes in mutants – MOMA and ROOM 28
2.4.2.2 Considering the evolution of fluxes with time - DFBA 30
2.4.2.3 Including thermodynamic restrictions - EBA 30
2.4.2.4 Including transcriptional regulation - rFBA 31
2.4.3 Genome-scale FBA based modelling: Industrially and scientifically interesting 32
2.5 Metabolic network alignment 34
3. Methods and Results 36
3.1 Analyses of all predictive prokaryotic GENREs and choice of the Case-Studies 36
3.2 Calculating the essential reactions for biomass formation 42
3.2.1 Conserved reactions among three GENREs from the three phylogenetic domains 42
3.2.2 Calculating essential reactions with Optflux 43
3.3 Set comparison and phylogenetic reconstruction 45
4. Discussion 52
4.1 Analysis of all predictive prokaryotic GENREs 52
4.2 Choice of Case Studies 53

4.3 Discovering conserved essential reactions 54

4.3.1 Conservation of reactions between Prokaryotes and Eukaryotes (Fig. 17) 54

4.3.2 Conservation of essential reactions in Prokaryotes (Fig. 18 and Table 5) 55

4.3.2.1 Chorismate synthase – the monofunctional, central and ancestral enzyme 56

4.4. Studying minimal essential metabolism, minimal life and early life with Systems Biology . 59

5. Conclusions 61

5.1 Theoretical and *in silico* testing and posterior *in vivo/in vitro* experimentation – Saving time and money in Molecular Biotechnology 62

References 64

Abbreviations

GENRE – Genome Scale Network Reconstruction

COG – Cluster of Orthologous Genes

LUCA – Last Universal Common Ancestor

FBA – Flux Balance Analysis

MOMA – Method of Minimization of Metabolic adjustment

KEGG – Kyoto Encyclopedia of Genes and Genome

List of figures

Fig. 1 – A minimal cell containing biological macromolecules and pathways proposed to be necessary and sufficient for replication from small molecule nutrients (Forster & Church, 2006b) 9

Fig. 2 – COG (clusters of orthologs groups) classification of proteins of different prokaryotic endosymbionts of insects (Nakabachi et al, 2006) 11

Fig. 3 – Major relationships between different omic datasets and the integration of Systems Biology and Bioinformatics.....13

Fig. 4 – Timeline of the major events known in the history of Life on Earth. 16

Fig. 5 – Steps in the construction of a genome-scale Metabolic Model. 20

Fig. 6 – Different topologies of biological networks.. 21

Fig. 7 – Mathematical calculation of flux distributions (solution spaces) and optimal fluxes (optimal solutions) requires the definition of constraints 23

Fig. 8 – Initial steps for FBA (Based on (Kauffman et al, 2003)).....26

Fig. 9 – Schematic representation of the feasible solution space for the flux distribution of wild type and knockout strains, with MOMA suboptimal solution.....29
Fig. 10 – Stoichiometric network reconstruction and analysis with FBA (Gianchandani et al, 2010) 34
Fig. 11 – Heatmap of metabolic reconstructions showing current validation levels. In (Oberhardt et al, 2009). 37
Fig. 12 – Distribution of prokaryotic phyla represented in Genome-Scale Metabolic Reconstructions 38
Fig. 13 – Phylogenetic Tree showing the relationships between all Prokaryotic species for which a predictive GENRE exists. 40
Fig. 14 – Conserved reactions among reconstructed metabolic models from the three domains of life 42
Fig. 15 – OptFlux software platform – Simulation environment.. 44
Fig. 16 – Venn's Diagram representing the intercepted sets of reactions in different GENREs analyzed.. 45
Fig. 17 – Conservation of metabolic reactions (grouped in metabolic subsystems). 47
Fig. 18 – Essential reactions conserved in different prokaryotic metabolic models.. 48
Fig. 19 – Phylogenetic tree of chorismate synthases (protein sequences). 51
Fig. 20 – Reaction catalyzed by chorismate synthase (Kitzing et al, 2004). 57
Fig. 21 – *Thermotoga maritima*'s Tryptophan, Tyrosine and Phenylalanine biosynthesis pathway, obtained from KEGG 58

List of tables

Table 1 – Number of molecules in a single *Escherichia coli* cell Adapted from (Deamer, 2009) ..4
Table 2 – Omics data types and examples of databases.....19
Table 3 – Significant events in the development of the FBA technique 24
Table 4 – Updated list of prokaryotic genome-scale metabolic reconstructions and their main features. 39
Table 5 – Most conserved essential reactions for biomass growth in prokaryotic metabolic networks.....49

1. Introduction

"Life, like a machine, cannot be understood simply by studying it and its parts; it must be put together from its parts."

(Forster & Church, 2006a)

1.1 Systems Biology and Minimal Cells

Systems Biology is starting to be accepted as the way in which Biology should be redesigned and reintegrated to become closer to the other natural Sciences (Westerhoff et al, 2009). Although with a long history, especially in immunology and developmental biology, systems biology, as the approach that systematically and dynamically analyses all the parts and their interactions in a biological system (cell, tissue, organ, and organism), is being reaffirmed in the new century, with genome sequencing and powerful computers that can account for an astonishing amount of information that one such system can include. This new power of bioinformatic tools helps the understanding of the cell's complex functioning, with mathematical and physical laws, integrating all genes, proteins and metabolites in interactive networks. This formal analysis reveals new functional states that arise when multiple molecules interact simultaneously. The applications are vast, especially in bioengineering that aims to understand and create new biological or bio-inspired systems and products. Metabolic engineering is an example of the great impact of systems biology, with strain improvement for the production of drugs and valuable metabolites.

The minimal cell concept puts together applied systems biology, for the design of synthetic cells for biotechnological applications, but also basic biology fundamentals – understanding the most basic, essential components of life. Ultimately, this latter approach can converge with the definition of the last universal common ancestor (LUCA), from which all life forms have evolved. In fact, both approaches need comparative systems biology, as they

require the most conserved features between species; these features are proved to be ancient by the darwinistic theory of evolution – take as an example ribosomal RNA, which is used to establish phylogenetic relationships between organisms.

1.2 Motivation and Goals

The motivation for this work appears in the interface of evolutionary and systems biology. The recent predictive genome-scale metabolic reconstructions open gates to the investigation of many fundamental and applied questions. This work consists on a theoretical and practical dissertation about minimal cellular functions related to ancestral-derived metabolic features. The minimal cell concept is not only conceptual anymore, as nowadays the perspective of building a synthetic cell for biotechnological applications is taken in serious consideration (Forster & Church, 2006a). Recently, an expanded review on specific biotechnological considerations on this matter was published (Foley & Shuler, 2010).

This dissertation aimed first at being an extended and critical review of the state-of-the-art in the construction of minimal cells through Systems Biology approaches (Chapter 2). These were until now mainly genomic approaches, as this was the first omics dataset to be easily available, which is highlighted. It was intended after to analyse the application of metabolic modelling towards the same goal - the creation of a minimal cell. It was also a goal to review the state of the art in the study of ancestral life and ancient metabolism as it converges with the systematic, top-down or bottom up approach that is also needed to infer minimal cell's features and important biological fundamental knowledge. After, Systems Biology specific field of Genome-Scale Metabolic Modelling was reviewed, especially Flux Balance Analysis (Varma & Palsson, 1993a) that is the main mathematical and computational technique used to analyse Genome-Scale Network Reconstructions (GENREs) and that was used as the main methodology of this work. There is still a reference to the state of the art in Metabolic Network alignment that was a possible direction to follow in the course of the practical work to be done.

The practical component of this thesis consisted on a comparative study of several different prokaryotic GENREs, focused on identification and comparison of essential or critical reactions. These reactions are those that cannot be cut off the network, otherwise the biomass produced by the model will be zero – therefore we aim the identification of minimal networks' components. For this purpose, the goals were to review all the available predictive GENREs for

a careful case-choice study based on validation level (how predictive the model has proved to be when compared to experimental data) and also phylogenetic reach. After, we calculate essential reactions for each case, and we compare the sets of essential reactions obtained, integrating the results with the state of the art of minimal cell design and construction. The results have interest for basic bioengineering, in molecular biotechnology, but also in the study of ancestral metabolic networks.

2. State of the art

2.1 Minimal cell – For long a Systems Biology concept

One simple prokaryotic cell can be viewed as a highly complex dynamic bio-nanomachine that bioengineers cannot assemble from its parts. *Escherichia coli*, for example, has millions of proteins, thousands of RNA molecules, millions of lipids, etc., all of them working together in subsystems and complexes. However, for each cell, there is only one circular double helix DNA genome (See Table 1)

Table 1 – Number of molecules in a single *Escherichia coli* cell Adapted from (Deamer, 2009) .

Molecular component	Number of molecules
Kinds of proteins	1,850 (mostly enzymes)
Total number of proteins	2,36 million
RNA in ribosomes	18,700
Transfer RNA	205,000
Messenger RNA	Variable depending on growth cycle
DNA	One circular double helix
Lipid	22 million
Lipopolysaccharide	1,2 million
Peptidoglycan	One (forms cell wall)
Glycogen	4,360 (energy storage)
Plasmids	0-200 (Watve et al, 2010)

The comprehension of the cell has boomed with modern biotechnology techniques, which have been focused on genome comprehension since the discovery of DNA and the sequencing of the first genome. Therefore, minimal cells have been studied mostly as minimal genome-containing cells. With Systems Biology affirmation in the several new omics datasets (proteomics, metabolomics, lipidomics, etc) cells are starting to be looked at as a whole, and the approach of this new discipline appears to be very appropriate to conceive minimal cells. The identification and interconnection of cellular parts in dynamic systems and ultimately, the construction of models are the goals of Systems Biology. In fact, the old approaches to minimal cells – Top down and Bottom up - are general terms for two strategies of information processing and knowledge ordering in any systems science, or philosophy. Therefore, systems biology has always been the approach in the search for the minimal cell.

Top down approaches aim at reducing genomes of existent minimal organisms to the smaller they could be while still living. The species used are mainly parasites or endosymbionts with the smallest genomes known, as *Mycoplasma genitalium* and *Buchnera aphidicola*, and the information is obtained from computational and experimental studies as a start (for a review, check (Henry et al, 2010)). Top down approaches have always come from genomes, never proteomes or other omics data-sets; for this reason, there is a general agreement that they will not take us to the minimal possible cells in chemical terms.

Bottom-up approaches aim at build artificial chemical systems that could be called alive. There is no successful experimental work done yet towards this goal. It has been said that metabolism is the stepchild in the family of bottom-up approaches (Szathmary, 2005).

It is commonly accepted that a minimal cell has to have some form of metabolism, genetic replication from a template and membrane production (Szathmary, 2005).

To better analyse the state of art of the minimal life concept, this section is divided in two: minimal genomes are explored first, along with the excitement they have generated in modern biotechnology; after, the concept of minimal metabolism (the set of essential biochemical reactions for life), more neglected also due to technical restrictions, is reviewed.

2.1.1 Minimal Genomes

The minimal genome concept is being refined over the last few years, since synthetic biology came to scene in life sciences, presenting new prospects in genome's synthesis *de novo*. Defining and understanding a functional minimal genome is a contextual task, since each species has its own set of genes providing the functions that allow that species to live well adapted in its environment, and play its role in its ecosystem. So if the aim is to design the simplest genome possible for life to happen, one would be dealing with an organism resembling a Prokaryote (simplest life forms known); but designing the simplest plant cell possible for the high yield photosynthetic-based production of a eukaryotic bio-product would probably be much harder. Designing the minimal genome required for rational intelligence, in the most extreme point of view, was not an easy task for evolution. *Homo sapiens* possesses a complex genome with a small set of protein-coding genes – only 1,5% (Lander et al, 2001); it is well known, however, how important are the huge amount of RNA-coding and regulatory sequences for a correct expression of genes. From another side, our survival and life quality is totally dependent on other species' life, in several aspects, from plant's photosynthesis and nitrogen fixation to intestinal microbial flora balance, to name just a few. Regarding this, establishing a conceptual minimal genome would completely depend on the **functions** desired for that genome, its **life context** and independence in habitats. A great barrier lies between the applied use of minimal genomes in biotechnology, or its investigation in basic life sciences aiming to understand life itself and its origins. High-tech cultures of microbes producing interesting compounds are advanced life forms totally dependent on one of the most complex forms – human life; these artificially engineered bacteria, even with a minimal genome, would probably be far away from the unknown cells that appeared when life began on Earth.

Imagination could drive us to the most amazing living machines with minimal genomes directed to numerous important applications, as bioremediation, drug and food production, and so on. But the technology required for synthesizing DNA still constrains these possibilities a lot. Going to the bottom question, the minimal genome is indeed seen today, realistically, as the minimal prokaryote genome able to grow, reproduce and evolve in a basic environment. So far, a living cell has not yet been synthesized from non-living matter. "How far can we push chemical self-assembly?" has recently been postulated as one of the big 25 questions in science for the next 25 years (Service, 2005).

2.1.1.1 Important steps in designing a minimal genome – Theoretical and experimental approaches

The 'minimal genome' approach usually aims to estimate the smallest number of genetic elements sufficient to build a modern-type free-living cellular organism (Mushegian, 1999). Establishing a minimal genome can cross different fields of Biology as Comparative Genomics, Genetics and Biochemistry, and Systems Biology. One of the first theoretical studies identifying a small set of essential genes used the first two sequenced genomes (*Haemophilus influenza* and *Mycoplasma genitalium*) to infer the number of approximately 256 genes as close to a minimal genome capable of producing a modern free-living organism (Mushegian & Koonin, 1996). On 16 June 2011, NCBI reported 1646 prokaryotic genomes fully sequenced (112 archaeal, 1534 bacterial), and one of the smallest that can be grown in axenic culture remains the second that has been sequenced – *Mycoplasma Genitalium* (Galperin, 2006). This feature adds to many others making this a reasonable model-organism to the study of the minimal genome – it is a wall-less prokaryote with a small genetic redundancy and an obligate parasite requiring small adaptive capacity. A recent study used transposon mutagenesis to systematically disrupt genes to improve the pre-existing knowledge about the number of minimal genes of *M. genitalium* (Glass et al, 2006). This study hypothesized that 387 protein-coding and 43 structural RNA genes could sustain a viable synthetic cell, not accounting for the synergistic effect that can possibly happen with several mutations. Other studies have used mutagenesis to inactivate, one by one, bacterial genes, establishing which of those are indispensable for growth in controlled conditions. *Bacillus subtilis* appears to need only ~271 genes which fall in relatively few but large categories: information processing, cell envelope, shape, division, and energetics (Kobayashi et al, 2003). In the latter study the authors demonstrated almost total conservation of these so-inferred essential genes among bacteria but also high conservation in eukarya and archaea (70%); also important were the discoveries about non-expected essential genes (Kobayashi et al, 2003). The comparisons were based solely on presence of genetic sequences (orthologs) in genomes. *Escherichia Coli* has also been target for similar studies (Gerdes et al, 2003), as well as *Buchnera aphidicola,* one endosymbiont of insects (Gil et al, 2002).

Mutagenesis studies test only for the need of one gene at the time, so saying that the non-indispensable genes in this manner constitute by themselves the essential set might be

pretty reductive: double or triple mutations can have a synergistic effect, while their respective individual mutants are viable; also, the long-term survival is usually not experimented. By another side, these results also provide "false essentials" due to the genes that were not mutated on the screen, creation of toxic partial complexes or pathways, unexpected effects of neighbour-genes, etc. Here Systems Biology and Synthetic Biology come to scene; both computer models and simulations and biotechnology are improving very fast, making possible today to test for the viability of altered genomes – *in silico* or in the lab.

2.1.1.2 Advances in large DNA sequences synthesis – How close are we to produce a minimal cell?

The technology for synthesizing DNA has suffered a tremendous progress in the last three decades. Since the total synthesis of a 207bp-long gene (a biologically functional tyrosine suppressor transfer RNA gene) in 1979, by Khorana's team (Sekiya et al, 1979), important steps have been taken. The chemical synthesis of the poliovirus full-length cDNA (~7440bp) (Cello et al, 2002) was probably the most important following mark, in 2002; it took many months to achieve, through a complex process. In the next year, 2003, Craig Venter's group was able to synthesize the complete infectious genome of bacteriophage ΦX174 (5,386bp) from a single pool of chemically synthesized oligonucleotides, in just two weeks (Smith et al, 2003), taking minimal genomes' synthesis to another level. However, viruses are not considered living-beings, so the serious advance was taken in 2008, by the same group, with the total synthesis, assembly and cloning of a *Mycoplasma genitalium* genome (Gibson et al, 2008). With 582 970bp, this genome was built from ~104 synthetic oligonucleotides each one with ~50 nucleotides; many errors can occur in the pathway to achieve such a goal, but this work brings new promises in dealing with previously-thought hard difficulties in Synthetic Biology.

Until then, large *in vitro* DNA assemblies have used type IIS restriction enzymes that create unique sticky ends on the components of the assembly, which are then joined by ligation; however with large pieces, it is increasingly difficult to find a type IIS enzyme that does not cleave within the piece. The astonishing work used *in vitro* recombination of overlaps between the ends of the fragments to be assembled, not depending on such enzymes and constructing the first synthetic genome (Gibson et al, 2008).

2.1.2 Minimal Metabolisms

The concept of minimal metabolism has been long neglected compared with the concept of minimal genome. In fact, metabolism, comprising all the chemical modifications made by the cell, is one of the most flexible elements of it - many alternative pathways exist to achieve a specific metabolic goal. Particularly, ATP can be synthesized using a variety of methods (Henry et al, 2010).

The excitement created around minimal genomes can't overcome the limitation that synthetic minimal genomes can only live in a *chassis* suitable for their replication. No complex genome would replicate without free new nucleotides, already existing polymerases, tRNAs, and most certainly a membrane to constrain it physically (to name just a few - Fig. 1).

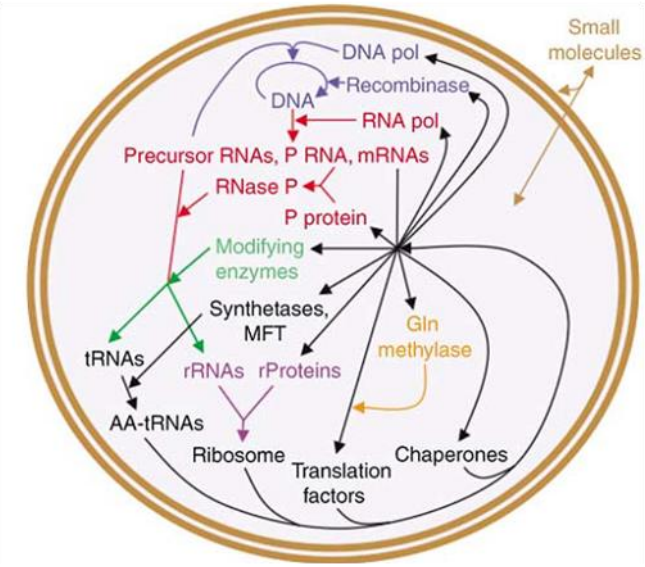


Fig. 1 – A minimal cell containing biological macromolecules and pathways proposed to be necessary and sufficient for replication from small molecule nutrients (Forster & Church, 2006b)

The interest that industrial biotechnology has in synthesizing minimal life goes beyond the scientific and theoretical goal of designing the minimal living entity possible; for industrial applications, the desirable would be to achieve a high-yield organism (in biomass or bioproduct production) at the expense of the cheapest, **minimal medium**. In fact, *Mycoplasma genitalium* is seen as close to a minimal cell as it has the smallest genome capable of being grown in pure culture, but *Escherichia coli*, with a genome **eight-times** bigger, grows **fifty-times faster**. This is due to the more efficient metabolism of the latter. So, which would be preferable to be used in Biotechnology processes? Also, being *E coli* the model bacterium in life sciences, the extended knowledge already available for its metabolic pathways and other physiological processes would certainly help a lot in perceiving the minimal functionalities required for life. In this manner, a minimal cell could be designed, not with the base of the

“minimum number of genes” but the minimum number of desired functions, which is still poorly explored. There’s no sufficient technology to produce a mutant for all the non-essential genes hypothesized, for example, in *E. coli*, although many advances have been achieved with transposon-deletion techniques (Goryshin et al, 2003).

One can also analyse the minimal metabolism of the minimalistic organisms for minimal genomes analyses. See Fig. 2 for the COG (clusters of orthologs groups) classification of proteins of insect endosymbionts, including the prokaryote with the smaller genome known to date, *Carsonella ruddi,* which has a single circular chromosome of 159,662 base pairs (Nakabachi, 2008). It is notorious the importance of translation and amino acid metabolism in all endosymbionts, but it is also significant to notice that, for being a symbiont, *Carsonella* lost important metabolic features as cell envelope, coenzyme and nucleotide metabolism.

2.1.2.1 The study of minimal metabolism *in silico*

A minimal gene-set obtained from the comparison of two small genomes (*Haemophylus influenzae* and *Mycoplasma genitalium* (Mushegian & Koonin, 1996)) motivated one dynamic, *in silico* study of metabolic viability of a minimal cell (Chiarugi et al, 2007). The authors refined the gene set to obtain a virtual cell that was proven to live *in silico* (ViCe) although they make the reservation that it remains to be tested experimentally. ViCe includes: a complete **glycolytic pathway** coupled with the synthesis of ATP through the ATP synthase/ATPase transmembrane system; a **Pentose Phosphate Pathway**; enzymes for **glycerol-fatty acids condensation** (but no pathways for fatty acids synthesis that need to be taken from the outside); ‘‘salvage pathways’’ for **nucleotide biosynthesis** (thymine is the only nucleotide the cell is able to synthesize de novo); a proper set of **carriers** for metabolites uptake; the necessary **enzymes for protein synthesis**, and the whole machinery necessary for DNA synthesis is also included in ViCe. Notoriously, this is an attempt to build a true minimal cell that is totally dependent on a rich medium.

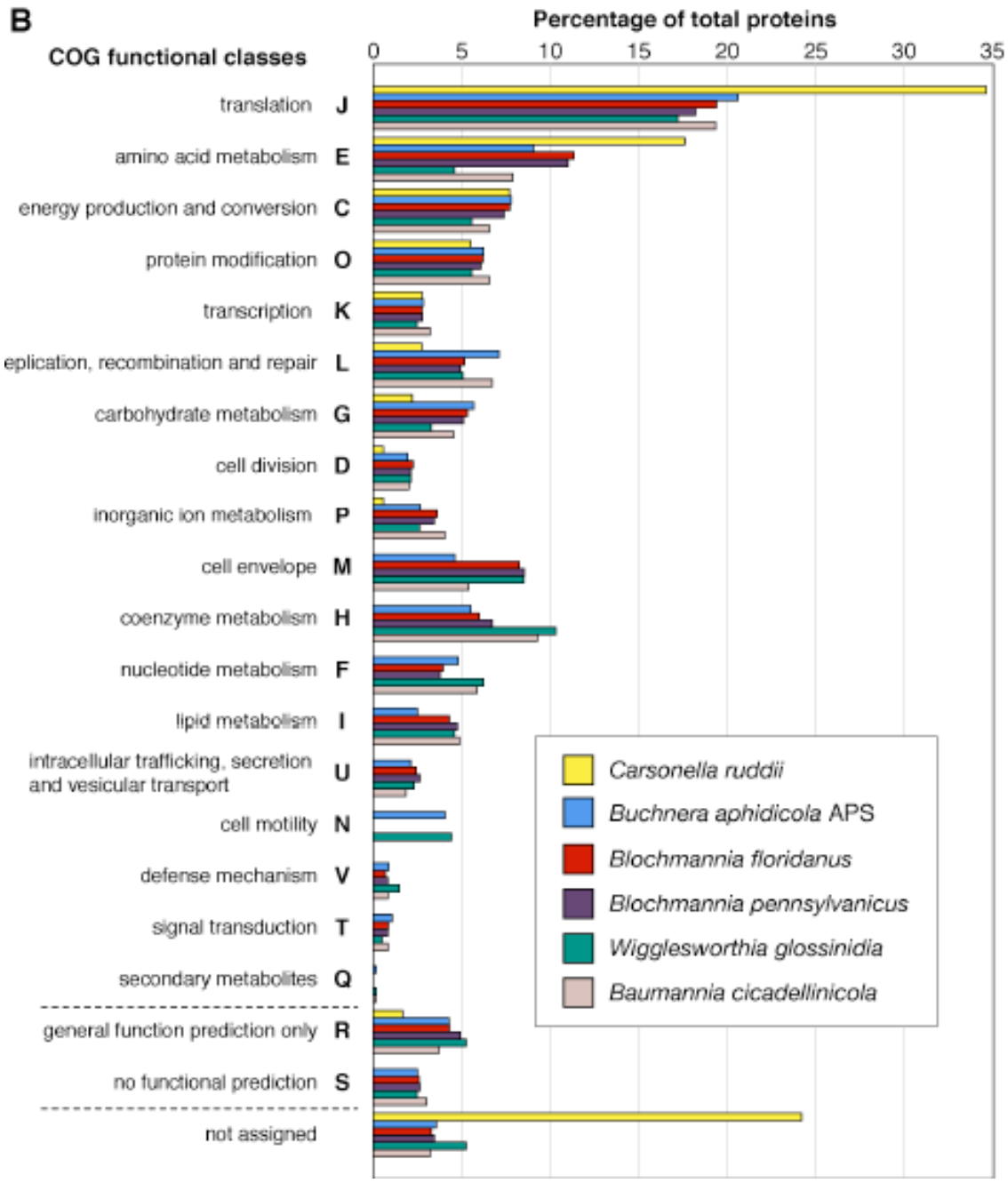


Fig. 2 – COG (clusters of orthologs groups) classification of proteins of different prokaryotic endosymbionts of insects (Nakabachi et al, 2006)

The limitations of comparative genomics in identifying minimal metabolic features have been recognized by the authors of another *in silico* study of a minimal metabolism (Gabaldon et al, 2007) – they recognize that comparative genomics yields sets of genes with few representatives of metabolic enzymes; another important point recognized by these authors is the relevance of the medium for the studied minimal cell.

In silico whole cell metabolic models have advanced fast in the last few years, and they are certainly a way to overcome certain experimental wet lab limitations (See Section 2.3 – Systems Biology and Genome-Scale metabolic models). The previously referred two works have relied on *in silico* metabolic modelling. Another important work should be referred in this context that used one *Escherichia coli in silico* whole-cell model (Edwards & Palsson, 2000a) to predict cell viability in minimal and complex media (Burgard et al, 2001). However, it has been criticized that this analysis should be repeated in a pan-genomic metabolic network, which implies the utilisation of other strains of *E. coli* (Henry et al, 2010).

The only work that compared predictions of different GENREs to infer minimal metabolic features was directed to the identification of autocatalytic components of metabolic networks (Kun et al, 2008); the authors also related the results of their work with the properties of ancestral networks.

2.1.3 Limitations and Prospects

Comparative Proteomics and Systems' Biology are coming together to understand what is common to all life forms and what is essential to each one of them – therefore approaching the theoretical, minimal cell. By another side, the great amount of high-tech experiments with Synthetic Biology (Cello et al, 2002; Gibson et al, 2008; Smith et al, 2003) opens gates to next-century experiments, which could connect to these more theoretically-sustained minimal cells. The **importance of the context** – the habitat where the cell grows and lives, believed today to influence strenuously the genome's behaviour – and the **genome's behaviour and reach** – (today seen much more organic than last century Biology predicted with the rigid central dogma) are parameters not well discussed today when the minimal genome's concept is reviewed. In Fig. 3 the major omics datasets are represented and is evident the control of the genome over other levels of information; one can see that the central dogma of biology has no longer signification in this context – the genome encodes proteins by mRNA, and proteins deal with metabolites; but metabolites act on proteins and DNA *per se,* being toxic or beneficial; without the necessary nutrients cell can't grow and divide and will not live anymore.

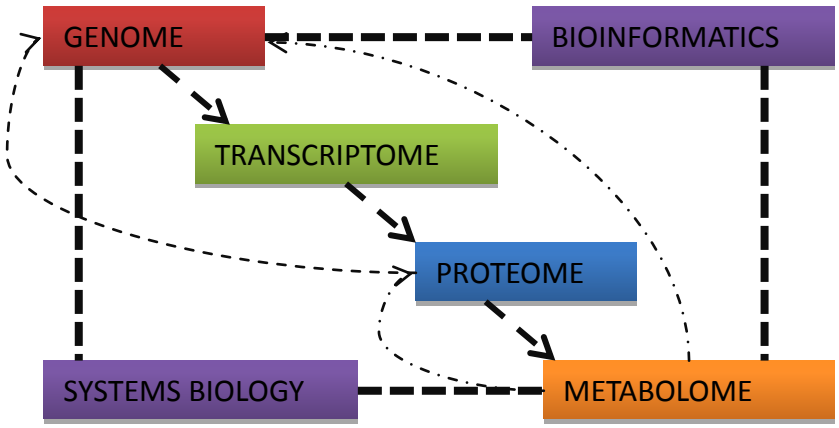


Fig. 3 – Major relationships between different omic datasets and the integration of Systems Biology and Bioinformatics. Metabolomics represents the ultimate level of information inside cells

If the aim is to design a minimal genome to the production of interesting compounds, one may be losing time and money using *Mycoplasma genitalium*–based minimal genome; it is defended today that putting foreign sequences in *E coli* would retrieve much higher yields with much smaller costs. This is the point where we start to understand that the minimal genome is so conceptual, that the primary approaches should probably be much more science-based and technology directed in a bottom-up approach. Understanding the genome's minimal functions and their relationships (as proteins function in complex networks) capable of sustaining life is a goal for next century Biology.

2.2 Primordial life and metabolism – first natural minimal networks in the light of Systems Biology

2.2.1 – Origin of Life - Generalities

Archebiosis: origin of living things from not-living materials

«[...] I should like to live to see Archebiosis proved true, for it would be a discovery of transcendent importance; or, if false, I should like to see it disproved, and the facts otherwise explained; but I shall not live to see all this»

Charles Darwin, Letter to Wallace. August 28, 1872

The Origin of Life is a totally obscure field in Science. While scientists and more recently bioengineers are trying to assemble life *in vitro* from its complex parts, they still don't know how it originated first in nature. Postulating the question in another way, the chemical laws and contexts that made appear the simpler prokaryotes from non-living matter are unknown. This question has generated great controversy, both inside the scientific community but also from the outside (Pereto, 2005). Modern evolutionary biologists have been proving Darwin's theory by several means, confirming that the first species reproduced and continuously evolved in a network of species (represented more commonly as a tree, the Tree of Life) ultimately giving way to modern species that we can observe today. However, it's not proved how life could evolve from non-life. The discovery of the first viruses brought the proposal that they would represent the missing link in the origin of life (Haldane, 1929). However, it has been shown that viruses cannot be included in the tree of life for some reasons that are worth of attention: viruses are polyphyletic (they have various evolutionary origins); their infection of distant hosts does not imply antiquity; some of them have metabolic and translation genes but these were proved to have been acquired from hosts (Moreira &

Lopez-Garcia, 2009). Viruses are accepted today as a reflection of a concept for long neglected – regressive evolution, the major process of parasite evolution. However, it remains to be explored the parasitic nature of viruses as a resemblance of an ancestral domination by RNA and/or DNA molecules of protein networks that might have been encapsulated with ribosomes and tRNA.

Did life start with replicative molecules, or a primitive auto-catalytic metabolic network? What we know for sure is that it started for long ago, at least before 3.5 billion years ago (See Fig. 4 and (Battistuzzi et al, 2004)). There is very little geological and fossil information from that period in Earth history, as plate tectonics is very dynamic considering these time spans. Biologists need to gather up with geologists and geochemists to understand what was going on chemically on Earth in the period of time when we know life begun. An excellent review on that matter was (Nisbet & Sleep, 2001). It is commonly assumed that early organisms inhabited an environment rich in organic compounds that was spontaneously formed in the prebiotic Earth. This theory is frequently referred to as the Oparin–Haldane theory or heterotrophic origin of life. Minimum biosynthesis is accepted within this theory; the first living systems would have originated directly from the primordial soup and evolved relatively fast up to a common ancestor, usually referred to as LUCA (Last Universal Common Ancestor). Defining the nature of LUCA is one of the central goals of the study of the early evolution on Earth; several attempts have been made in this direction but the nature of this common ancestor is still under debate (Fani & Fondi, 2009). It is not the purpose of this dissertation to analyse the complex theories of the origin of life; instead we focus on minimal metabolic networks that might have arisen first in the prebiotic environment and their evolution to the extant metabolic networks.

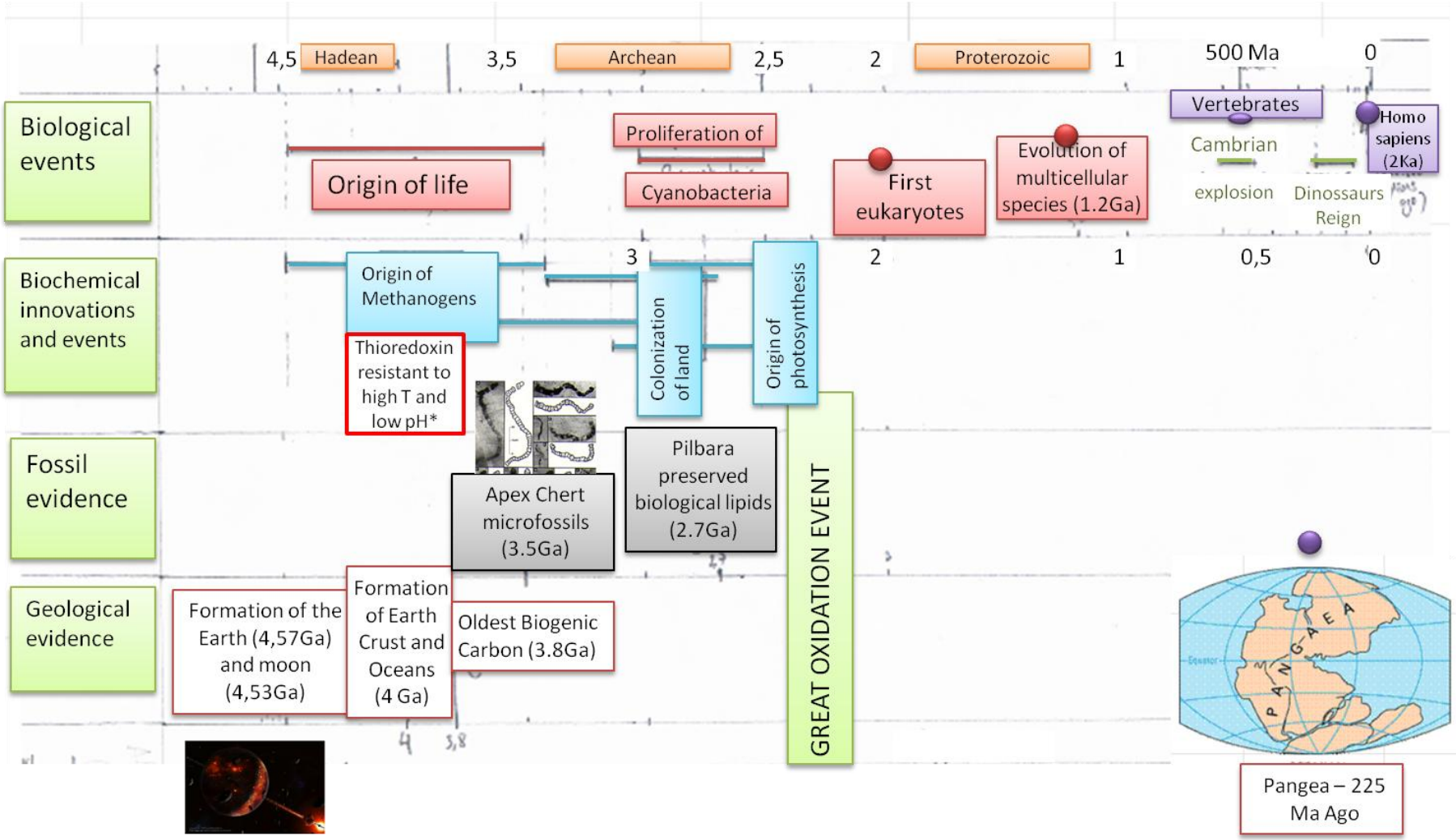


Fig. 4 – Timeline of the major events known in the history of Life on Earth. In Blue, events from (Battistuzzi et al, 2004): Red box, (Perez-Jimenez et al, 2011)

2.2.2 – Primordial metabolism

The different metabolic and ecological abilities of organisms dictate if their genetic record prevails, or not, in species evolution. The question of how metabolic and genetic functions got so intricately intertwined in a common ancestor was considered fundamental to Biological sciences (Pereto, 2005). Miller and Urey proved with their famous experiment that amino acids can form under conditions that resemble primitive Earth (Miller, 1953). A recent experiment also showed how RNA pyrimidine ribonucleotides can form in prebiotic conditions (Powner et al, 2009).

It is assumed that in an early Earth populated with prebiotic chemical systems based on minimal metabolism, the depletion of essential nutrients imposed a selective pressure in these "micro-systems" to be able to synthesize those nutrients (Fani & Fondi, 2009). Therefore, the evolution of metabolic pathways represents a crucial step in molecular and cellular evolution. In fact, metabolism is one of the most conserved cellular processes; it is recognized that very little is known about how the chemistry of primitive enzymes arose (Perez-Jimenez et al, 2011) and which were the first enzymes appearing. However, it has been shown that enzymes were active on earth already around four billion years ago (4.000.000.000 years, more or less 62 times the time passed since dinosaurs' predicted extinction, 65.000.000 years ago).

In today's life forms, enzymes act in complex **networks**. The approach to such a complex question as how were constituted these networks first must be systematic; one way is to analyse extant networks, of organisms that live today, trying to minimize their features while they are still living. A Systems Biology computational approach analysed, recently, theoretical properties of inferred prebiotic networks (Shenhav et al, 2005). The authors concluded that molecular ensembles with high complexity may have arisen very early in life's evolution.

Comparative analysis of the metabolism has the potential to study ancestor's metabolic networks; however, the state of the art in metabolic network alignment is still incipient compared to genetic alignment (see Section 2.5).

Systems Biology Genome-scale metabolic reconstructions (GENREs) are representative of essential metabolism for cell growth and division. In this work the power of these new tools

to infer minimal/common metabolic essential reactions that might reflect ancient important steps in metabolism evolution was analysed. The identification of these central enzymes or "hubs" can also be important for other evolutive studies, for example with comparative genomics, as the already referred breakthrough work published in Nature this year (Perez-Jimenez et al, 2011). More than reconstructing the sequences of thioredoxins as ancient as 4 billion years old (see Fig. 4 and notice how early in life's history enzymes were already existant and active); the authors also reconstructed these and tested them in the laboratory. These enzymes were shown to be highly resistant to Temperature and low pH, concordantly with what is known from Earth's environment at that time (Perez-Jimenez et al, 2011).

2.3 Systems Biology and Genome-Scale Metabolic Models

In the last couple of decades, Life's Sciences and Technologies have improved exponentially, in a cyclic, mutually dependent way. The advances in Biotechnology have opened way to fast and high-throughput experimental results. A recent and excellent review shows how the era of the new "-omics" is confirmed (Joyce & Palsson, 2006); the recently established Systems Biology should deal with the disciplines comprised in it (see Table 1). Systems Biology is seen today as non-converging with classical Biology, as it can discover its own fundamental quantitative laws, approaching physics as a science (Westerhoff et al, 2009); metabolic networks constitute one of the preferential fields for Systems biology, especially in the case of the relatively simple prokaryotic organisms.

Metabolic engineering followed up Systems Biology development closely, but it is also accepted that the power of available molecular biological techniques was, in the beginning of the last decade, very far from the ability to rationally analyze biochemical networks (Stephanopoulos, 1994). We started by knowing a lot about individual enzymes, metabolites, reactions and genes; on the other side, we couldn't study their interactions, regulations and dynamics in a whole system. This is starting to be possible with many biotechnological tools, among which the use of pure mathematics and bioinformatics to describe systems; these are essentialities of Systems Biology, and of Metabolic Engineering for a long time.

Table 2 – Omics data types and examples of databases

Data type	Online database	Web page address	Features
Genomics	Entrez Genome Database	http://www.ncbi.nlm.nih.gov/sites/genome	Information on genomes including maps, chromosomes, assemblies, and annotations
	Ensembl	http://www.ensembl.org/index.html	Software system; produces and maintains automatic annotation on vertebrate genomes
	Ensembl Genomes	http://www.ensemblgenomes.org/	Ensembl sister Project for non-vertebrate genomes
	Genomes OnLine Database	http://www.genomesonline.org/	Complete and ongoing genome projects, as well as metagenomes and metadata
	ENA genomes server	http://www.ebi.ac.uk/genomes/	Repository of completed genomes
	Microbial Genome Database (MBGD)	http://mbgd.genome.ad.jp/	Japanese database for comparative analysis of completely sequenced microbial genomes
Proteomics	ExPASy Proteomics Server	http://expasy.org/	Sequences, structures, and 2-D PAGE analysis
	PRoteomics IDEntifications database (PRIDE)	http://www.ebi.ac.uk/pride/	Public repository for protein and peptide identifications and evidence supporting these identifications
	Global Proteome Machine Database	http://gpmdb.thegpm.org/	Database constructed to validate the peptide MS/MS spectra as well as protein coverage patterns
Transcriptomics	Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/	Public functional genomics data repository; tools available to query and download experiments and curated gene expression profiles
	Stanford Microarray database	http://smd.stanford.edu/	Data from microarray experiments, and the corresponding image files
	ArrayExpress Archive	http://www.ebi.ac.uk/arrayexpress/	Database of functional genomics experiments
Lipidomics	LIPID Metabolites And Pathways Strategy (LIPID MAPS)	http://www.lipidmaps.org/	Free, comprehensive genome-scale lipids database
Localizomics	Pathway Localization database (PathLocdb)	http://pathloc.cbi.pku.edu.cn/	Repository of subcellular localizations of metabolic pathways and their participant enzymes
	Yeast GFP Fusion Localization Database	http://yeastgfp.yeastgenome.org/	Global analysis of protein localization studies in *S. cerevisiae*
Metabolomics	The Human Metabolome Project	http://www.metabolomics.ca/	Database containing detailed information about small molecule metabolites found in the human body
	Metabolomics Fiehn Lab Databases	http://fiehnlab.ucdavis.edu/db/	mass spectra, retention indices, structures and links to external metabolic databases for over 1,000 identified compounds
GENRE's	BiGG	http://bigg.ucsd.edu/bigg/home.pl	Knowledgebase of Biochemically, Genetically and Genomically structured genome-scale metabolic network reconstructions

Metabolic engineering designs new metabolic potentials or improves existing ones for existing cells through the construction of mathematical models (Fig. 5); when these models are too complex (many reactions involved) the computer is used to facilitate calculations, so it is called an *in* silico model (Patil et al, 2004).

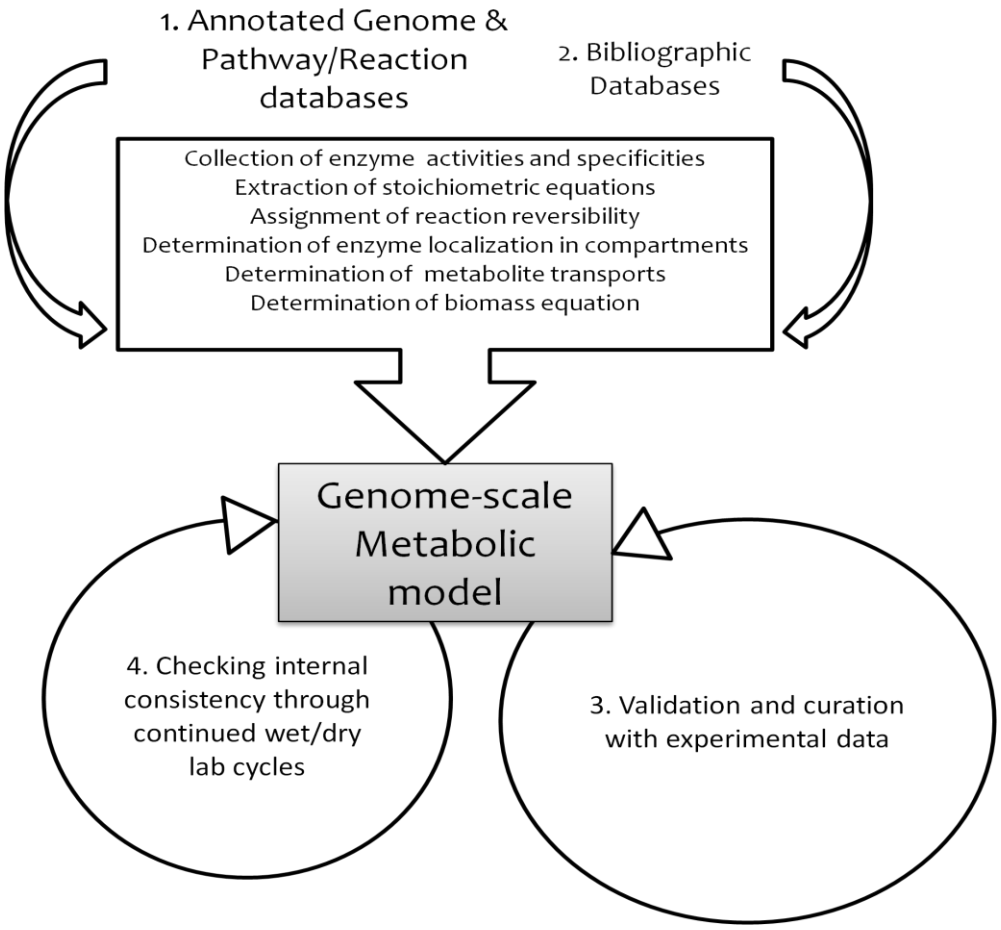


Fig. 5 – Steps in the construction of a genome-scale Metabolic Model.

Although the knowledge of the regulatory functions (genetic and through protein interactions) is still in its beginning, the topology of the biochemical network of metabolic reactions (enzymes, metabolites and their interactions) is very well characterized, in genome-scale models, for example for *Escherichia coli* (Edwards & Palsson, 2000a; Feist et al, 2007; Reed et al, 2003; Vickers et al, 2011) and yeast (Duarte et al, 2004; Forster et al, 2003; Nookaew et al, 2008). Many biological networks are described as scale-free, which means their topology is dominated by a few highly connected nodes (hubs) which link the rest of the less connected nodes to the system (Jeong et al, 2000); however, hierarchical structures have been proved to be best suited for capturing most of the features of metabolic networks (Ravasz et al, 2002) (Fig. 6); in hierarchical networks, there is a scale-free topology with embedded

modularity. In both kinds of networks any two nodes in the system can be connected by relatively short paths. Due to this interconnected nature of metabolic networks, all steps in the metabolism can (in principle) influence all other steps. Therefore, understanding, simulating and predicting both dynamic and steady state operations of metabolic network are challenging tasks.

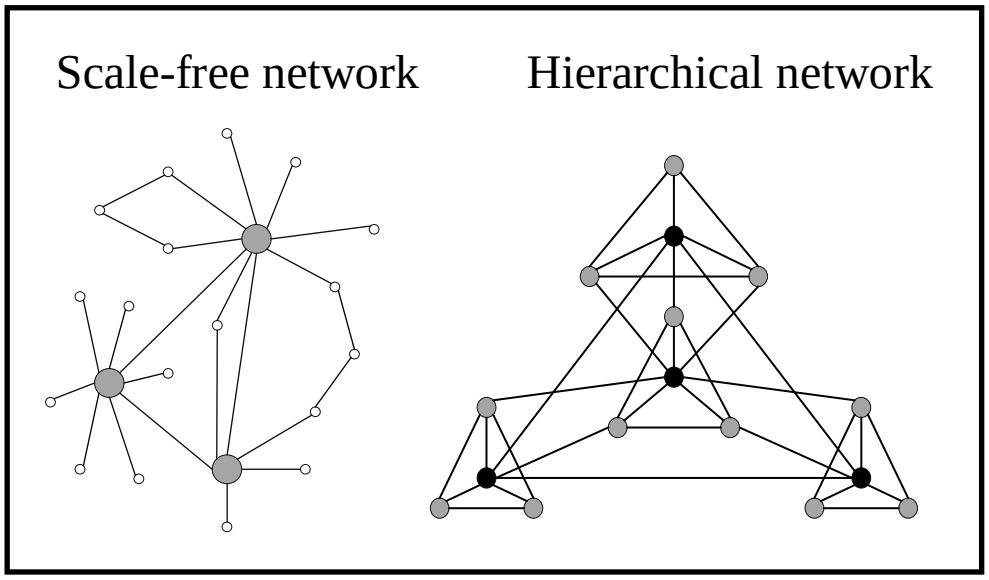


Fig. 6 – Different topologies of biological networks. Scale free networks follow a power law-like distribution. Hierarchical networks allow hubs and modular structures to be embedded in the scale-free topology**.**

2.3.1 Building a genome-scale metabolic model

To date all of the genome-scale metabolic models have been built manually in a four-step process (Oberhardt et al, 2009) (Fig. 5): (1) Initial reconstruction built from gene annotation data coupled with information from databases that store various enzyme data such as ligand molecules (cofactors, substrates, products, inhibitors and activators), reaction formulae and metabolic pathways: KEGG (Kanehisa et al, 2006), EXPASY (Gasteiger et al, 2003), BRENDA (Chang et al, 2009), EcoCyc (Keseler et al, 2009) etc; (2) examination of the primary literature to improve the initial reconstruction and conversion of all the knowledge achieved to a mathematical model (that will be possible to analyze with a constraint-based approach – see next chapter); (3) validation of the model through comparison of its predictions to phenotypic information; (4) submission of the model to continued wet/dry lab cycles improving its

accuracy and testing hypotheses. Flux balance analysis is the basis of the constraint based approaches able to deal with genome-scale models.

2.4 Flux Balance Analysis in genome-scale metabolic modelling

Two main categories of techniques exist to analyze metabolic models – kinetic/mechanistic and stoichiometric; however, the last one was obligatory to deal with genome-scale models until very recently (Smallbone et al, 2010). This happens because kinetic models take in consideration all the parameters for each of the enzymatic reactions in the cell, and in fact, one reaction can have several kinetic parameters (Smallbone et al, 2007), and even in the best-understood organisms the majority of kinetic parameters remain undetermined (Terzer et al, 2009). The way metabolic engineering found to deal with the problem was by using mathematical constraints that don't change much over time and are much easier to identify than kinetic parameters. Constraint-based methods use physicochemical constraints such as mass balance, energy balance and flux limitations, basic concepts in metabolic engineering (Stephanopoulos, 1998). For further understanding these approaches, it is indispensable to define clearly the term "flux". In metabolic context, flux for a certain reaction refers to the amount of substrates processed (or products produced) per unit time, for example, ethanol produced per unit time per unit biomass. The integrated effect of fluxes in the course of different reactions in a network has as a consequence the phenotype of the cell. As all of the constraints on a cellular system are never completely available, multiple steady-state solutions are possible, so there is the need to identify a biologically viable steady state (Fig. 7).

Flux Balance Analysis (FBA) is one of the techniques used to predict phenotypes from network reconstructions; through stoichiometric constants, it employs a linear programming (LP) strategy to generate a flux distribution that is optimized toward a particular so-called objective/ phenotypical goal (See FBA – Foundations of the Technique – step IV). This strategy was introduced on the basis of the Darwinian principle that states organisms' optimization during evolution (Ruppin et al, 2010; Varma & Palsson, 1993a).

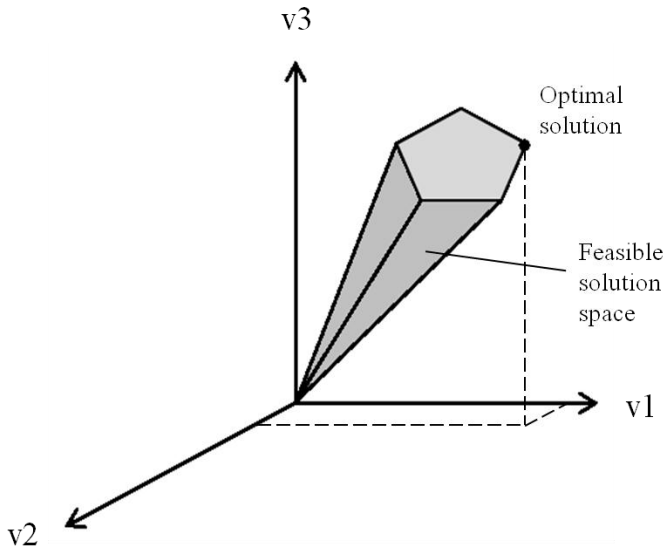


Fig. 7 – Mathematical calculation of flux distributions (solution spaces) and optimal fluxes (optimal solutions) requires the definition of constraints
Without constraints any distribution of the three fluxes is possible within the solution space. When constraints are applied (as the stoichiometric matrix S or cell capacity constraints (flux upper and lower bounds) a feasible solution space is defined. When an objective is defined (e.g. maximum growth rate, maximum ATP production) an Optimal Solution (optimal flux distribution) can be identified.

In the last few years flux-balance analysis (FBA) has been the most successful and widely used technique for studying metabolism at system level (Braunstein et al, 2008) as it makes possible to predict the biomass yields of organisms or the product yields for a biotechnologically important metabolite (Orth et al, 2010) easily. Table 2 (adapted from (Kauffman et al, 2003)) shows some of the main events in the history of the development of this technique.

Table 3 – Significant events in the development of the FBA technique

Year	Significant events in FBA History	Reference
1984	Papoutsakis used linear programming to calculate maximal theoretical yields	(Papoutsakis, 1984)
1986	Fell and Small used linear programming to study lipogenesis	(Fell & Small, 1986)
1990	Majewski and Domach studied acetate overflow during aerobic growth	(Majewski & Domach, 1990)
1992	Savinell and Palsson performed detailed analysis and development of FBA theory	(Savinell & Palsson, 1992a; Savinell & Palsson, 1992b)
1993	Varma and Palsson used FBA to describe *E. coli* properties	(Varma et al, 1993b; Varma & Palsson, 1993a; Varma & Palsson, 1993b)
1997	Pramanik and Keasling studied growth rate dependence on biomass concentration	(Pramanik & Keasling, 1997; Pramanik & Keasling, 1998)
2000	Edwards and Palsson carried out a gene deletion, phase plane, robustness study of *E. coli* Lee et al. identification of alternative optima Schilling et al. integrated FBA with extreme pathway analysis	(Edwards & Palsson, 2000a; Edwards & Palsson, 2000b) (Lee S, 2000) (Schilling et al, 2000)
2001	Burgard and Maranas examined performance limits of *E. coli* and minimal reaction sets Covert, Schilling and Palsson added regulatory constraints to FBA models	(Burgard & Maranas, 2001) (Covert et al, 2001)
2002	Papin et al. studied network redundancy in *Haemophilus influenzae* (alternate optima) Ibarra, Edwards and Palsson looked at adaptive evolution of *E. coli* Mahadevan, Edwards and Doyle studied dynamic FBA Beard, Liang and Qian considered the addition of energy balance constraints to FBA	(Papin et al, 2002) (Ibarra et al, 2002) (Mahadevan et al, 2002) (Beard et al, 2002)

2.4.1 Flux balance analysis: Foundations of the technique

One very recent, complete review, explains in a very user-friendly way how to prepare a Flux Balance Analysis *in silico* experiment (Orth et al, 2010). The steps explained in the next paragraphs are based on this review, with more information added.

2.4.1.1 Model Construction/System definition

All of the metabolic reactions and metabolites should be defined and included in the model. Although the pathways will be regulated, depending on the environmental context with the definition of only a subset of reactions happening at the time, FBA doesn't take in consideration the regulation explicitly (Kauffman et al, 2003). Another important consideration at this point is the inclusion of all the transport mechanisms (or exchange reactions) – compounds that diffuse through the membrane, pores in the membrane or active transports across the membrane.

2.4.1.2 Mathematical representation

All the reactions previously defined are now put into a stoichiometric matrix (normally labelled **S**); each column represents a reaction, each row represents a metabolite. An extra column can be added to represent the biomass production, for example. The flux vector is normally labelled **v**, and represents the unknown flux distribution to be determined.

2.4.1.3 Mass Balance

Flux balance analysis considers the cell as in steady state, which allows the equality

$$\mathbf{S.v=0} \qquad (1)$$

that automatically constrains the solution space (Fig. 7 and Fig. 8) and defines a system of linear equations. The key challenge here is that most of biological systems contain more reactions than metabolites (Gianchandani et al, 2010), which makes them undetermined. So at

this point, the model contains an undetermined system of linear equations. One way to obtain additional constraints is to define upper and lower bounds for some fluxes (Fig. 7)

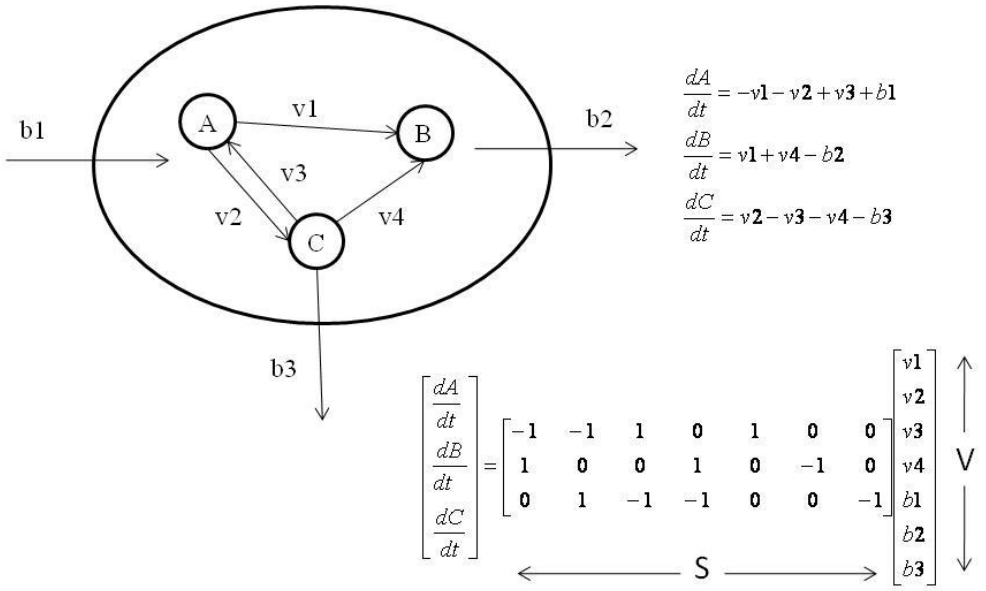


Fig**. 8 – Initial steps for FBA (Based on (Kauffman et al, 2003)).** A model System with three metabolites, A, B and C is represented. Internal fluxes are vi and external fluxes are bi. All the flux mass balance equations are represented both in a system of equations and in the matrix form. At steady state the variation of concentrations of each metabolite is zero, the equation S.v=0 is possible; notice that the length of V is bigger than the number of columns in S, which makes this an undetermined system at this point.

2.4.1.4 Optimization/ Definition of the objective function

Optimization is the alternative used by FBA to estimate one exact internal flux distribution of the cell. This is done assuming that the metabolism of the cell is optimized to a certain objective. This opens a lot of doors, allowing questions like ‘what are the fluxes when the cell wants to reproduce the maximum it can?’ Or ‘which are the larger fluxes when the cell is optimized to produce a certain antibiotic, and is it possible to enhance that reaction to improve the process income?’

If the objective function defined is linear, the optimization problem becomes then a Linear Programming (LP) problem (Kauffman et al, 2003); FBA simulations that used growth flux/biomass equation as the objective function have been proven to be consistent with experimental data (Edwards et al, 2001; Ibarra et al, 2002; Segre et al, 2002).

The objective function **Z** can be described by the equality:

$$Z=c^{T}v \qquad (2)$$

where **c** is a vector of weights for each of the fluxes **v** that defines how each individually contributes to the objective. In the case of biomass being the objective function, **Z** is equal to the flux of this reaction only (the other coefficients in the vector are zero). Other objective functions have been used in literature, as maximizing ATP production (not only for bacteria (Teusink et al, 2006) but also for a myocardial model (Luo et al, 2006) and mitochondrial models (Ramakrishna et al, 2001; Vo et al, 2004)), maximizing or minimizing the production of a particular metabolic product (Blank et al, 2008; Izallalen et al, 2008; Lee et al, 2007; Portnoy et al, 2008; Varma et al, 1993a) and maximizing or minimizing the rate of nutrient uptake (Cakir et al, 2007; Kim et al, 2007).

2.4.1.5 Flux Calculations by Linear Programming

The flux distribution that optimizes the objective function is identified with Linear Programming within the space of allowable fluxes. The point that corresponds to the desired flux distribution is situated on an edge or corner of the solution space. It is often possible that more than one solution leads to the optimization of the objective function. In this case such alternate solutions can be identified with FVA (Flux variability analysis) a method derived of FBA that maximizes and minimizes all the reactions in the network (Mahadevan & Schilling, 2003); another way to deal with this problem is using MILP (mixed-integer linear programming-based algorithm) (Lee S, 2000; Phalakornkule et al, 2001).

2.4.1.6 Computational resources

To analyze the huge amount of information in a genome-scale metabolic model (hundreds of metabolites and thousands of reactions), strong computational resources and statistical techniques are needed. The COBRA Toolbox (Becker et al, 2007) is a freely available Matlab toolbox that can be used to perform a variety of FBA-based methods. Another software relatively simpler than Matlab has been used for FBA calculations – R, a statistical computing free software (R Development Core Team, 2009). Recently, a Portuguese team within the

authors group developed OptFlux, a user-friendly, free and very complete software for FBA calculations (Rocha et al, 2010); in the paper of the release a complete table compares this tool with other available tools (Rocha et al, 2010), including COBRA.

2.4.2 After basic FBA: 2nd generation FBA models

The basic FBA approach has several limitations. It is only suitable for steady-state predictions, and as a result it doesn't predict metabolite concentrations. Also, usually it doesn't take in consideration regulatory effects that can suffer a high impact after single mutations and even more in a synergistic multiple-mutant. Basic FBA may not be sufficient to study minimal metabolism, and the improvements/extensions it has suffered were worth a review for possible directions to take in the practical course of this thesis.

2.4.2.1 Determining fluxes in mutants – MOMA and ROOM

One of the earliest extensions of FBA was the development of the Method of Minimization of Metabolic adjustment – MOMA (Segre et al, 2002). This innovative approach questions that the assumption of optimality for a wild-type microbe is acceptable, but the same argument may not be valid for genetically engineered knockouts or other bacterial strains that were not exposed to long-term evolutionary pressure. MOMA is based on the same stoichiometric constraints as FBA, but relaxes the assumption of optimal growth flux for gene deletions. In this manner, MOMA states that the knockout mutant should display a suboptimal flux distribution, an intermediate between the wild-type optimum and the mutant optimum (Fig. 9). Classical FBA would assume the knock-out optimum, but MOMA calculates this intermediary state, the flux minimal response to the perturbation, which requires quadratic programming (QP).

Gene knockout simulations based on MOMA were experimentally proved being capable of refining candidate mutations for increasing the production level of lycopene, an efficient anti-oxidant compound in *E. coli* (Alper et al, 2005). MOMA can be, in this way, particularly relevant for characterizing initial transient behaviour after gene manipulation

(illustrating this hypothesis, MOMA predicts point c in Fig. 9, but it is expected that the cell will adapt to reach the point b).

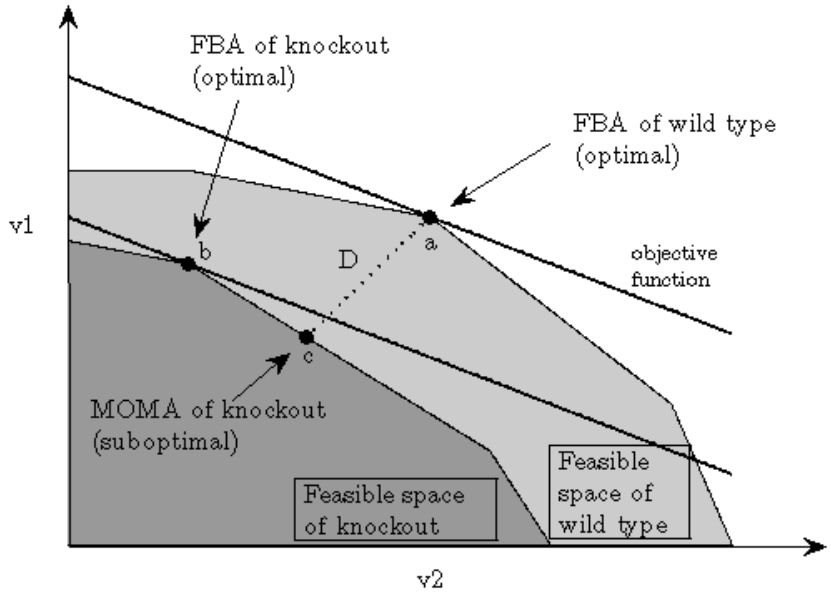


Fig. 9 – Schematic representation of the feasible solution space for the flux distribution of wild type and knockout strains, with MOMA suboptimal solution. Wild type - light+ dark grey; knock-out - dark grey. Coordinates are two arbitrary representative fluxes. Black lines -objective function; a – optimal FBA LP prediction for WT; b – optimal FBA LP prediction for knock-out; c – alternative MOMA knock- out solution calculated through Quadratic Programming, also the projection of the FBA optimum into the feasible space of the mutant. The b and c points are in general distinct. Adapted from (Segre et al, 2002).

ROOM (regulatory on/off minimization) is a recent development of FBA for mutant strains that identifies the metabolic flux state by minimizing the number of significant flux changes from the wild-type flux distribution (Shlomi et al, 2005). The significance is defined by a set of threshold values the user specifies, considering manageable computation time. ROOM allocates a cost to the expression of a gene and finds the feasible flux distribution with a minimal number of Boolean flux changes from the flux distribution of the wild-type strain. The resulting optimization problem involves mixed integer linear programming (MILP). ROOM accurately predicted the post-adaptation states (flux distribution after a significant period of time) for *E. coli* metabolism (Shlomi et al, 2005).

ROOM shares with MOMA the calculation of a flux distribution for mutants based on a minimal adjustment metric. The main difference between these two approaches is in the motivation behind them: MOMA has a mathematical origin in the formulation of a minimal

response to perturbations, while ROOM uses a more qualitative, biological approach to control gene expression, assuming that a cell, in the long run, tries to minimize the number of significant flux changes (Terzer et al, 2009).

2.4.2.2 Considering the evolution of fluxes with time - DFBA

Dynamic Flux Balance Analysis (DFBA) (Mahadevan et al, 2002) presents a new analysis frame for the quantitative study of the dynamic reprogramming of metabolic networks; it's more quantitative than rFBA. DFBA was introduced in two distinct formulations, a dynamic optimization approach that required solving a nonlinear programming (NLP) problem, and a static optimization approach that required solving an LP problem (Mahadevan et al, 2002). The NLP approach performs a single optimization spanning the time period of interest and yields temporal profiles of fluxes with correspondingly metabolite concentrations (Mahadevan et al, 2002). The static approach separates the time window into uniform time intervals, solving at the beginning of each interval one instantaneous optimization problem to predict the fluxes at that time point, followed by integration over the interval to compute species concentrations over time.

MOMA and DFBA were coupled in a more recent study (already referred as using the maximum ATP production as the objective function) in a strategy called M-DFBA (Luo et al, 2006). The myocardial metabolic network was studied altering the objective function from maximization of ATP production to minimization of fluctuation of metabolite concentrations between normal and ischemic conditions. The results with the coupled techniques showed greater consistency with experimental data than those obtained with DFBA alone, but also supported the hypothesis that metabolic systems relapse to suboptimal states during transient perturbations.

2.4.2.3 Including thermodynamic restrictions - EBA

Reaction thermodynamics were incorporated in an extension of FBA called energy balance analysis (EBA) (Beard et al, 2002). EBA eliminates thermodynamically non-viable results associated with FBA by incorporating energy-balance loop equations analogous to Kirchhoff's voltage laws for electrical networks(Beard et al, 2002). Equations balancing the global potential energy of a network are overlaid on the underlying network stoichiometry, and inequality constraints for each flux are also added to ensure that entropy production is

positive for every reaction. EBA requires the solution of a nonlinear optimization problem that estimates the growth rate and the intracellular metabolic fluxes. The resulting feasible solution space is a subset of the space predicted by a traditional FBA; however, as it uses nonlinear optimization, EBA doesn't ensure an optimal solution. For *E. coli* metabolism it was shown that the combination of energy balance analysis with FBA gave the same optimal growth rate, but the observed fluxes were substantially different (Beard et al, 2002). The energy balance analysis was also able to explain why certain genes that FBA identified as nonessential were in fact essential — major changes were required in the observed fluxes to compensate for these knockouts (Beard et al, 2002).

2.4.2.4 Including transcriptional regulation - rFBA

Gene regulatory constraints have already been incorporated into metabolic models, leading to a modification of FBA called regulatory flux balance analysis (rFBA) (Covert & Palsson, 2002; Covert & Palsson, 2003; Covert et al, 2001). These constraints were imposed as Boolean operators representing temporary flux constrictions that arise due to a specific environment (which genes are on or off in response to a specific signal); this kind of constraints should be assimilated from literature and overlaid on an existing stoichiometric model of metabolism (Covert & Palsson, 2003). rFBA predictions yielded 91% agreement with experimental measurements for *E. coli* genome-scale metabolism, up from 86% for FBA without the gene regulatory constraints (Covert & Palsson, 2003).

More recently, another extension of FBA called integrated FBA (iFBA) was developed to build an FBA model of *E. coli* central metabolism (Covert et al, 2008). The integrated model described biomass, 77 metabolites, 151 genes and 113 reactions, and it combined a kinetic model of *E. coli* phosphotransferase (PTS) catabolite repression (Kremling et al, 2007) with an rFBA model of the same system (Covert & Palsson, 2002). The resulting model had significant advantages over either the rFBA or ODE-based models alone, particularly in predicting the consequences of gene perturbation.

2.4.3 Genome-scale FBA based modelling: Industrially and scientifically interesting

FBA has a special application interest for metabolic engineering, as it can predict flux distributions for various gene knockout or knockdown conditions, which allows the identification of specific changes that may facilitate optimal by-product yields. The results can then be used to engineer strains (e.g., through up- or down-regulation of target genes) that have desired phenotypes. Several examples exist, some already referred. One, is a photosynthetic algae, *Chlamydomonas reinhardtii,* that was analyzed by FBA simulation and some changes in hydrogen production rates were correlated to some reaction knockouts (Manichaikul et al, 2009). *E. coli*, as the preferred and first model organism for FBA (Edwards & Palsson, 2000a) was also engineered to overproduce in high yields the amino acids threonine (Lee et al, 2007) and valine (Park et al, 2007), lactic acid (Fong et al, 2005) and succinic acid. (Lee et al, 2005)

Constraint-based models have been experimentally validated for prediction of *E. coli* growth and by-product secretion (Edwards et al, 2001) but also for the prediction of the outcome of adaptive evolution of laboratory strains(Ibarra et al, 2002), with promising results. In the latter innovative study it was shown that under growth selection pressure, the growth rate of *E. coli* in glycerol evolved over 700 generations to achieve the *in silico* whole-cell model predicted optimal growth rate(Ibarra et al, 2002). These results are conclusive and can explain why many times incorrect predictions of *in silico* models based on the optimizing of the objective function may occur. As the optimal states are acquired through an evolutionary process, the authors propose that the non-coincident results may happen because of incomplete adaptive evolution under the conditions tested. This work opens way to the research of possible adaptive evolved states predicted *in silico* and shows the powerful application of FBA in Science as well as in Engineering.

Although its predictions may not always be accurate, Flux Balance Analysis (FBA) is still the reference method for metabolic network analysis and simulation, especially in a large scale as the one of genome-scale models. On its birth, FBA counted only with stoichiometric constraints; since then, many other developments had been made, with the introduction of genetic-regulatory, kinetic and thermodynamic constraints and the development of special approaches to deal with genetically modified/non-adapted strains, among others. The applications these developments have in the analysis of GENREs are vast. Furthermore, the

Industry based on Metabolic Engineering can rely on predictions made by FBA to optimize the production of several metabolites.

GENRE's don't apply only to bacteria; yeast has had quite the same attention by FBA experiments, and at least one attempt of building one MILP-FBA human metabolic network was already done (Shlomi et al, 2008), but there are several other models available (Gianchandani et al, 2010).

Several improvements still need to be done to optimize the FBA technique, and the development of the omics techniques makes promising statements in this area; for example, microarray data was already integrated in an FBA model (Akesson et al, 2004).

Ideally, one would measure all the fluxes directly from the cell, but Fluxomics is one of the poorest described fields in Systems Biology to date. MFA (Metabolic Flux Analysis) is dedicated to this problem, measuring cellular fluxes with isotope-labelling and NMR, GC-MS and LC-MS techniques. MFA has a great deal of information to provide to FBA, but also FBA can indicate interesting targets for MFA. This iterative process already made possible a more accurate calculation of *S. Cerevisiae* flux distributions (Blank et al, 2005).

Probably the big issue in genome-scale metabolic reconstruction approaches is the lack of relevant kinetic information to obtain more precise and specific answers from FBA models; one recent study built a kinetic model using flux distribution obtained by FBA (Smallbone et al, 2010).

In Fig. 10 (Gianchandani et al, 2010) the whole process of stoichiometric network reconstruction and analysis with FBA is summarized. It is clearly shown that Systems Biology is a discipline that relies on a huge variety of knowledge and technology and encourages the interactivity between different fields of Life Sciences.

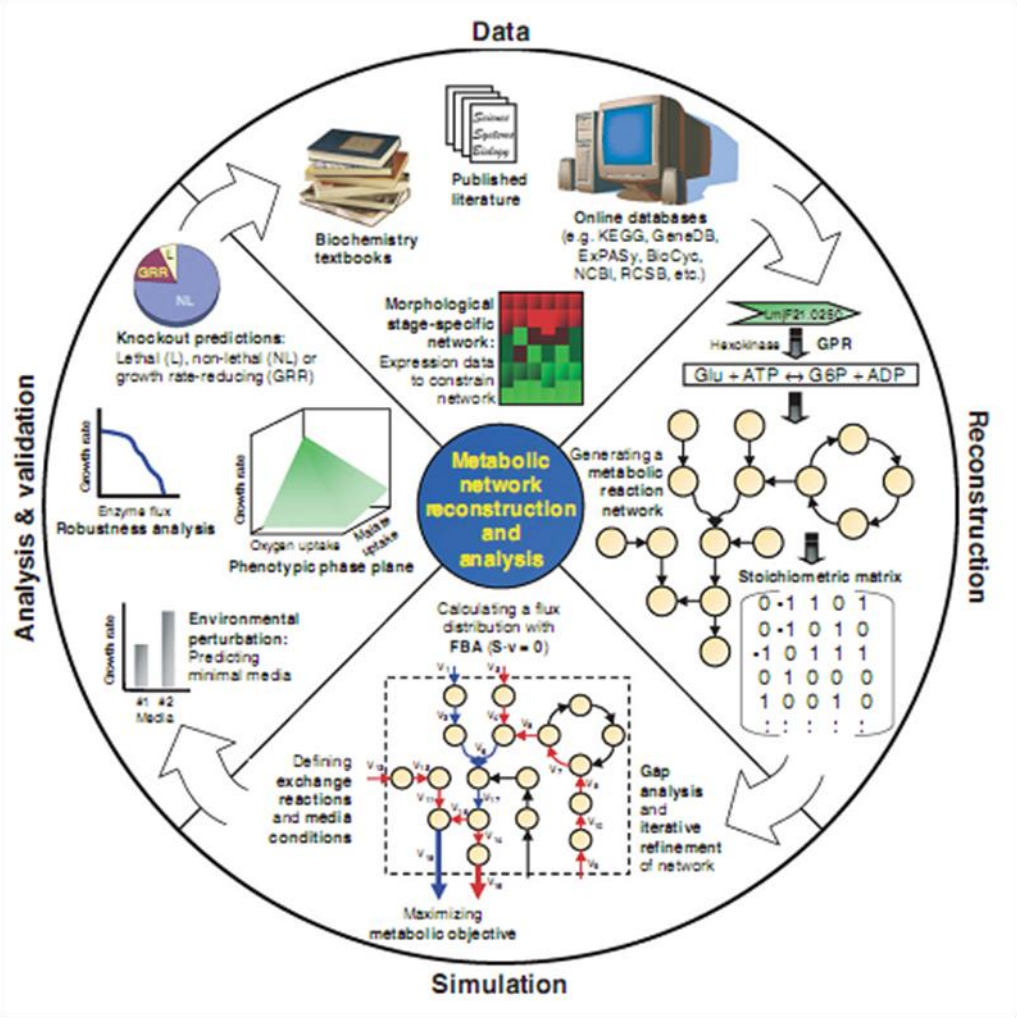


Fig. 10 – Stoichiometric network reconstruction and analysis with FBA (Gianchandani et al, 2010)

2.5 Metabolic network alignment

Metabolic network alignment can be viewed as a systems biology tool when considering multiple-alignment (more than two comparisons, at least). In genetic alignments, this has been achieved with user-friendly tools as ClustalW, its implementation on MEGA5 (Tamura et al, 2011), etc. However, alignment tools for biological networks are only now starting to become user-friendly (Flannick et al, 2006; Kelley et al, 2004). However, just one author has claimed to have developed an algorithm that aligns metabolic pathways in a genome-scale, based solely on topology (Kuchaiev et al, 2010; Kuchaiev et al, 2011). The use of network topology for comparison can give insights into evolution and relationship of metabolic

networks. The conclusions obtained by this method are different from the inferences made when aligning genetic sequences, mainly because similarities at the sequence level do not necessarily correspond to similarities at the network topology level. Metabolic capabilities can be regarded as the result of several evolutionary processes, but metabolism evolution can be interpreted independently from those processes, due to the functional role of metabolic networks. In fact, when comparing organisms, the enzymatic reactions can be the same, while the sequences of the proteins involved are very different. This kind of analysis can also help to overcome the mistaken phylogenetic inferences that exist to date (Kuchaiev et al, 2010; Yamada & Bork, 2009). Also, while genetic sequences give us insights on close phylogenetic relationships (only suitable for recent/conserved proteins) and protein sequences are used to make inferences for older proteins, metabolic pathway comparison can give us insight about even more ancient features, probably existent since the origin of life. Metabolic network alignment has been used and validated to make phylogenetic inferences (Kuchaiev et al, 2010; Ma & Zeng, 2004; Oh et al, 2006), but not yet using GENREs.

Also, we believe there is a much broader range of applications for this methodology such as in expediting model construction/improvement, biotechnology optimization, etc. The tools developed, however, are only now starting to be intelligible to bioengineers and bioinformaticians and not only to programming experts; so, their utilization in answering important biological and biotechnological questions is still in its inauguration. Here, the state of the art in metabolic network alignment was analysed for a possible application towards the goals of this work; however, it is still an incipient area and it was not possible to implement none of the algorithms because of time constraints. We also assumed that the identification of essential pathways is truly beneficial before a posterior multi-pathway, genome-scale alignment. A total genome-scale alignment of the entire metabolism can be an unfruitful task, as these networks are vast and complex (enzymes are hundreds or thousands and metabolites are also on the order of thousands, while genetic/protein alignments only have 4 or 20 entities to account for); edges are normally complex reactions with chemical modifications, with complex nomenclature that is still not homogeneous.

3. Methods and Results

In this section, the practical work of this thesis is described. It started with the analysis of prokaryotic GENREs available to the choice of the case studies for comparative analysis. After, the essential or critical reactions for those models were calculated. Finally, these reactions were compared to identify the common essential reactions. In the end, the most conserved essential reactions were analysed specifically considering primary structures and specific features of the enzymes/proteins associated.

3.1 Analyses of all predictive prokaryotic GENREs and choice of the Case-Studies

The choice of the case studies for essential reaction calculations was based on literature analysis, according to the main goal proposed – approaching the identification of primordial/ basic metabolic functions that allow a cell to survive. Oberhardt, Palsson and Papin's review is probably the best to date about the existent genome-scale metabolic network reconstructions (GENREs) (Oberhardt et al, 2009). From there, we collected the primary information for case study analysis. This review covers 26 prokaryotic GENREs with information regarding their states of validation, through different techniques (Fig. 11)

Fourteen other prokaryotic GENREs were added to the group for further comparison, found in other reviews (Durot et al, 2009; Ruppin et al, 2010) and online-updated Supplementary Table 1 from (Palsson et al, 2009) . Table 4 shows the results of this research, including species and phyla covered, number of genes in the species, models available for that species, validation techniques performed for each model (in the case of models from Oberhardt review), number of genes represented in the model, number of metabolites and reactions, references and the usage or not of the species in another different study that

compares different GENREs (Kun et al, 2008). Fig. 12 shows the distribution of prokaryotic phyla within GENREs and Fig. 13 is the phylogenetic tree of all prokaryotic species with a predictive GENRE (built with iTOL (Bork & Letunic, 2007)).

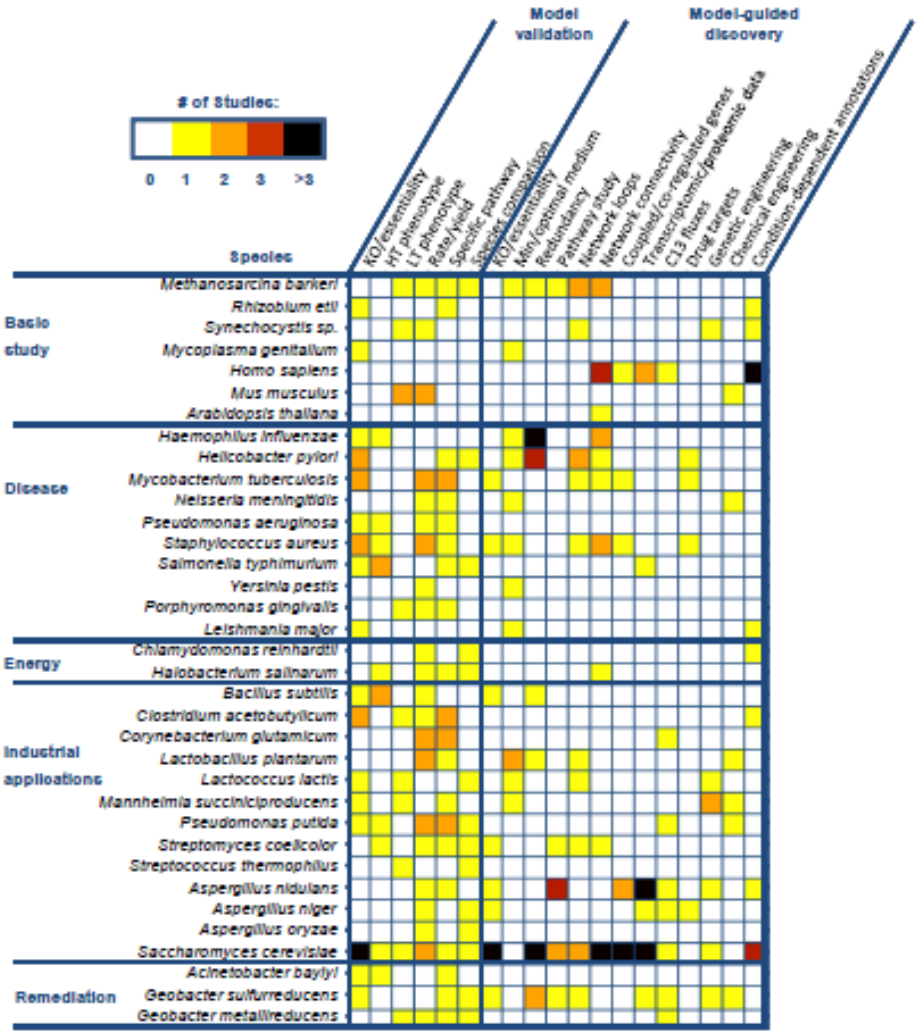


Fig. 11 – Heatmap of metabolic reconstructions showing current validation levels. In (Oberhardt et al, 2009).

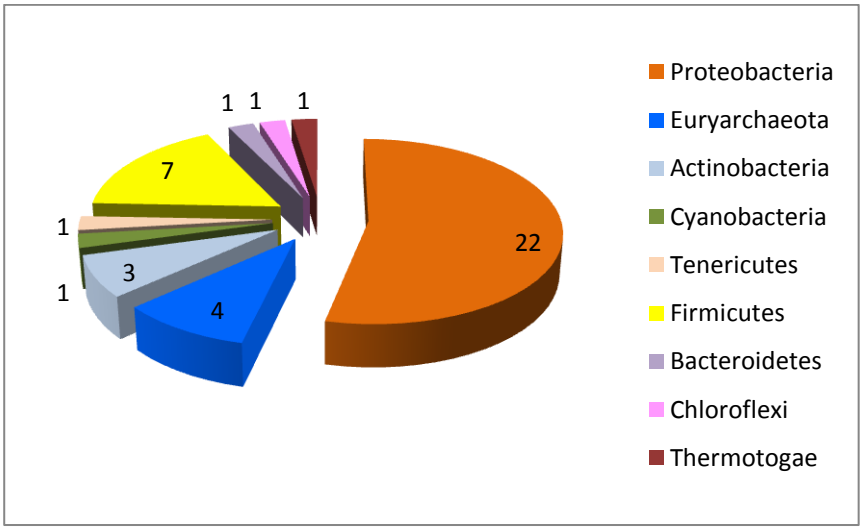


Fig. 12 – Distribution of prokaryotic phyla represented in Genome-Scale Metabolic Reconstructions

Table 4 – Updated list of prokaryotic genome-scale metabolic reconstructions and their main features. Eukaryotic *Sacharomyces cerevisiae* is included for comparison; KO – Knock-out validated model.

	Organism	PHYLUM	Models	KO	# different validation techniques	Gene number/Gene number in model	best relation (previous column)	# metabolites	# reactions	ref.	Kun(08) study
BASIC STUDY	*Methanosarcina barkeri*	Euryarchaeota	**iAF692**		4	**5072/**692	0,14	558	619	Feist, 2006	×
	Rhizobium etli	Proteobacteria	iOR363	X	2	**3168/**363	0,11	371	387	Resendis-antonio, 2007	
	Synechocystis sp.	Cyanobacteria	Kun,2008| Fu,2008| iSyn669| iSyn811		2	**3221/X**;663|669|811	0,25	879|704|790|911	916|831|882 |956	Kun, 2008; Fu, 2008; Montagud, 2010; Montagud, 2011	×
	Mycoplasma genitalium	Tenericutes	**iPS189**	X	1	**521/**189	0,36	274	262	Suthers, 2009	
DISEASE	*Haemophilus influenzae*	Proteobacteria	iJE303; iCS400	X	2	**1775/**296|400	0,23	343|451	488|461	Edwards, 1999; Schilling and Palsson, 2000	
	Helicobacter pylori	Proteobacteria	iCS291|**iIT341**	XX	4	**1632/**291|341	0,21	340|485	388|476	Schilling, 2002; Thiele, 2005	×
	Mycobacterium tuberculosis H37Rv	Actinobacteria	iNJ661|GSMN-TB |**iNJ661m**	XX	6	**4402/**661|726|663	0,16	828|739|838	939|849 |1049	Jamshidi&Palsson, 2007; Beste, 2007;Fang, 2010	×
	Neisseria meningitidis	Proteobacteria	Baart, 2007		2	**2226/**555	0,25	471	496	Baart, 2007	
	Pseudomonas aeruginosa	Proteobacteria	iMO1056	X	4	**5640/**1056	0,19	760	883	Oberhardt, 2008	
	Staphylococcus aureus N315	Firmicutes	**iSB619**|iMH551 |Lee,2009	XX	6	**2588/**619|551|546	0,24	571|604|1431	641|712|1493	Becker&Palsson, 2005; Heinemann, 2005; Lee, 2009	×
	Salmonella typhimurium LT2	Proteobacteria	iRR1083|iMA945| **STM_v1.0**	X	5	**4489/**1083|945|1270	0,28	774|1036|1119	1087|1964 |2201	Raghunathan, 2009; AbuOun, 2009; Thiele, 2011	
	Yersinia pestis 91001	Proteobacteria	iAN818m		1	**4037/**818	0,20	825	1020	Navid&Almaas, 2009	
	Porphyromonas gingivalis	Bacteroidetes	iVM679		3	**2015/X**		564	679	Mazumdar, 2009	
ENERGY	*Halobacterium salinarum*	Euryarchaeota	Gonzalez, 2008		4	**2867/**490	0,17	557	711	Gonzalez, 2008	
INDUSTRIAL APPLICATIONS	*Bacillus subtilis*	Firmicutes	model_v3 |Goelzer, 2008 |**iBsu1103**	X	4	**4114/**844|534|1103	0,27	988|456|1138	1020|563 |1437	Oh, 2007; Goelzer, 2008;Henry, 2009	
	Clostridium acetobutylicum	Firmicutes;	Sanger, 2008; Lee, 2008	XX	6	**3848/**474|432	0,12	422|479	552|502	Sanger, 2008; Lee, 2008	
	Corynebacterium glutamicum	Actinobacteria	Kjeldsen, 2009; Shinfuku, 2009		4	**3002/X**|227	0,08	411|423	446|502	Kjeldsen, 2009; Shinfuku, 2009	
	Lactobacillus plantarum	Firmicutes	Teusink, 2006		3	**3009/**721	0,24	531	643	Teusink, 2006	
	Lactococcus lactis	Firmicutes	Oliveira, 2005	X	3	**2310/**358	0,15	422	621	Oliveira, 2005	×
	Mannheimia succiniciproducens	Proteobacteria	Hong, 2004 | Kim, 2007	X	3	**2384/**335|425	0,18	332|519	373|686	Hong, 2004; Kim, 2007	
	Pseudomonas putida	Proteobacteria	iNJ746;iJP815	X	7	**5350/**746|815	0,15	911|886	950|877	Nogales, 2008; Puchalka, 2008	
	Streptomyces coelicolor	Actinobacteria	Borodina, 2005; Alam, 2010		4	**7825/**700|789	0,10	500|759	700|1015	Borodina, 2005; Alam, 2010	×
	Streptococcus thermophilus	Firmicutes	Pastink, 2009		2	**1889/**429	0,23	???????	552	Pastink, 2009	
	Saccharomyces cerevisiae		iFF708|iND750| iLL672| iIN800| iMH805/775 |iMM904	XXXXX	>9	**6183/**708|750|672 |800|832|904	0,15	584|646|636|1013|1168|713	1175|1149 |1038|1446 |1857|1412	Forster, 2003; Duarte, 2004; Kuepfer, 2005; Nookaew, 2008; Herrgård, 2008; Mo, 2009	×
REMEDIATION	*Acinetobacter baylyi*	Proteobacteria;	iAbaylyiV4	X	3	**3287/**774	0,24	701	875	Durot, 2008	
	Geobacter sulfurreducens	Proteobacteria;	Mahadevan, 2006	X	3	**3530/**588	**0,17**	541	523	Mahadevan, 2006	×
	Geobacter metallireducens	Proteobacteria	Sun, 2009		4	**3532/**747	0,21	769	697	Sun, 2009	
Other models than Oberhardt review											
	Clostridium thermocellum	Firmicutes	iSR432			**3307/**432	0,13	525	577	Roberts, 2010	
	Rhodoferrax/Albidoferax ferrireducens	Proteobacteria	Risso, 2009			**3168/**744	**0,23**	790	762	Risso, 2009	
	Acinetobacter baumannii	Proteobacteria	AbyMBEL891			**3760/**650	0,17	778	891	Kim, 2010	
	Natronomonas pharaonis	Euryarchaeota	Gonzalez, 2010			**2892/**654	**0,23**	597	683	Gonzalez, 2010	
	Buchnera aphidicola	Proteobacteria	iGT196			**574/**196	**0,34**	240	263	Thomas, 2009	
	Chromohalobacter salexigens	Proteobacteria	iOA584			**3352/**584	0,17	1411	1386	Ates, 2011	
	Dehalococcoides ethenogenes	Chloroflexi	iAI549			**2061/**549	0,27	549	518	Islam, 2010	
	Francisella tularensis	Proteobacteria	iRS605			**1802/**683	0,38	586	605	Raghunathan, 2010	
	Shewanella oneidensis	Proteobacteria	iSO783			**5066/**783	0,15	634	774	Pinchuck, 2010	
	Thermotoga maritima	Thermotogae	Zhang, 2009			**1917/**478	0,25	503	562	Zhang, 2009	
	Vibrio vulnificus	Proteobacteria	VvuMBEL943			**2896/**673	0,23	792	943	Kim, 2011	
	Zymomonas mobilis	Proteobacteria	ZmobMBEL601| iZM363			**1808/**347|363	0,20	579|704	601|747	Lee, 2010| Widiastuti, 2011	
	Methanosarcina acetivorans	Euryarchaeota	iVS941			**4540/**941	0,21	708	705	Satish, 2011	
	E. coli K12	Proteobacteria	iJE660|**iJR904**| **iAF1260**			**4405/**660|904|1260	0,29	438|625|1039	627|931 |2077	Edwards, 2000; Reed,2003; Feist 2007	×
	E. coli W (ATCC…)	Proteobacteria	**iCA1273**			**4764/**1273	0,27	1111	2477	Archer, 2011	

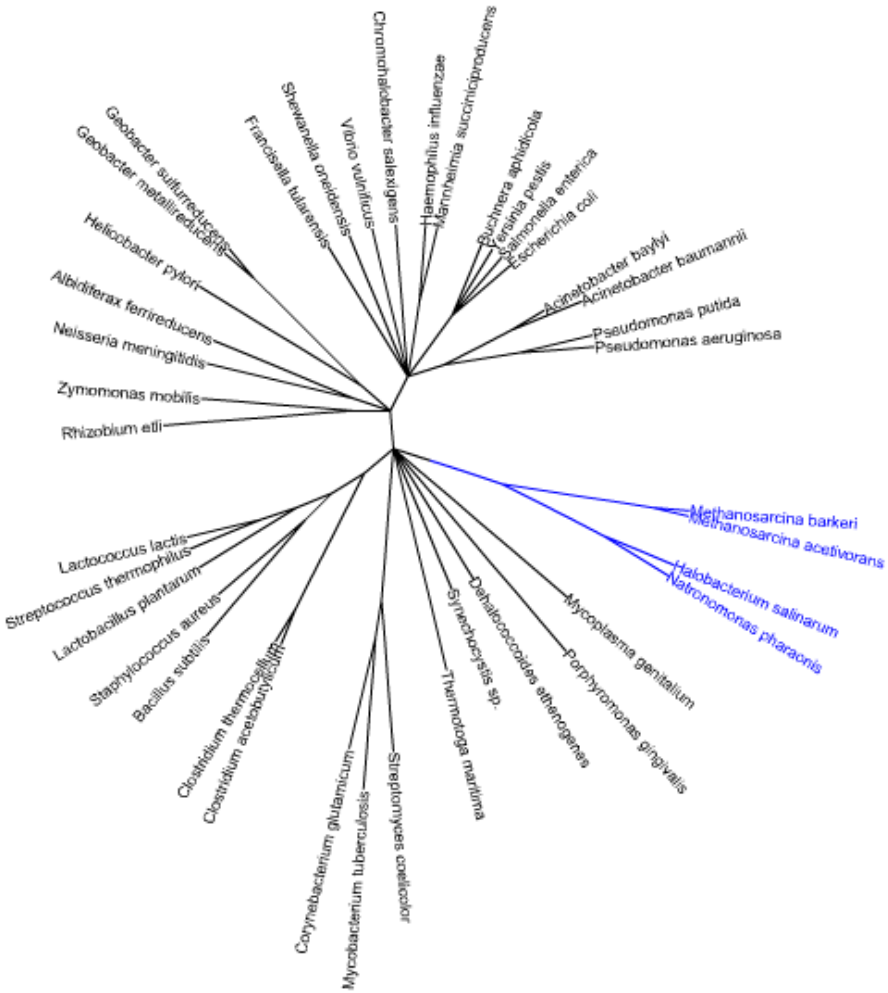


Fig. 13 – Phylogenetic Tree showing the relationships between all Prokaryotic species for which a predictive GENRE exists.
Generated with iTOL (Bork & Letunic, 2007). Species in blue belong to Archaea domain

One of the models chosen was ***Buchnera aphidicola***'s iGT196 (Thomas et al, 2009). This species' genome was until very recently the smallest bacterial genome known (Gil et al, 2002), nowadays exceeded by *Carsonella ruddii* (Nakabachi, 2008) which has impressively only 159,662 base pairs. iGT196 model is very small, with only 196 gene products, 240 compounds and 263 reactions (39% and 27% of *Escherichia coli*, respectively). All reactions were active for simulation, a realistic condition, as this organism almost totally lacks transcriptional regulation.

Representing the Archaea domain, we chose ***Methanosarcina barkeri***, an organism with interest for industry for methane production and basic science also. *M. barkeri*'s model used (iAF692) has 692 metabolic genes with 509 reactions associated, 558 metabolites and

extra 110 reactions with no gene associated. It was the first archaeal genome-scale metabolic model, and also the first methanogen's. The model was validated to predict growth phenotypes and robustness of the network (Feist et al, 2006).

Escherichia coli, the best validated model was also chosen for comparison (scientific, industrial and biomedical interesting). We used ***Escherichia coli* W,** as it was built from *Escherichia coli* K12 most recent reconstruction (the best validated model, see Table 4 and (Palsson & Feist, 2008), and this strain has more genes/ reactions to test for essentiality. It was isolated from soil (which indicates less adaptation specific to host and therefore older metabolic features) and has several advantages for industrial applications (Vickers et al, 2011).

We also used one interesting study (Rodrigues & Wagner, 2009) that tries to overcome the ambiguity of the great carbon sources diversity in prokaryotes. The authors explore the plasticity and robustness of ***Escherichia coli* K12** network, varying the carbon sources, using 101 possible metabolites for this purpose. Still, these 101 represent a tiny fraction of all carbon-containing metabolites in *E coli* - 18.3% (110/550) (Rodrigues & Wagner, 2009). The authors present the reactions from *E coli* K12 that are essential in all cases studied. They start with the iJR904 network (931 reactions) (Reed et al, 2003) and calculate that 67 reactions are essential in all random networks tested. We used this set of 67 reactions for further comparison.

Finally, we also used ***Thermotoga maritima***'s metabolic model (Godzik et al, 2009). This model was done in the context of a breakthrough study which integrated the GENRE with 3D structural information of the proteins of the central metabolic network of this organism, providing a systematic point of view to Structural Biology. *T. maritima* represents the deepest known lineage of Eubacteria with one of the smallest genomes known for a free living organism; it has been the subject of extensive experimental analysis making it an ideal model organism for Systems Biology (Godzik et al, 2009).

3.2 Calculating the essential reactions for biomass formation

3.2.1 Conserved reactions among three GENREs from the three phylogenetic domains

Feist and co-authors (Feist et al, 2006) constructed the first genome-scale metabolic model for an archaeal species, *Methanosarcina barkeri*, at the same time the first methanogen large-scale simulation. Along with the construction of the model, they compared the reaction content with a model of the prokaryotic domain - *Escherichia coli* iJR904 (Reed et al, 2003) and one of the eukaryotic domain *Sacharomyces cerevisiae* iND750 (Duarte et al, 2004). Results are summarized in Fig. 14 (adapted from (Feist et al, 2006)). The models used in this comparison are the best characterized, curated and validated for each phylogenetic domain they represent (Oberhardt et al, 2009) (Table 4). From this comparison, the authors concluded that 211 reactions are conserved across the three species, plus 69 reactions conserved only between Archaea and Bacteria, which gives us 280 reactions conserved across the Prokaryotes.

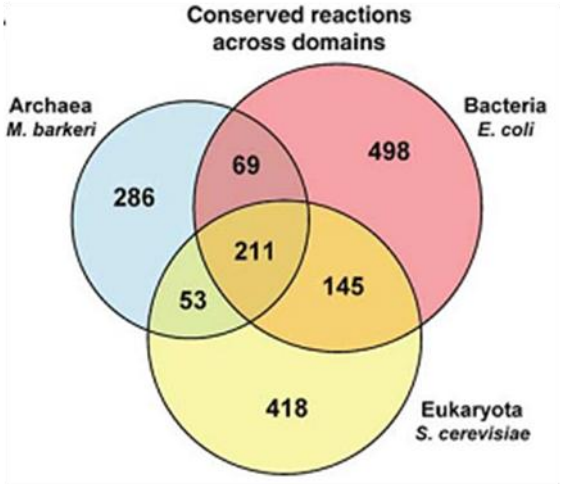


Fig. 14 – Conserved reactions among reconstructed metabolic models from the three domains of life

3.2.2 Calculating essential reactions with Optflux

To obtain the essential reaction sets for further comparison, we used the Optflux software platform (Rocha et al, 2010). OptFlux allows simulations based on Flux Balance Analysis, calculating the Biomass Flux to a desired condition. In our case, we wanted to determine whether the Biomass Flux would be still positive if reaction X was not functioning in the model (if we cut that node from the network the reaction would be non-essential for cell growth) or it would be zero (the reaction would be essential for cell growth). The basic procedure is using the Optimization option in OptFlux – "Calculate Critical Reactions", as follows:

1. Downloading the desired model in SBML format and checking for compatibility with the Optflux platform.
2. Create a new project for the model in Optflux and upload the model in SBML format
3. Check model specificities (environmental conditions, reaction nomenclature, etc)
4. Calculate Critical Reactions
5. Extract reactions (save outside OptFlux) and check for nomenclature – are the abbreviations in agreement with other models used?
6. Compare reaction set with previously calculated ones.

We calculated the critical reactions in Optflux for *Methanosarcina barkeri*¸ *Escherichia coli W*, *Thermotoga maritima* and *Buchnera aphidicola*. Fig. 15 shows the environment of Optflux while running a simulation for calculation of critical reactions.

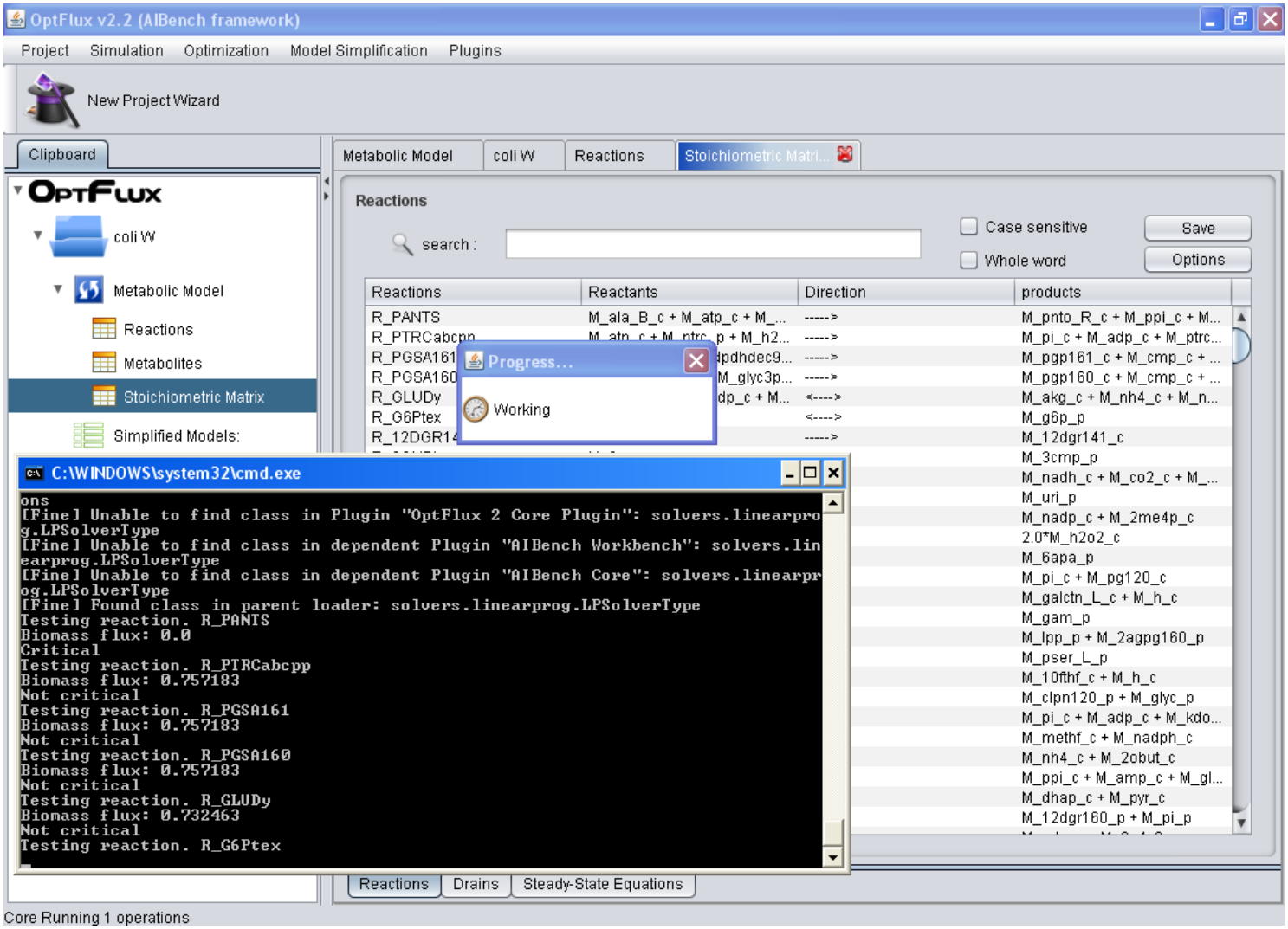


Fig. 15 – OptFlux software platform – Simulation environment. In the figure it's possible to see one operation running, the calculation of the essential reactions for Escherichia coli W. In front the command line window showing the reaction being tested and the biomass flux for the metabolic network with that reaction deleted. In the back the Stoichiometric Matrix is selected; each reaction is represented with the associated reactants and products and possible directions.

3.3 Set comparison and phylogenetic reconstruction

In order to compare the sets of essential reactions obtained previously, we used MS Excel (see file Supplementary Data). We analysed the sets of reactions illustrated in Fig. 16.~

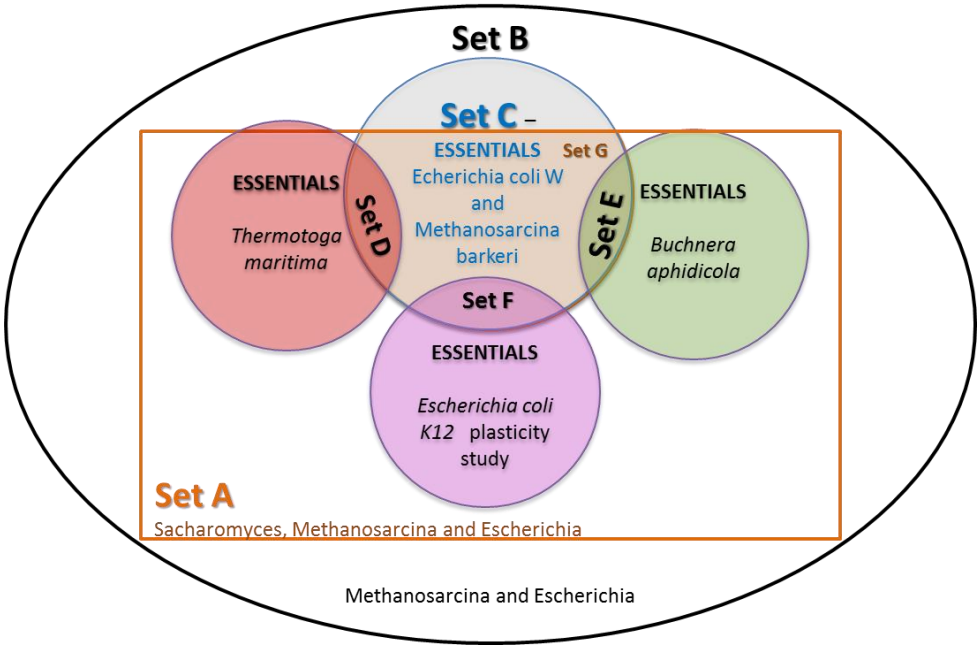


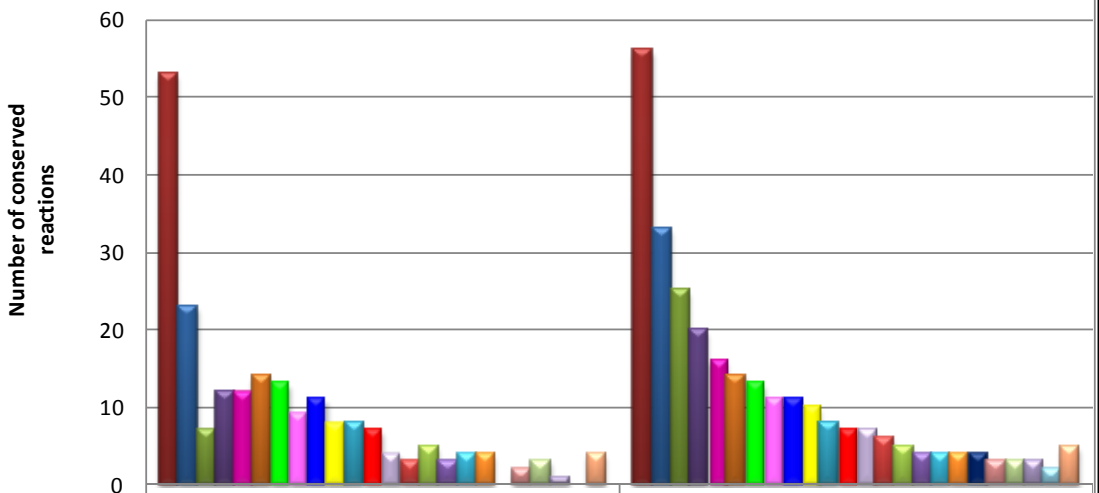
Fig. 16 – Venn's Diagram representing the intercepted sets of reactions in different GENREs analyzed. Set A - Reactions present in the GENREs of *Sacharomyces cerevisiae, Escherichia coli* and *Methanosarcina barkeri* (Feist et al, 2006). Set B - Reactions present in the GENREs of *Escherichia coli* and *Methanosarcina barkeri* (Feist et al, 2006). Set C - Common essential reactions of *Escherichia Coli* W and *Methanosarcina barkeri.* Set D - Set C intercepted with Essential reactions of *Thermotoga maritima*. Set E - Set C intercepted with Essential reactions for *Buchnera aphidicola*. Set F - Set C intercepted with essential reactions calculated in (Rodrigues & Wagner, 2009) – only essential in all networks. Set G - Set C intercepted with Set A. Circles are not contained in the parabola or rectangle, just superimposed on them for the identification of Set G, that doesn't necessarily contain set D, F or E. See Supplementary Data for list of reactions.

The results are summarized in Fig. 17 (metabolic subsystems' distributions from set A and B) and Fig. 18 (metabolic subsystems' distribution in essential reactions from different metabolic networks) and finally, in Table 5, where we list the most conserved essential reactions in all sets analyzed. There were 14 reactions common to all sets studied; the rest of the list of conserved essential reactions in the different networks can be found in Supplementary Data.

To further explore the results and analyse conservation patterns of the identified proteins, we left the GENREs and performed an analysis of the associated enzymes for each reaction identified (Table 5): we compared the number of amino acids for each one of four species: *Methanosarcina barkeri, Escherichia coli* (K12), *Thermotoga maritima* and *Aquifex aeolicus*. *Aquifex* was inserted in this analysis because it is proven to be one of the most ancient bacterial genera living today – divergence around 3.46 Ga ago (Braunstein et al, 2008), along with *Thermotoga* - 3.3 Ga (Braunstein et al, 2008). We wanted to compare the presence of these sequences in both different ancient bacteria, to analyse if ancient sequences represented smaller number of amino acids or not, and which were the differences from these ancient bacteria to the modern bacteria *E. Coli* and to the archaea *M. barkeri*.

In one special case there is no protein recognized to perform the reaction in any species (pyrimidine phosphatase). It remains a fundamental question in Cellular Biology to identify the enzyme that performs this reaction, although it is suggested nowadays that it might be catalyzed by a phosphatase of broad substrate specificity. In other cases, there is no ortholog or protein assigned just for one species – *Methanosarcina barkeri* or *Thermotoga maritima* (see discussion for details).

To analyse the distribution, similarity and probable ancestry of one of the enzymes, we constructed one phylogenetic tree of chorismate synthase, based on protein sequences (Fig. 19). This enzyme was chosen for two main reasons: it is central (a hub) in the most represented pathway in the results (See Discussion for details) and all the species have proteins with similar size, which is good for phylogenetic alignment. We used NCBI database to collect the sequences and MEGA5 (Tamura et al, 2011) to perform the alignment and construct the Neighbour-Joining phylogenetic tree. We used 1000 Bootstrap replicates for statistical analysis.



	SET A - Eukaryotes&Prokaryotes	SET B - Prokaryotes
Nucleotide Met.	53	56
Vitamins & Cofactor Biosyn.	23	33
Transport	7	25
Central Met.	12	20
Tyrosine, Tryptophan & Phenylalanine Met.	12	16
Valine, leucine & isoleucine Met.	14	14
Alanine & aspartate met.	13	13
Glycolysis/Gluconeogenesis	9	11
Histidine Met.	11	11
Glutamate met.	8	10
Coenzyme A Biosyn.	8	8
Amino Acid Met.	7	7
Lipid & Cell Wall Met.	4	7
Threonine & Lysine Met.	3	6
Citrate Cycle	5	5
Cysteine Met.	3	4
Glutamine Met.	4	4
Glycine & Serine Met.	4	4
Lipid Biosyn.	0	4
Arginine & Proline Met.	2	3
Methionine Met.	3	3
Tetrahydramethanopterin Biosyn.	1	3
Aminosugars Met.	0	2
Other	4	5

Fig. 17 – Conservation of metabolic reactions (grouped in metabolic subsystems). Conservation in the three domains of life – Eukaryotes & Prokaryotes (*Sacharomyces cerevisiae* for eukaryotes, *Escherichia coli* for bacteria and *Methanosarcina barkeri* for Archaea) is compared with conservation only in Prokaryotes (*Escherichia coli* and *Methanosarcina barkeri* only)

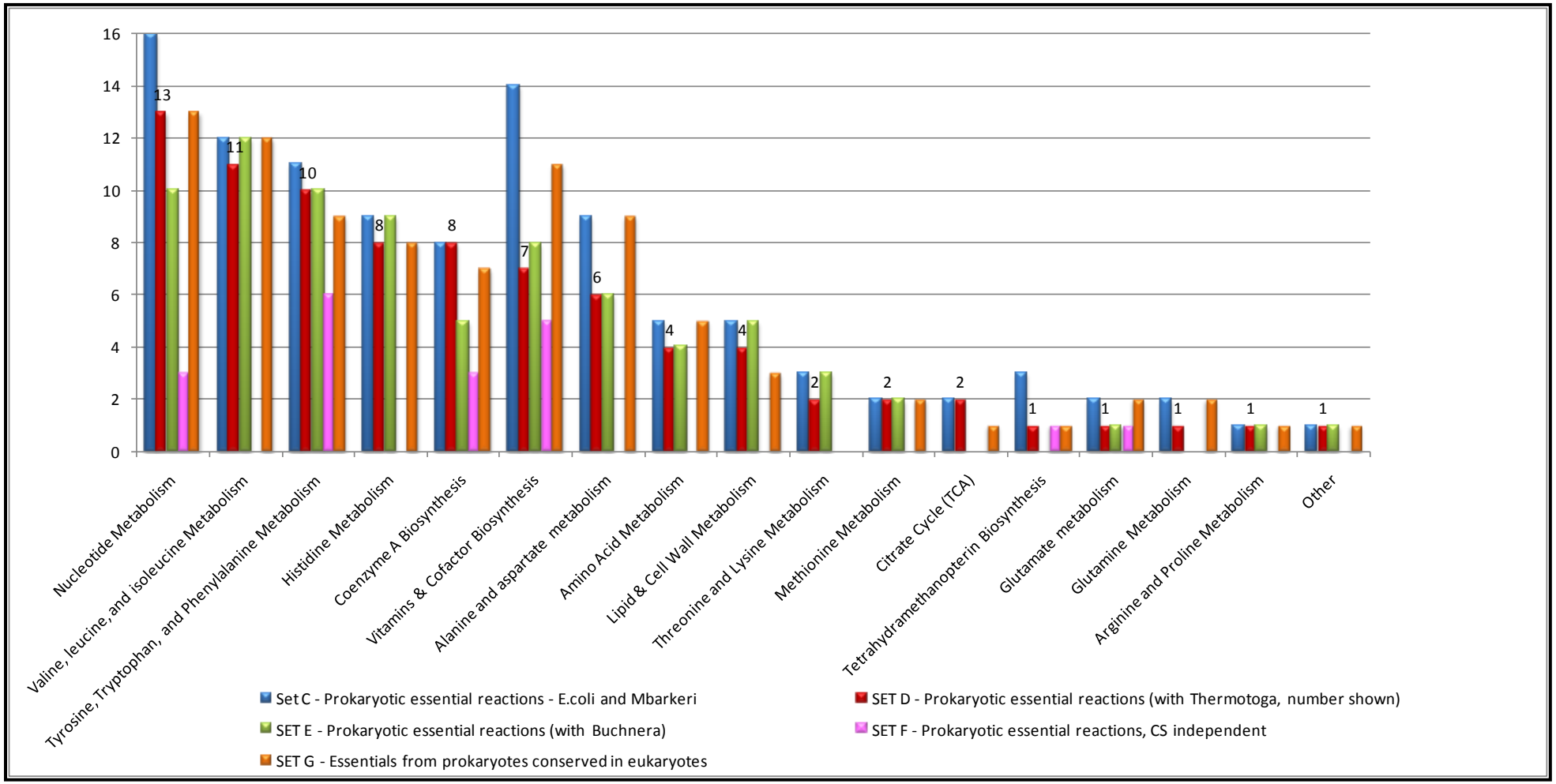


Fig. 18 – Essential reactions conserved in different prokaryotic metabolic models. Set E, D, F and G represent set C intercepted with a specific set.

Table 5 – Most conserved essential reactions for biomass growth in prokaryotic metabolic networks

Reaction name (essential in *Ecoli* W and *Mbark*)	Essential in *E coli* K12 plasticity study	Essential in *T mar.*	Essential in *B aph.*	Metabolic Subsystem	Number of amino acids in annotated sequence	Notes	Ref sequence Thermotoga maritima
anthranilate phosphoribosyl-transferase	✓	✓	✓	Tyrosine, Tryptophan, and Phenylalanine Metabolism	*Methanosarcina*: 366; *Thermotoga*: 589; *Escherichia*: 532; *Aquifex*: 340	In thermotoga maritima, it is named: anthranilate synthase component II and anthranilate phosphoribosyltransferase	GenBank: CAA52203.1
chorismate synthase	✓	✓	✓	Tyrosine, Tryptophan, and Phenylalanine Metabolism	*Methanosarcina*: 365; *Thermotoga*:376; *E. coli*: 356; *Aquifex*: 398	Catalyses the final step of the shikimate pathway.	NP_228158.1
3,4-Dihydroxy-2-butanone-4-phosphate synthase	✓	✓	✓	Vitamins & Cofactor Biosynthesis	*Methanosarcina*: 246; *Thermotoga*:388; *E.coli*:217; *Aquifex*:406	Flavin biosynthesis . In Thermotoga maritima it's named: Bifunctional 3,4-dihydroxy-2-butanone 4-phosphate synthase/GTP cyclohydrolase II; Ecoli's is not very similar; Ecoli's GTP cyclohydrolase doesn't have ortholog	NP_229623.1
dihydrofolate synthase	✓	✓	✓	Vitamins & Cofactor Biosynthesis	***Methanosarcina*: ?** *Thermotoga*:430; *E.coli*:422; *Aquifex*:410	Folate biosynthesis. Bifunctional (annotated for Ecoli and Tmar) : folylpolyglutamate synthase/dihydrofolate synthase; Ecoli's is not very similar; Aquifex - folylpolyglutamate synthase only	NP_227981.1
dephospho-CoA kinase	✓	✓	✓	Coenzyme A Biosynthesis	***Methanosarcina*: ?** *Thermotoga*:180; *E.coli*:206; *Aquifex*:196	Catalyzes the final step in CoA biosynthesis; not present in archaeons!	NP_229188
dTMP kinase	✓	✓	✓	Nucleotide Metabolism	*Methanosarcina*: 205; *Thermotoga*:197; *E.coli*:213; *Aquifex*:195	Thymidine monophosphate kinase	Swiss-Prot: Q9X0I3.1
guanylate kinase (GMP:ATP)	✓	✓	✓	Nucleotide Metabolism	***Methanosarcina*: ?** *Thermotoga*:207; *E.coli*:207; *Aquifex*:109 (hypothetical)	Catalyzes the reversible phosphoryl transfer from ATP to GMP to yield ADP and GDP. It plays an essential role in the biosynthesis of guanosine triphosphate (GTP). Also important for the activation of some antiviral and anticancer agents, such as acyclovir, ganciclovir, carbovir, and thiopurines.	NP_229489.1

indole-3-glycerol-phosphate synthase	✓	✓	✓	Tyrosine, Tryptophan, and Phenylalanine Metabolism	*Methanosarcina*: 268; *Thermotoga*: 230; ***E.coli*:452(fused);** *Aquifex*: 257	Enzyme in the tryptophan biosynthetic pathway; active as a separate monomer in most organisms, but also found fused to other enzymes as part of a bifunctional or multifunctional enzyme involved in tryptophan biosynthesis	NP_227955.1
NAD kinase	✓	✓	✓	Vitamins & Cofactor Biosynthesis	*Methanosarcina*: 275; *Thermotoga*:258; *E.coli*:292; *Aquifex*: 274	Catalyses the phosphorylation of NAD to NADP utilising ATP and other nucleoside triphosphates as well as inorganic polyphosphate as a source of phosphorus	NP_229531.1
nucleoside-diphosphate kinase (ATP:dTDP)	✓	✓	✓	Nucleotide Metabolism	*Methanosarcina*: 149; ***Thermotoga*:???**; *E.coli*:143; *Aquifex*: 142	NDP kinase domains are present in a large family of structurally and functionally conserved proteins from bacteria to humans that generally catalyze the transfer of gamma-phosphates of a nucleoside triphosphate (NTP) donor onto a nucleoside diphosphate (NDP) acceptor through a phosphohistidine intermediate.	Aquifex: NP_214093.1
pyrimidine phosphatase	✓	✓	✓	Vitamins & Cofactor Biosynthesis	No enzyme associated!		
3-phosphoshikimate 1-carboxyvinyl-transferase	✓	✓	✓	Tyrosine, Tryptophan, and Phenylalanine Metabolism	*Methanosarcina*: 443; *Thermotoga*:421; *E.coli*:427; *Aquifex*: 431	Catalyses the reaction between shikimate-3-phosphate (S3P) and phosphoenolpyruvate (PEP) to form 5-enolpyruvylshkimate-3-phosphate (EPSP), an intermediate in the shikimate pathway leading to aromatic amino acid biosynthesis.	NP_228156.2
riboflavin synthase	✓	✓	✓	Vitamins & Cofactor Biosynthesis	*Methanosarcina*: 154 (α); *Thermotoga*:190 (α); *E.coli*:213; *Aquifex*:207(α)	Catalyzes the final reaction of riboflavin biosynthesis.	NP_229624.1
shikimate kinase	✓	✓	✓	Tyrosine, Tryptophan, and Phenylalanine Metabolism	*Methanosarcina*: 169; ***Thermotoga*:492;** *E.coli*:173; *Aquifex*:168	The fifth enzyme in the shikimate pathway, a seven-step biosynthetic pathway found in bacteria, fungi and plants. Chorismic acid is an important intermediate in the synthesis of aromatic compounds, such as aromatic amino acids, p-aminobenzoic acid, folate and ubiquinone. In Thermotoga maritima: bifunctional shikimate kinase/3-dehydroquinate synthase	NP_228159.1

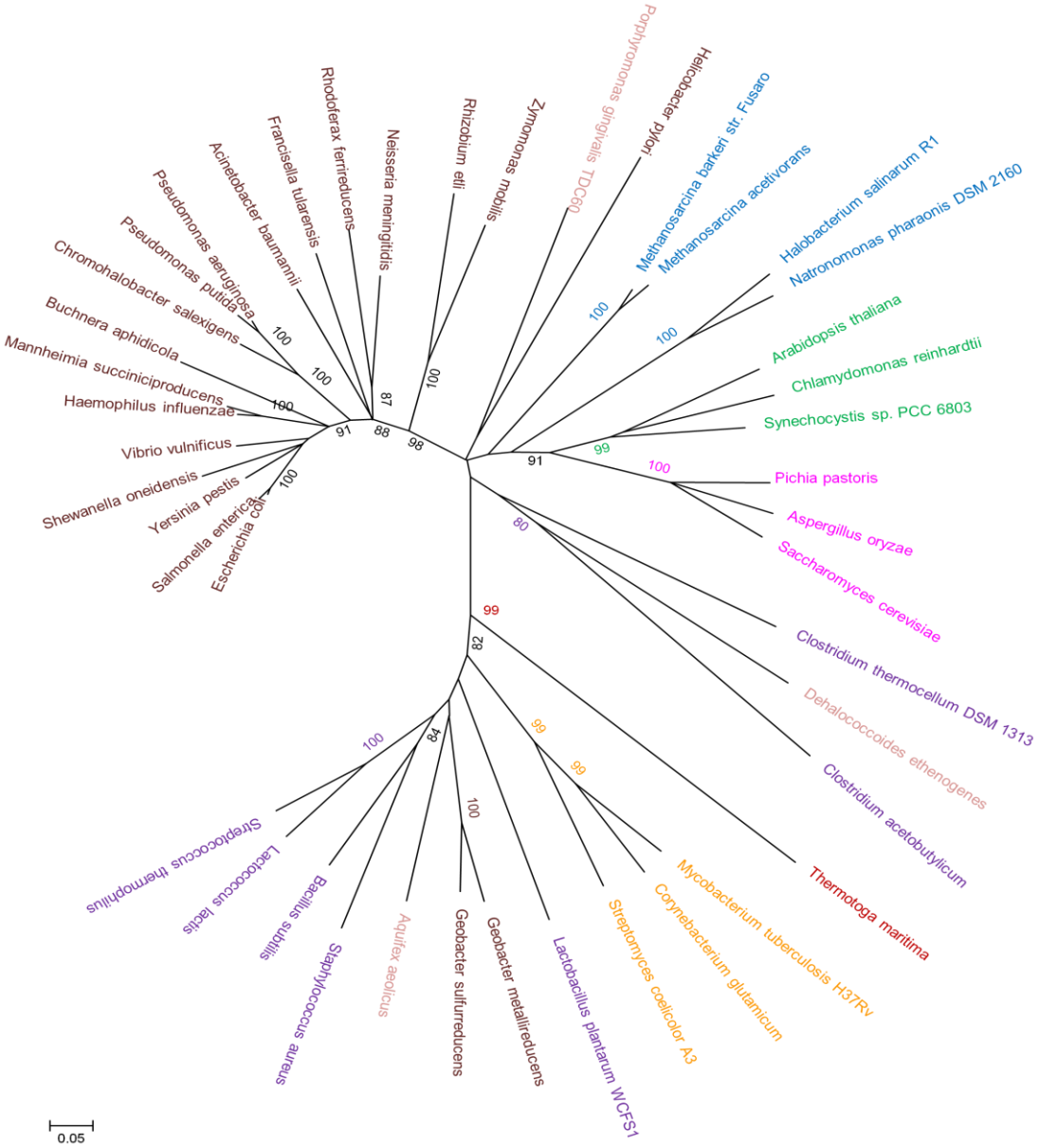


Fig. 19 – Phylogenetic tree of chorismate synthases (protein sequences).
Species used are only those for which a predictive GENRE exists, except for *Aquifex aeolicus*. Brown – Proteobacteria; Blue – Archaea; Green – Plant, algae and cyanobacteria; Dark pink – Fungi; Purple – Firmicutes; Yellow – Actinobacteria; Light pink – phyla with only one representative, except for *Thermotoga maritima* that is highlighted in red.

4. Discussion

4.1 Analysis of all predictive prokaryotic GENREs

There is a notable requirement of an online database dedicated only to validated GENREs, which could facilitate not only comparative studies like this one, but also the homogenization of GENRE building and annotation process, as the existent models vary significantly when it comes to nomenclature, format and features. Bigg (Schellenberger et al, 2010) is still poor, with few models, and Model SEED (Bockstege & Hazekamp, 2011) is still directed to model construction, without information on validation, and lacking many validated models (we couldn't find there *Thermotoga maritima*'s, neither *Escherichia coli* W, two of the models chosen for this study). The best source for a comprehensive list of all the GENREs is, to our opinion, Supplementary Table 1 entitled "Available predictive genome-scale metabolic network reconstructions" from (Palsson et al, 2009), that is kept online updated in http://gcrg.ucsd.edu/In_Silico_Organisms/Other_Organisms. A community effort to build a database with all the GENREs, improving the included models with the fulfilment of some basic requirements would make it easier for Systems Biology to evolve. The building process of one such database, which could integrate interesting information as taxonomy, genetic/proteic sequences and structures, associated to each reaction, would, per se, yield an explosion of information and open questions to answer.

Until now, there is one convention for naming GENREs, still not used by all the authors of the GENREs found (Table 4, Methods and Results). This convention is **iXXxxx(a)** (Reed et al, 2003). The 'i' refers to an *in silico* model; it is followed by the initials (XX) of the first author of the model; then the number of genes included in the model are indicated (xxx). Any lower-

case letters following the number of genes (a) indicate that slight modifications were made to the model.

The analysis of the current state of the art in genome-scale metabolic modelling shows that the distribution of the models among bacterial phyla is not so biased as it would be expected - 63% of the models belong to Proteobacteria, and this is, in fact, the phylum with the largest number of entries in NCBI Taxonomy browser (41,3% of all bacterial species). In the phylogenetic tree built for all prokaryotic GENREs, the well-distributed representation of all phyla is also evident (Fig. 13 – Methods and Results). To our knowledge, for a comparative study to be performed, the major problem with GENREs are the validation methods, analysed in (Oberhardt et al, 2009). The authors recognize that "no standard exist for how a model should be validated", although several model predictions have been proved to be concordant with experimental ones through different techniques (Oberhardt et al, 2009). Validation can occur with KO/essentiality predictions (highlighted in our research as we wanted to predict essential reactions, although some of the listed validations refer to gene knock-outs - Fig. 11 and Table 4); high/low-throughput phenotypes (phenotypes in many growth conditions or just one); rate/yield predictions (biomass growth rate, production of specific metabolite), more quantitative and reliable; etc.

4.2 Choice of Case Studies

Case studies were chosen between prokaryotic GENREs according to their validation level, mainly, but also to fulfil the purposes of this work - to study minimal/ancestral metabolism (see also Methods and Results section). We used two *in silico* subspecies of ***Escherichia coli*** (Set C – strain W and Set F – strain K12), the species which has been most used and validated in genome-scale modelling - (Palsson & Feist, 2008). This is also a species with great bioindustrial/biomedical interest. ***Thermotoga maritima*** is one of the deepest lineages in Eubacteria, estimated to have diverged 3.3 billion years ago (Sheridan et al, 2003) and having the highest percentage (24%) of genes similar to archaeal genes (Godzik et al, 2009). ***Buchnera aphidicola*** has one of the smallest genomes known (Gil et al, 2002), and although being an endosymbiont of higher organisms, its symbiosis with host began about 200 million years ago, much earlier than the association of the human parasite very much used in minimal genome/cell studies, *Mycoplasma genitalium*, which is also modelled and could have been

used. As we needed to represent Archaea in our study for ancestral inferences, we chose ***Methanosarcina barkeri***'s model for being well-validated. Moreover, as it came from a comparative study, all of its reactions were already concordantly named with *Escherichia coli*'s.

In Table 4 it's visible that *Escherichia coli* W, *Buchnera aphidicola* and *Thermotoga maritima* have all, at least, 25% of the genes represented in the model (highlighted in yellow in "best relation" column - Table 4). *Methanosarcina barkeri* has a considerably big genome - 5072 genes - and only 14% are represented in the model; however, the model has been validated through at least four different analysis techniques (Fig. 11; green highlighted in "# different techniques" column - Table 4).

4.3 Discovering conserved essential reactions

In Table 4 the utilization of GENREs in another comparative study was highlighted (Kun et al, 2008). This study is the only one we found comparing predictions of several GENREs from different phyla, and interestingly, it is also highly connected to the theme of the origin of life and metabolism. The authors used various GENREs and one minimal metabolic network (Gabaldon et al, 2007) to identify autocatalytic components of metabolic networks (chemicals that are necessary for the synthesis of more of themselves). In fact, these are also "hubs" or important centres in metabolic networks that are likely to be very ancient and that are indispensable for the design of minimal cells – there is no known cell that can live without their presence. The authors show that intermediary metabolism is invariably autocatalytic for ATP, and in some cases, NAD+, coenzyme A and tetrahydrofolate. In our case, we studied the essentiality of chemical **reactions** in the networks. We performed our analysis in what is the essential **chemistry** of the cytoplasm as the previous referred authors, but while Kun's work was dedicated to the study of essential organic chemicals (metabolites) (Kun et al, 2008), we analysed the enzymes associated with essential reactions.

4.3.1 Conservation of reactions between Prokaryotes and Eukaryotes (Fig. 17)

Comparing the GENREs of *Sacharomyces cerevisiae*, *Methanosarcina barkeri* and *Escherichia coli*, (Feist et al, 2006) it is remarkable the conservation of **Nucleotide metabolism**

associated reactions; from the 56 conserved reactions in prokaryotes, 95% are still present in *Sacharomyces cerevisiae*. The other categories that show total conservation (100%) are the metabolism of: valine, leucine and isoleucine; Alanine and aspartate; Histidine; Coenzyme A; Aminoacid (other); Citrate cycle; Glutamine; Glycine and Serine and Methionine. After, there is: Glycolysis/gluconeogenis (82%); Glutamate (80%); Tyrosine, Tryptophan and Phenylalanine (75%); Cysteine (75%); Vitamins and Cofactor metabolism (70%). It is remarkable that the Transport category, although being highly represented with 25 reactions common in prokaryotes only, shows a small number of 28% of conservation in the eukaryote; this represents a probable main change when eukaryotes appeared.

4.3.2 Conservation of essential reactions in Prokaryotes (Fig. 18 and Table 5)

In Fig. 18 the conservation of essential reactions for biomass growth of different metabolic networks is represented. We calculated the intersection between essential reactions of *Escherichia coli* W and *Methanosarcina barkeri* first, as their set was the biggest (blue bars – set C); after, we calculated the intersection of this set with the set of essential reactions from *Thermotoga maritima* (red bars – set D), *Buchnera aphidicola* (green bars – set E) and one set of essential reactions calculated in another study (Rodrigues & Wagner, 2009) (see Methods and results for details) - pink bars – set F. Finally, we also identified the conservation of essential reactions in *Sacharomyces cerevisiae* network (orange bars, set G). In terms of the major metabolic subsystems, the essential reactions identified were notably mainly from Nucleotide Metabolism and Amino acids metabolism. There is also a good representation of Vitamins and Cofactor biosynthesis-associated reactions. Notice that Trypthophan, Phenylalanine and Tyrosine pathway is the only amino acids pathway that is represented also in set F with more than one reaction.

In the final set of conserved essential reactions, common to all five networks studied (Table 5) there were fourteen enzymes; five of them related to the Tryptophan, Tyrosine and Phenylalanine pathway; five enzymes related to Vitamins & Cofactor biosynthesis, three enzymes belonging to Nucleotide pathways and one enzyme from coenzyme A pathway. Interestingly, *Methanosarcina's* genome lacks the identification of three of these enzymes. When building the metabolic model, the authors realized that these reactions are essential for

the growth of the microorganism, but it remains to be analysed *in vivo,* if these reactions are actually occurring, which enzymes are performing them?

Another important feature of the final set analysed in Table 5 is the over-representation of bifunctional enzymes – it should be highlighted that this reflects the centrality of these enzymes in global networks– they perform more than one function in the cell and removing them from the network will cause extra perturbation. Although this doesn't confirm the ancestry of the bifunctionality per se, both functions they represent are certainly ancient. It remains to be analysed if ancestral enzymes were predominantly bifunctional, or not.

In Fig. 21 we show *Thermotoga maritima*'s Tryptophan, Tyrosine and Phenylalanine pathway retrieved from KEGG. The results of our work are highlighted with the enzymes present in Table 5 that are represented in the pathway (circled in red); one extra enzyme relating to Folate pathway is marked. Five more enzymes that were conserved in four networks (not shown in Table 5, see Supplementary Data) are circled in orange. Curiously or not, 6 of the 14 identified enzymes in all the networks of this study and five more that are common to four organisms can be represented in the pathway for synthesis of aromatic amino acids - Tryptophan, Tyrosine and Phenylalanine pathway, which centralises in Chorismate. Chorismate synthase is essential in all networks studied and a central enzyme in the results of this study as one can see in Fig. 21 and as we discuss in the next section.

4.3.2.1 Chorismate synthase – the monofunctional, central and ancestral enzyme

Chorismate synthase catalyzes the final step in the shikimate pathway; this pathway links the metabolism of carbohydrates to the biosynthesis of the three aromatic amino acids and several other aromatic secondary metabolites. The pathway is **not present in mammals** making this enzyme a prime target for the development of antimicrobial and herbicidal agents. Chorismate synthase is biochemically unique – its mechanism of catalysis is the only one of its kind in nature - an 1,4-anti-elimination of the 3-phosphate group and the C-(6proR) hydrogen from 5-enolpyruvylshikimate-3-phosphate (Fitzpatrick et al, 2001) (see Fig. 20).

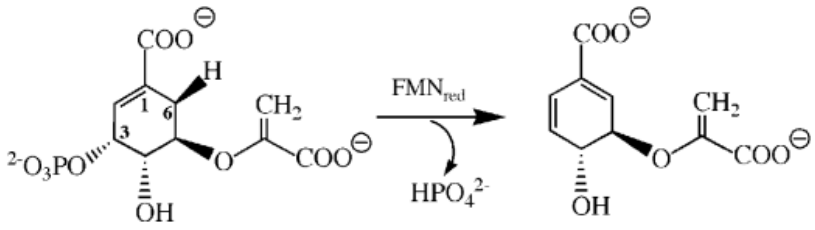


Fig. 20 – Reaction catalyzed by chorismate synthase (Kitzing et al, 2004).

The enzyme has an absolute requirement for reduced flavin mononucleotide (FMN), the principal form in which riboflavin is found in cells and tissues (observe that Flavin biosynthesis is also represented in Table 5). Phylogenetic analysis have shown before that all chorismate synthases known are of monophyletic origin (descend all from the same ancestral sequence) (Macheroux et al, 1999).

We built a phylogenetic tree for this protein with all the species with a predictive GENRE, based on aminoacidic sequences. It should be referred that *Mycoplasma genitalium*, the species used more to study minimal life (See State of art – Minimal genome) doesn't have this enzyme – it probably obtains all the aromatic amino acids from the host. It's important to refer that ViCe – one minimal *in silico* organism built from a top down approach (Chiarugi et al, 2007) (See Section 2.1.2.1) also doesn't have the chorismate synthase enzyme – this is due to the top-down approach it used- identifying essential reactions from *Mycoplasma genitalium*.

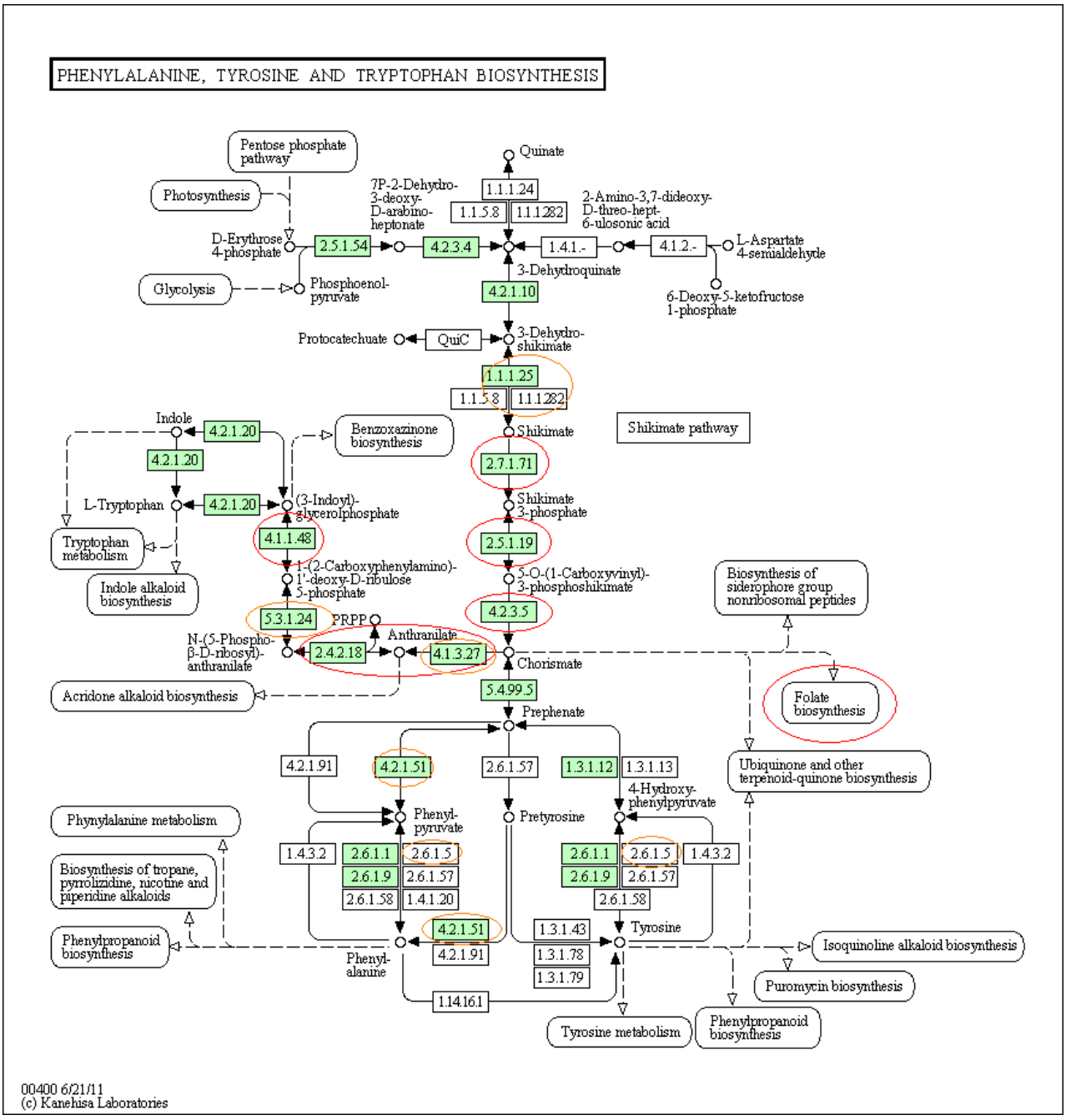


Fig. 21 – *Thermotoga maritima*'s Tryptophan, Tyrosine and Phenylalanine biosynthesis pathway, obtained from KEGG (Kanehisa et al, 2006). Red circles – conserved essential reactions in all five networks studied; Orange circles – essential in four networks.

4.4. Studying minimal essential metabolism, minimal life and early life with Systems Biology

From the results of our work, it can be reaffirmed that the minimal life concept needs to be refined in terms of what are we intending to define as minimal. As discussed in the state of the Art Section, minimal life is a concept totally distinct from ancestral life; these concepts separate in the **medium** on which the cell is living. If the medium is rich, the cell might not need to produce some components - we saw that the principal enzyme identified in our work is not present in the model-species for the study of minimal life - *Mycoplasma genitalium*. However, life in general needs aromatic amino acids, and the extremely evolved *M. Genitalium* still needs to get it from the host. In terms of bio-industrial applications, it might make more sense to cut reactions that don't imply the supply of more expensive nutrients, as the case of essential amino acids.

As in this study the common essential reactions are compared, it relies on the veracity of the common ancestor theory. Although many times in the history of evolution the same functions appeared in distant lineages, we found that at least one of the most indispensable enzymes to metabolic networks (chorismate synthase) appeared much earlier than the divergence of prokaryotes and eukaryotes, close to LUCA's appearance. We believe that our approach directs us to the ancestral metabolism that appeared on LUCA (Fig. 19). This statement is not only visible in the Chorismate Synthase inferences, but also in the major representativeness of Nucleotide metabolism in the most conserved essential reactions, which is compatible with the theory of RNA and proteins first; it is relevant to remember that although all life forms known have thousands of single RNA molecules, ribosomes and proteins floating in their cytoplasm, only one giant DNA molecule is present in the cell. The hypothesis that this might have been the turning point in the assembly of true cells – the association of plasmids to the RNA/protein cellular machinery – can be postulated. Looking at the results of this work – ancestral enzymes that deal with nucleotide metabolism are central and essential in all prokaryotic networks; the evolution of these enzymes might have turned possible an efficient metabolism that integrated nucleotides with proteins.

Phylogenetic inferences based on DNA sequences have been proven to have problems when they attempt to be detailed, which is much more important when we try to study

sequences that don't exist anymore (LUCA and other ancestral sequences). These problems of sequence-based phylogeny are mainly due to horizontal gene transfer between species, but also because there are different rates of evolution for different proteins and for different species (between others). Although this study didn't intend to produce phylogenetic inferences, it shows that metabolism comparison can overcome some of these problems. One can study metabolic and protein networks, realize which enzymes are essential in pathways, and then compare them *per se*. In fact, in our study we saw that, for example, guanylate kinase is not present in archaeal genomes, although it is supposed to be essential in *Methanosarcina barkeri*'s GENRE and in all other GENREs studied. The authors of *Methanosarcina barkeri*'s model report the addition of 110 reactions with no association with any gene product in the annotation, because they have been reported in prior literature, or because they were required to fill a gap in the reconstructed network (Feist et al, 2006). It has been proposed that guanylate kinase is not present in archaea because it might not have been inherited from LUCA to all life domains, but evolved in Bacteria and had laterally spread into eukaryotes (Leipe et al, 2003). But it remains to be explored why it was considered essential in the archaeal GENRE and which protein is responsible for this reaction in the archaea domain. Here, we reaffirm the need to study these specific reactions *in vivo* to associate proteins and genes to them, not only to improve physiological knowledge but also to validate the predictions of the archaeal GENRE.

5. Conclusions

In this work we studied and tried to improve the concept of minimal cell, which has been defined in literature as the minimal chemical system able to grow, divide and evolve. However, the influence of the environment is not present in this concept. A minimal environment or medium is totally subjective – when directed to biotechnological applications, it should take in consideration nutrient cost and availability and its impact on specific productivity. Therefore, the study should be specific for that industrial goal, using the species that is known to be the best for the application. In most cases, it has been *Escherichia coli*, which grows fast and is easily transformable; there are other cases where special metabolic pathways are recognized in other species which are then tested and used. Still in this case, if the species grows slow and its biomass production rate is important for the process, it is preferable not to use the species but rather to use genetic transformation for the insertion of these metabolic pathways in *E. coli's* genome which is therefore used in the industrial process. In fact, one recent review on the design of synthetic cells for biotechnological applications emphasized that the utility of a minimal cell in a biotechnological context would be limited to its retention of properties required for robust and predictable growth (Foley & Shuler, 2010).

Under a biological, systematic approach, the research on minimal cells has been until now based on top-down, genome-reducing approaches. The appearance of new omics datasets, as well as new frames and tools allows a new approach to the minimal cell – genome-scale modelling. While not all genes are represented in *in silico* models yet, proteins and chemical reactions that are necessary for biomass growth are. In this work, we used different metabolic GENREs to predict which are the essential reactions for biomass growth in different prokaryotic species – one archaea, one minimal genome of an endosymbiont, *Escherichia coli* and one ancient species of bacteria. We identified features common to all networks, mainly in the essentiality of Nucleotide metabolism-associated reactions and amino

acid biosynthesis. More specifically, all of our networks showed essentiality for enzymes in the pathway of Tryptophan, Tyrosine and Phenylalanine. We also identified gaps in the knowledge to be explored, as some essential reactions didn't have any identified enzyme in the proteome or genome of the archaea. One specific reaction doesn't have an enzyme associated in any species. It remains to be analysed why, and where are these reactions occurring? Are they spontaneous? Can we control them or influencing their levels will have an impact on growth?

We also implemented this analysis in the study of primordial metabolism. Conserved features among species have always been used to infer relationships and age of species divergence, in Darwin's original theory but also with modern 16S rRNA sequencing. Aligning metabolic pathways is still computationally hard and incipient, as these are highly complex. However, comparing GENREs predictions is quantitative and easy with user-friendly software. This is therefore one open door in Evolutionary Biology and in the study of ancestral life. With our study, we confirmed that at least one modern prokaryotic species conserved essential reaction (chorismate synthase) is truly ancient, from before the divergence of archaea and eukarya from bacteria (See Fig. 19). Moreover, we saw that a biosynthetic ability may disappear in higher mammals and some parasites or symbionts but they will still need to get the essential metabolites produced in that pathway from their outside-nutrition.

The study of ancestral networks needs to be more sustained in other fields of knowledge, as geochemistry, that could inform about the correct composition of earth's surface at the time life is believed to have originated. However, the fast advancement of biotechnology and bioinformatics seems to overpass geochemistry, as there are studies now, that from inferred protein composition can infer temperature and ph ranges from billions of years ago (Gaucher et al, 2008; Perez-Jimenez et al, 2011).

5.1 Theoretical and *in silico* testing and posterior *in vivo/in vitro* experimentation – Saving time and money in Molecular Biotechnology

It is important to reaffirm the very cheap but yet very promising approach of *in silico* bacterial modelling. Even more than in testing metabolic features, the exploration of minimal metabolism/minimal genomes is truly effort-consuming in wet lab; it requires extensive laboratory human resources, time and money. Notice a notable study in which *Bacillus subtilis* genes were systematically inactivated (Kobayashi et al, 2003); it is indispensable for validation

of *in silico* work with individual deletions, but it is impressive that once the metabolic network of *Bacillus subtilis* is completely reconstructed with its GPR information, it would take only one person and few hours to compute the exact same results of this study, with 99 authors.

When applications are desired and one GENRE is already built and validated, if one hypothesis does not pass one *in silico* test on this validated model, it will be unlikely that it passes the *in vivo* experiment. It is hard to sustain the opposite, though: we cannot confirm one hypothesis passing one *in silico* test without a wet-lab validation of the proposal. So both approaches are necessary and complementary.

Future work in Systems biology is impressively vast; there is a lot to do in organizing knowledge of different biosystems and integrating it in accessible frameworks. Moreover, this kind of work identifies gaps in knowledge that can be hot topics for experimental, specific researches, as happened with this work. Bioengineering can only grow from true Systems Biology, as the parts and respective interactions that the latter describes are the parts that the first integrates into novel or improved existing bio-engines.

The comparison and systematic analysis of metabolism across phylogenetically different GENREs remains an unexplored field. Future work to include in phylogenetic/ancestry inferences should embrace comparison of minimal media from different prokaryotes, levels of internal metabolites and fluxes also. From all that has been analysed in this integrative work, we propose that a bottom up approach based on previous top down approaches from more datasets than the genome to be the best in the quest for minimal cells. Basically, we propose the identification of the essential life features through comparison of several organisms and not only one, achieving a level of understanding of the parts that can be directed to specific applications and also to ancestry inferences. After this identification, minimal cells could be tested *in silico* with the building of a novel model that could be simulated. This requires more work to be done in the identification of specific parts in the systems but also the recognition of their dynamic interactions for posterior manipulation.

References

Akesson M, Forster J, Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering* **6:** 285-293

Alper H, Jin YS, Moxley JF, Stephanopoulos G (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. *Metabolic Engineering* **7:** 155-164

Battistuzzi FU, Feijao A, Hedges SB (2004) A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *Bmc Evol Biol* **4:** -

Beard DA, Liang SD, Qian H (2002) Energy balance for analysis of complex metabolic networks. *Biophys J* **83:** 79-86

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* **2:** 727-738

Blank LM, Ebert BE, Buhler B, Schmid A (2008) Metabolic capacity estimation of Escherichia coli as a platform for redox biocatalysis: constraint-based modeling and experimental verification. *Biotechnol Bioeng* **100:** 1050-1065

Blank LM, Kuepfer L, Sauer U (2005) Large-scale C-13-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biology* **6:** -

Blazeck J, Alper H (2010) Systems metabolic engineering: Genome-scale models and beyond. *Biotechnol J* **5:** 647-659

Bockstege B, Hazekamp N (2011) The Model SEED Plug-in for Viewing and Editing Metabolic Models. *10th Annual Celebration for Undergraduate Research and Creative Performance (2011) Paper 2 http://digitalcommonshopeedu/curcp_10/2/*

Bork P, Letunic I (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23:** 127-128

Braunstein A, Mulet R, Pagnani A (2008) Estimating the size of the solution space of metabolic networks. *Bmc Bioinformatics* **9:** -

Burgard AP, Maranas CD (2001) Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions. *Biotechnol Bioeng* **74:** 364-375

Burgard AP, Vaidyaraman S, Maranas CD (2001) Minimal reaction sets for Escherichia coli metabolism under different growth requirements and uptake environments. *Biotechnol Progr* **17:** 791-797

Cakir T, Efe C, Dikicioglu D, Hortacsu A, Kirdar B, Oliver SG (2007) Flux balance analysis of a genome-scale yeast model constrained by exometabolomic data allows metabolic system identification of genetically different strains. *Biotechnol Prog* **23:** 320-326

Cello J, Paul AV, Wimmer E (2002) Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template. *Science* **297:** 1016-1018

Chang A, Scheer M, Grote A, Schomburg I, Schomburg D (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* **37:** D588-592

Chiarugi D, Degano P, Marangoni R (2007) A computational approach to the functional screening of genomes. *Plos Comput Biol* **3:** 1801-1806

Covert MW, Palsson BO (2002) Transcriptional regulation in constraints-based metabolic models of Escherichia coli. *Journal of Biological Chemistry* **277:** 28058-28064

Covert MW, Palsson BO (2003) Constraints-based models: Regulation of gene expression reduces the steady-state solution space. *Journal of Theoretical Biology* **221:** 309-325

Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213:** 73-88

Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics* **24:** 2044-2050

Deamer D (2009) On the origin of systems Systems biology, synthetic biology and the origin of life. *Embo Rep* **10:** S1-S4

Duarte NC, Herrgard MJ, Palsson BO (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research* **14:** 1298-1309

Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *Fems Microbiol Rev* **33:** 164-190

Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol* **19:** 125-130

Edwards JS, Palsson BO (2000a) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* **97:** 5528-5533

Edwards JS, Palsson BO (2000b) Robustness analysis of the Escherichia coli metabolic network. *Biotechnol Prog* **16:** 927-939

Fani R, Fondi M (2009) Origin and evolution of metabolic pathways. *Phys Life Rev*

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3:** 121

Feist AM, Scholten JCM, Palsson BO, Brockman FJ, Ideker T (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Molecular Systems Biology***:** -

Fell DA, Small JR (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem J* **238:** 781-786

Fitzpatrick TB, Killer P, Thomas RM, Jelesarov I, Amrhein N, Macheroux P (2001) Chorismate synthase from the hyperthermophile Thermotoga maritima combines thermostability and increased rigidity with catalytic and spectral properties similar to mesophilic counterparts. *Journal of Biological Chemistry* **276:** 18052-18059

Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S (2006) Graemlin: General and robust alignment of multiple large interaction networks. *Genome Research* **16:** 1169-1181

Foley PL, Shuler ML (2010) Considerations for the design and construction of a synthetic platform cell for biotechnological applications. *Biotechnol Bioeng* **105:** 26-36

Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnology and Bioengineering* **91:** 643-648

Forster AC, Church GM (2006a) Towards synthesis of a minimal cell. *Molecular Systems Biology***:** -

Forster AC, Church GM (2006b) Towards synthesis of a minimal cell. *Molecular Systems Biology*

Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Research* **13:** 244-253

Gabaldon T, Pereto J, Montero F, Gil R, Latorre A, Moya A (2007) Structural analyses of a hypothetical minimal metabolism. *Philos T R Soc B* **362:** 1751-1762

Galperin MY (2006) The minimal genome keeps growing. *Environmental Microbiology* **8:** 569-573

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31:** 3784-3788

Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451:** 704-U702

Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *Journal of Bacteriology* **185:** 5673-5684

Gianchandani EP, Chavali AK, Papin JA (2010) The application of flux balance analysis in systems biology. *Wiley Interdiscip Rev Syst Biol Med* **2:** 372-382

Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, Merryman C, Young L, Noskov VN, Glass JI, Venter JC, Hutchison CA, Smith HO (2008) Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319:** 1215-1220

Gil R, Sabater-Munoz B, Latorre A, Silva FJ, Moya A (2002) Extreme genome reduction in Buchnera spp.: Toward the minimal genome needed for symbiotic life. *P Natl Acad Sci USA* **99:** 4454-4458

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. *P Natl Acad Sci USA* **103:** 425-430

Godzik A, Zhang Y, Thiele I, Weekes D, Li ZW, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Palsson B, Osterman A (2009) Three-Dimensional Structural View of the Central Metabolic Network of Thermotoga maritima. *Science* **325:** 1544-1549

Goryshin IY, Naumann TA, Apodaca J, Reznikoff WS (2003) Chromosomal deletion formation system based on Tn5 double transposition: Use for making minimal genomes and essential gene analysis. *Genome Research* **13:** 644-653

Haldane JBS (1929) The Origin of Life. *Rationalist Annual***:** 3-10

Henry CS, Overbeek R, Stevens RL (2010) Building the blueprint of life. *Biotechnol J* **5:** 695-704

Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420:** 186-189

Ideker T (2004) Systems biology 101 - what you need to know. *Nature Biotechnology* **22:** 473-475

Izallalen M, Mahadevan R, Burgard A, Postier B, Didonato R, Jr., Sun J, Schilling CH, Lovley DR (2008) Geobacter sulfurreducens strain engineered for increased rates of respiration. *Metab Eng* **10:** 267-275

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. *Nature* **407:** 651-654

Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **7:** 198-210

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34:** D354-357

Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. *Curr Opin Biotechnol* **14:** 491-496

Kelley BP, Yuan BB, Lewitter F, Sharan R, Stockwell BR, Ideker T (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Research* **32:** W83-W88

Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Peralta-Gil M, Santos-Zavaleta A, Shearer AG, Karp PD (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res* **37:** D464-470

Kim TY, Kim HU, Park JM, Song H, Kim JS, Lee SY (2007) Genome-scale analysis of Mannheimia succiniciproducens metabolism. *Biotechnol Bioeng* **97:** 657-671

Kitano H (2002) Systems biology: A brief overview. *Science* **295:** 1662-1664

Kitzing K, Auweter S, Amrhein N, Macheroux P (2004) Mechanism of chorismate synthase - Role of the two invariant histidine residues in the active site. *Journal of Biological Chemistry* **279:** 9451-9461

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Debarbouille M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan R, Eschevins C, Fukushima T, Haga K, Harwood CR, Hecker M, Hosoya D, Hullo MF, Kakeshita H, Karamata D, Kasahara Y, Kawamura F, Koga K, Koski P, Kuwana R, Imamura D, Ishimaru M, Ishikawa S, Ishio I, Le Coq D, Masson A, Mauel C, Meima R, Mellado RP, Moir A, Moriya S, Nagakawa E, Nanamiya H, Nakai S, Nygaard P, Ogura M, Ohanan T, O'Reilly M, O'Rourke M, Pragai Z, Pooley HM, Rapoport G, Rawlins JP, Rivas LA, Rivolta C, Sadaie A, Sadaie Y, Sarvas M, Sato T, Saxild HH, Scanlan E, Schumann W, Seegers JFML, Sekiguchi J, Sekowska A, Seror SJ, Simon M, Stragier P, Studer R, Takamatsu H, Tanaka T, Takeuchi M, Thomaides HB, Vagner V, van Dijl JM, Watabe K, Wipat A, Yamamoto H, Yamamoto M, Yamamoto Y, Yamane K, Yata K, Yoshida K, Yoshikawa H, Zuber U, Ogasawara N (2003) Essential Bacillus subtilis genes. *P Natl Acad Sci USA* **100:** 4678-4683

Kremling A, Bettenbrock K, Gilles ED (2007) Analysis of global control of Escherichia coli carbohydrate uptake. *BMC Syst Biol* **1:** 42

Kuchaiev O, Milenkovic T, Memisevic V, Hayes W, Przulj N (2010) Topological network alignment uncovers biological function and phylogeny. *J R Soc Interface* **7:** 1341-1354

Kuchaiev O, Stevanovic A, Hayes W, Przulj N (2011) GraphCrunch 2: Software tool for network modeling, alignment and clustering. *Bmc Bioinformatics* **12:** -

Kun A, Papp B, Szathmary E (2008) Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biology* **9:** -

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH,

Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang HM, Yu J, Wang J, Huang GY, Gu J, Hood L, Rowen L, Madan A, Qin SZ, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan HQ, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang WH, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz JR, Slater G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Conso IHGS (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921

Lee KH, Park JH, Kim TY, Kim HU, Lee SY (2007) Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol Syst Biol* **3:** 149

Lee S PC, Domach MM, Grossmann IE (2000) Recursive MILP model for finding all alternate optima in LP models for metabolic networks. *Computers Chemical Engineering* **24:** 711-716

Lee SJ, Lee DY, Kim TY, Kim BH, Lee JW, Lee SY (2005) Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation. *Applied and Environmental Microbiology* **71:** 7880-7887

Leipe DD, Koonin EV, Aravind L (2003) Evolution and classification of P-loop kinases and related proteins. *J Mol Biol* **333:** 781-815

Luo RY, Liao S, Tao GY, Li YY, Zeng S, Li YX, Luo Q (2006) Dynamic analysis of optimality in myocardial energy metabolism under normal and ischemic conditions. *Mol Syst Biol* **2:** 2006 0031

Ma HW, Zeng AP (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. *Mol Phylogenet Evol* **31:** 204-213

Macheroux P, Schmid J, Amrhein N, Schaller A (1999) A unique reaction in a common pathway: mechanism and function of chorismate synthase in the shikimate pathway. *Planta* **207:** 325-334

Mahadevan R, Edwards JS, Doyle FJ, 3rd (2002) Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys J* **83:** 1331-1340

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5:** 264-276

Majewski RA, Domach MM (1990) Simple constrained-optimization view of acetate overflow in E. coli. *Biotechnol Bioeng* **35:** 732-738

Manichaikul A, Ghamsari L, Hom EFY, Lin CW, Murray RR, Chang RL, Balaji S, Hao T, Shen Y, Chavali AK, Thiele I, Yang XP, Fan CY, Mello E, Hill DE, Vidal M, Salehi-Ashtiani K, Papin JA (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nature Methods* **6:** 589-U553

Miller SL (1953) A Production of Amino Acids under Possible Primitive Earth Conditions. *Science* **117:** 528-529

Moreira D, Lopez-Garcia P (2009) Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* **7:** 306-311

Mushegian A (1999) The minimal genome concept. *Current Opinion in Genetics & Development* **9:** 709-714

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *P Natl Acad Sci USA* **93:** 10268-10273

Nakabachi A (2008) The 160-kilobase genome of the bacterial endosymbiont Carsonella (vol 318, pg 267, 2006). *Science* **319:** 901-901

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M (2006) The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* **314:** 267-267

Nisbet EG, Sleep NH (2001) The habitat and nature of early life. *Nature* **409:** 1083-1091

Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, Cheevadhanarak S, Nielsen J, Bhumiratana S (2008) The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. *Bmc Syst Biol* **2:** -

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* **5:** 320

Oh SJ, Joung JG, Chang JH, Zhang BT (2006) Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *Bmc Bioinformatics* **7:** -

Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* **28:** 245-248

Palsson BO, Feist AM (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nature Biotechnology* **26:** 659-667

Palsson BO, Feist AM, Herrgard MJ, Thiele I, Reed JL (2009) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* **7:** 129-143

Papin JA, Price ND, Edwards JS, Palsson BO (2002) The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. *Journal of Theoretical Biology* **215:** 67-82

Papoutsakis ET (1984) Equations and calculations for fermentations of butyric acid bacteria. *Biotechnol Bioeng* **26:** 174-187

Park JH, Lee KH, Kim TY, Lee SY (2007) Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 7797-7802

Patil KR, Akesson M, Nielsen J (2004) Use of genome-scale microbial models for metabolic engineering. *Curr Opin Biotechnol* **15:** 64-69

Pereto J (2005) Controversies on the origin of life. *Int Microbiol* **8:** 23-31

Perez-Jimenez R, Ingles-Prieto A, Zhao ZM, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A, Sanchez-Ruiz JM, Gaucher EA, Fernandez JM (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol* **18:** 592-596

Phalakornkule C, Lee S, Zhu T, Koepsel R, Ataai MM, Grossmann IE, Domach MM (2001) A MILP-based flux alternative generation and NMR experimental design strategy for metabolic engineering. *Metabolic Engineering* **3:** 124-137

Portnoy VA, Herrgard MJ, Palsson BO (2008) Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient Escherichia coli strain. *Appl Environ Microbiol* **74:** 7561-7569

Powner MW, Gerland B, Sutherland JD (2009) Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459:** 239-242

Pramanik J, Keasling JD (1997) Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol Bioeng* **56:** 398-421

Pramanik J, Keasling JD (1998) Effect of Escherichia coli biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol Bioeng* **60:** 230-238

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramakrishna R, Edwards JS, McCulloch A, Palsson BO (2001) Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *Am J Physiol Regul Integr Comp Physiol* **280:** R695-704

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297:** 1551-1555

Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol* **4:** R54

Rocha I, Maia P, Evangelista P, Vilaca P, Soares S, Pinto JP, Nielsen J, Patil KR, Ferreira EC, Rocha M (2010) OptFlux: an open-source software platform for in silico metabolic engineering. *Bmc Syst Biol* **4:** 45

Rodrigues JFM, Wagner A (2009) Evolutionary Plasticity and Innovations in Complex Metabolic Reaction Networks. *Plos Comput Biol* **5:** -

Ruppin E, Papin JA, de Figueiredo LF, Schuster S (2010) Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr Opin Biotech* **21:** 502-510

Savinell JM, Palsson BO (1992a) Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *J Theor Biol* **155:** 201-214

Savinell JM, Palsson BO (1992b) Optimal selection of metabolic fluxes for in vivo measurement. II. Application to Escherichia coli and hybridoma cell metabolism. *J Theor Biol* **155:** 215-242

Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *Bmc Bioinformatics* **11:** -

Schilling CH, Edwards JS, Letscher D, Palsson BO (2000) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng* **71:** 286-306

Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* **99:** 15112-15117

Sekiya T, Takeya T, Brown EL, Belagaje R, Contreras R, Fritz HJ, Gait MJ, Lees RG, Ryan MJ, Khorana HG (1979) Total Synthesis of a Tyrosine Suppressor Transfer-Rna Gene .16. Enzymatic Joinings to Form the Total 207-Base Pair-Long DNA. *Journal of Biological Chemistry* **254:** 5787-5801

Service RF (2005) How far can we push chemical self-assembly. *Science* **309:** 95-95

Shenhav B, Solomon A, Lancet D, Kafri R (2005) Early systems biology and prebiotic networks. *Lect Notes Comput Sc***:** 14-27

Sheridan PP, Freeman KH, Brenchley JE (2003) Estimated minimal divergence times of the major bacterial and archaeal phyla. *Geomicrobiol J* **20:** 1-14

Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *P Natl Acad Sci USA* **102:** 7695-7700

Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology* **26:** 1003-1010

Smallbone K, Simeonidis E, Broomhead DS, Kell DB (2007) Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J* **274:** 5576-5585

Smallbone K, Simeonidis E, Swainston N, Mendes P (2010) Towards a genome-scale kinetic model of cellular metabolism. *Bmc Syst Biol* **4:** 6

Smith HO, Hutchison CA, 3rd, Pfannkoch C, Venter JC (2003) Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc Natl Acad Sci U S A* **100:** 15440-15445

Stephanopoulos G (1994) Metabolic engineering. *Curr Opin Biotechnol* **5:** 196-200

Stephanopoulos G (1998) Metabolic engineering. *Biotechnol Bioeng* **58:** 119-120

Szathmary E (2005) Life - In search of the simplest cell. *Nature* **433:** 469-470

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol*

Terzer M, Maynard ND, Covert MW, Stelling J (2009) Genome-scale metabolic networks. *Wiley Interdiscip Rev Syst Biol Med* **1:** 285-297

Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, Smid EJ (2006) Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. *J Biol Chem* **281:** 40041-40048

Thomas GH, Zucker J, Macdonald SJ, Sorokin A, Goryanin I, Douglas AE (2009) A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. *Bmc Syst Biol* **3:** -

Varma A, Boesch BW, Palsson BO (1993a) Biochemical production capabilities of Escherichia coli. *Biotechnol Bioeng* **42:** 59-73

Varma A, Boesch BW, Palsson BO (1993b) Stoichiometric Interpretation of Escherichia-Coli Glucose Catabolism under Various Oxygenation Rates. *Applied and Environmental Microbiology* **59:** 2465-2473

Varma A, Palsson BO (1993a) Metabolic Capabilities of Escherichia-Coli .1. Synthesis of Biosynthetic Precursors and Cofactors. *Journal of Theoretical Biology* **165:** 477-502

Varma A, Palsson BO (1993b) Metabolic Capabilities of Escherichia-Coli .2. Optimal-Growth Patterns. *Journal of Theoretical Biology* **165:** 503-522

Vickers CE, Archer CT, Kim JF, Jeong H, Park JH, Lee SY, Nielsen LK (2011) The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. *Bmc Genomics* **12**

Vo TD, Greenberg HJ, Palsson BO (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J Biol Chem* **279:** 39532-39540

Watve MM, Dahanukar N, Watve MG (2010) Sociobiological Control of Plasmid Copy Number in Bacteria. *Plos One* **5:** -

Westerhoff HV, Winder C, Messiha H, Simeonidis E, Adamczyk M, Verma M, Bruggeman FJ, Dunn W (2009) Systems Biology: The elements and principles of Life. *FEBS Lett* **583:** 3882-3890

Yamada T, Bork P (2009) Evolution of biomolecular networks - lessons from metabolic and protein interactions. *Nat Rev Mol Cell Bio* **10:** 791-803