An introduction to metabolic networks and their structural analysis

Vincent Lacroix [1,2,3], Ludovic Cottret [1,2], Patricia Thébault [1,2,4] and Marie-France Sagot[1,2]

July 8, 2008

[1] Équipe BAOBAB, Laboratoire de Biométrie et Biologie Évolutive (UMR 5558); CNRS; Univ. Lyon 1, 43 bd du 11 nov 1918, 69622, Villeurbanne Cedex, France.

[2] Projet Helix, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France

[3] Centre de Regulació Genòmica, Genome Bioinformatics Research Group - CRG, PRBB, Aiguader 88, 08003 Barcelona, Spain

[4] Équipe MABioVis, Laboratoire Bordelais de Recherche en Informatique (UMR5800); CNRS; Univ. Bordeaux 2, 351, cours de la Libration, 33405, Talence Cedex, France

Corresponding authors: lacroix/cottret/thebault/sagot@biomserv.univ-lyon1.fr

Abstract

There has been a renewed interest for metabolism in the computational biology community, leading to an avalanche of papers coming from methodological network analysis as well as experimental and theoretical biology. This paper is meant to serve as an initial guide for both the biologists interested in formal approaches and the mathematicians or computer scientists wishing to inject more realism into their models. The paper is focused on the structural aspects of metabolism only. The literature is vast enough already, and the thread through it difficult to follow even for the more experienced worker in the field. We explain methods for acquiring data and reconstructing metabolic networks, and review the various models that have been used for their structural analysis. Several concepts such as modularity are introduced, as are the controversies that have beset the field these past few years, for instance, on whether metabolic networks are small-world or scale-free, and on which model better explains the evolution of metabolism.

keywords: metabolism, network, metabolic data, modelling, topological analysis, modularity, evolution

Contents

1 Introduction

In his December 11, 1907 lecture for the Nobel prize in chemistry, Eduard Buchner said: “We are seeing the cells of plants and animals more and more clearly as chemical factories, where the various products are manufactured in separate workshops. The enzymes act as the overseers. Our acquaintance with these most important agents of living things is constantly increasing”. An even perfunctory look at the computational biology literature will indicate that interest of the community for such cellular “chemical factories” is also constantly increasing to the point where anyone may feel overwhelmed by the amount and variety of papers pertaining to the topic. This survey is meant as an initial guide into chemical factories, that is into metabolism which is defined as “the sum of the physical and chemical processes in an organism by which its material substance is produced, maintained, and destroyed, and by which energy is made available” [1]. The guide is “initial” because providing a complete even if simplified roadmap through the whole of the computational biology literature on metabolism would be beyond the scope of a single paper. We decided therefore to focus our attention on the simplest mathematical model one may draw of metabolism, that of a graph (or hypergraph), and on the questions that may be asked using such a model. These are essentially structural questions on what corresponds to a static representation of metabolism. We shall see that, even within such restrictions, the literature is already vast and controversial enough. We also chose to focus our attention on metabolism exclusively and to intentionally disregard other cellular processes such as gene regulation and signalling. The main reason, besides a problem of space (talking about such processes even in simplified ways would take us too far), is that, at least for now, metabolism is far better documented (with more abundant and reliable data) and thus provides a good stepping stone for initiating the modelling of complex cell factories.

Although this paper concerns "only" structural aspects of metabolism, the topic is of importance as illustrated by the numerous *in silico* analyses that will be discussed at some length in this paper, various of which represent also experimentally fully checked "success stories" (see [18, 159, 169, 176] for a few examples and references on both *in silico* and *in vitro* or *in vivo* work) where the study of structural features such as input-output relationships and maximisation of a product, modularity, pathway redundancy and phenotypic behaviour were able to bring significant insight in the understanding of a metabolic system and in some cases even means to modify and monitor it. As an additional motivation for this review, it is worth stressing that in all these cases, the use of formal models was an essential requirement.

The term metabolism is derived from the Greek word "metabolē" for "change". Its scientific study appears to have started some 400 years ago with experiments performed by Santorio Sanctorius on human, indeed on himself. The experiments involved observing weight fluctuations in his body over the course of a day and during various metabolic processes such as occur when digesting, sleeping, and eating. Through these experiments, published in *Ars de statica medicina* in 1614, Sanctorius introduced the quantitative aspect into medical research, and at the same time founded the modern study of metabolism with, however, an exclusively vitalistic view to explain it. It was two centuries later only, by studying the fermentation of sugar to alcohol by yeast, that Louis Pasteur showed that the organic compounds and chemical reactions found in cells are no different in principle from chemistry. It was however really the discovery of enzymes by Eduard Buchner in the early 20th century that separated the study of the chemical reactions composing the metabolism of an organism from the biological study of its cells.

Such reactions are traditionally grouped into so-called metabolic pathways, which may in turn be classified as anabolic or catabolic. Anabolism is the synthesis of molecules through the use of energy and consumption of reducing agents (a reducing agent is a substance that chemically reduces other substances by donating one or several electrons) while catabolism corresponds to the degradation of molecules yielding energy and the production of reducing agents. The pathways may either be studied in isolation or, since they are overlapping, be combined together to yield what is referred to as a metabolic network. The benefits of studying the whole network rather than individual pathways are numerous and include, for instance, the possibility to explore alternative pathways.

Before getting to a network representation of metabolism however, the first task is to acquire the data. This is no trivial matter and will be described at some length in Section 2. The process can be time-consuming but it is time worth spending in order to avoid the risk of fully or partially invalidating the results obtained with any later analysis, above all their correct biological interpretation. Once data are acquired, modelling them into a network may then start. We discuss in Section 2.3 the various graph-related models (the only ones considered in this survey for the reasons given above) that have been or could be used.

With a graph representation of metabolism in hand, initial analyses are possible. These concern mainly the network topology. Although all such analyses are potentially interesting, one stands out perhaps more than others. This concerns the issue of modularity. It is generally agreed that the notion of modularity is relevant in biology, indeed important for understanding function and evolution. For metabolism in particular, this notion is old if one considers that the metabolic pathways into which the chemical reactions taking place in a cell are traditionally organised and depicted correspond to modules. Glycolysis, and the urea and tricarboxylic acid cycles are probably the most ancient such pathways to have been discovered, the first by Otto Meyerhof around 1940 following the work of Louis Pasteur and others on fermentation, and the last two by Hans Krebs in 1932 and 1937 respectively. The tricarboxylic acid cycle, TCA cycle for short, which corresponds to the sequence of chemical reactions that produces energy in cells, is also known as the Krebs cycle and earned Hans Krebs a Nobel Prize in 1953. The identification of metabolic pathways presents however a certain arbitrariness, particularly at the frontiers of the pathways that are often ill defined. Automatic and formal ways of identifying metabolic pathways, such as those that have been familiar to biochemists since the last century, have therefore been explored, as have other definitions of modules, some of which are operationally rather than biologically motivated. All are presented in Section 3 together with the other network measures that have been applied to metabolism.

Although doubtless useful, in particular for understanding biological processes and also for validating some computational methods developed for analysing metabolism, the search for modules should not overshadow the need sometimes to work with the full network. This will depend on the biological question that is asked. Studying the evolution of metabolism may be one occasion where using the full network is in some cases required. Work on this essential topic is discussed in Section 4. It may be argued that structural analyses should not be divorced from evolutionary considerations as the latter are important for keeping a check on the soundness of the first. For the sake of exposition however, such separation appears to be helpful.

The reader who wishes to skip information on how data is acquired and is knowledgeable on how metabolic networks may be represented using graph or constraint-based models, may safely go directly to Sections 3 or 4.

2 Acquisition of metabolic data

2.1 Defining the entities

A *metabolic network* may be formally defined as a collection of objects and the relations among them. The objects correspond to chemical compounds, biochemical reactions, enzymes and genes (see Figure 1).

Chemical compounds, also called *metabolites*, are small molecules that are imported/exported and/or synthesised/degraded inside an organism. For most metabolites, the amount observed varies depending on the tissue and cell compartment inside which the compound is present. Tissues and cells indeed contain a number of liquid compartments separated from one another by selectively permeable membranes.

Biochemical reactions produce a set of one or more compounds (called the *products*) from another set of one or more compounds (called the *substrates*). In theory, a chemical reaction can occur in both directions. However, under particular physiological conditions, some reactions occur in only one direction. In this case, they are defined as being *irreversible* if all other conditions remain constant. Inside a cell, some reactions are spontaneous but most are *catalysed* by one or several enzymes which strongly accelerate their speed. An *enzyme* is a protein or a protein complex, coded by one or several *genes*. A single enzyme may accept distinct substrates and may catalyse several reactions, and conversely, a single reaction may be catalysed by several enzymes. Elucidating the links between genes, proteins and reactions (the so-called GPR relationship) is not a trivial task and is a major concern in metabolic reconstruction [?] as is discussed in the next section.

The presence of small molecules, called *cofactors*, is sometimes essential to allow the catalysis of a reaction by an enzyme. Such molecules can vary for a given reaction among several organisms. By binding the enzyme, cofactors can enhance or decrease the activity of the enzyme. They are called, respectively, *allosteric activators* or *allosteric inhibitors.* The term allostery indicates that the regulatory site of an allosteric protein is separate from its active site.

Data on individual reactions have not always been available and the concept of metabolic pathway has often been used to informally characterise the set of reactions involved in the synthesis or degradation of a molecule of interest (glycolysis for the transformation of glucose into pyruvate for instance). The concept of metabolic pathway remains very much employed for historical reasons but lacks formal definition. In particular, there is no consensus on the boundaries of a pathway. Efforts were made in recent years to propose a more formal definition [52, 98, 152, 157] but no general agreement has been reached yet.

2.2 Metabolic reconstruction

Reconstructing a metabolic network consists in inferring the relations between genes, proteins (enzymes) and reactions in a given metabolic system. This is usually achieved using comparative genomics but also, often as a refinement step, using metabolomic data (the latter referring to the type and quantity of metabolites present in the metabolism of an organism).

Other types of relations may be more difficult to assess. This is the case, for instance, of the allosteric effects of an enzyme which are rarely known. The precision needed in the definition of each relational link depends also on the question one wishes to answer. In some cases, it suffices to obtain a list of metabolites associated with a list of the chemical transformations to which they are linked. On the other hand, if the aim is to study, for instance, the relationship between genotype and phenotype, then it becomes necessary to establish a precise correspondence between reactions and the enzymatic genes whose protein products catalyse them.

Inference from comparative genomic data usually suggests a list of metabolic reactions present in an organism of interest, not the whole network. Although the process has been highly automated in recent years, it still requires in most cases an expert manual intervention sustained by data painstakingly collected from the literature. Once a model for the whole network, such as a graph, has been obtained a study of its general mathematical properties can start (see Section 3).

The quality of such a reconstruction obviously strongly depends on the quality of the genome annotation for that organism and, to a lesser degree, on the taxonomic position of the input organism. Indeed, there exist few organisms for which the set of genes composing their genomes and their associated functions (metabolic or other) are well known. This is the case, for instance, of *Escherichia coli.* EcoCyc [91], the part of the metabolic database BioCyc [85] dedicated to *E. coli*, thus offers an excellent level of accuracy due in no small part to its numerous links to experimental evidence. This kind of database is however an exception among genome-scale metabolic databases. Besides being often much less accurate, most of the pathways in BioCyc or in KEGG, another widely used database for metabolism [7], are further biased towards the microbial and plant kingdoms. The quality of the metabolic reconstruction of any animal is thus expected to be worse than the one of a bacterium evolutionarily close to *Escherichia coli.*

Metabolic reconstruction from comparative genomics is traditionally divided in two parts. The first provides a functional annotation of the metabolic genes, *i.e.* determines the catalytic activity of the enzymes the genes code for. The second defines the relation between functional annotations and biochemical reactions, *i.e.* establishes the list of reactions enabled by the annotated catalytic

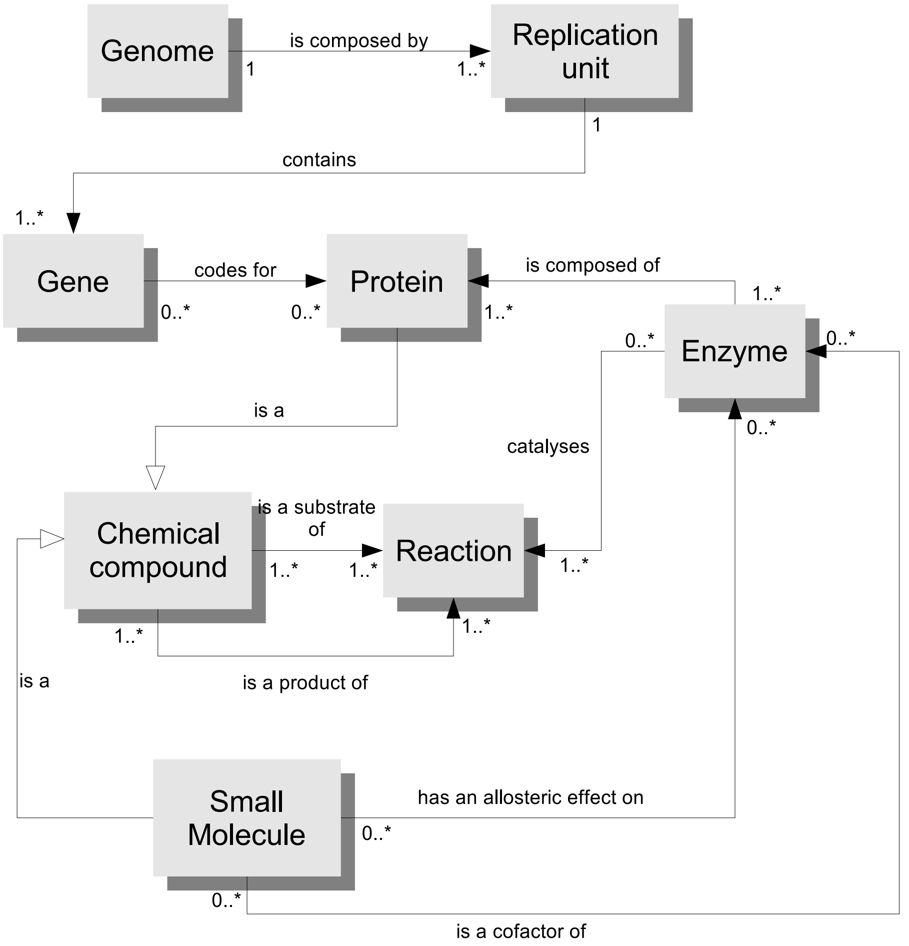


Figure 1: Simplified UML schema of the various objects involved in a metabolic network. The indications on each side of an arrow representing a relation between two objects define the cardinalities of each object in the relation: “1” means exactly one, “0..*” means zero or more, “1..*” means one or more. For instance, the indications on each side of the relation “codes for” between “Gene” and “Protein” mean that one gene can produce several proteins (in the case of alternative splicing for instance) or none if the gene is not a protein gene (this explains the “0” in the cardinality for the proteins), and that a protein can either be produced by more than one gene or supplied by the environment (this explains the “0” in the cardinality for the genes).

activities.

2.2.1 Functional annotation of the metabolic genes

Recent high-throughput genome sequencing techniques have given access to many complete genome sequences. For instance, the EBI genomic database (http://www.ebi.ac.uk/genomes/) contained at the time of writing this paper the complete genomes of 575 bacteria or viruses, 51 archea and 74 eukaryotes. Most genomes have been annotated using fully automated methods. The first step consists in detecting the boundaries of the genes and in assigning a function to the protein(s) such genes code for. Several complementary methods are currently used to identify the limits of a gene, which proceed by putting together information from, among others, the detection of open reading frames and of conserved motifs around the junctions between introns and exons (in the case of eukaryotes), as well as alignments with already known genes. The latter should enable also to determine the function(s) of the genes. Indeed, this is most often predicted by sequence comparison of the related protein with proteins of known function(s) in already sequenced genomes.

In the case of a metabolic network reconstruction, it is particularly important to be able to establish the catalytic function(s), if any, of a protein. Functions are specified either based on experimental clues or by using automatic methods. Because of the frequency of new completely sequenced genomes, it is nowadays impossible to provide annotations based on experiments for each protein, and automatic methods are therefore the most common means of assigning function(s) to a protein. As pointed out in [58], the first challenge is however in arriving at an unambiguous and operational (for the purpose of an automated prediction) definition of “function”.

Protein function The ambiguity comes from the fact that protein function may depend on which aspect, biochemical or physiological for instance, is considered. It can also be highly context-dependent as exemplified by the so-called “moonlighting proteins” which have different functions inside a same organism depending on their cellular localisation, cell-related level of expression, oligomeric state, and cellular concentration of a ligand, substrate, cofactor or product [78]. Such characteristic can go unsuspected for years.

Ontologies such as GO (“Gene Ontology”) [10], by establishing a common and controlled vocabulary for the functional description of homologous genes in multiple organisms, have made gene annotation easier and in general more consistent, and have improved data exchange and curation. A so-called EC (for “Enzyme Commission”) numbering system is used to classify enzymes by type and function by means of a hierarchical code with 4 digits separated by dots attributed by the IUBMB (International Union of Biochemistry and Molecular Biology). Each digit represents a progressively finer class [177]. The first one defines the type of reaction catalysed (1. oxidoreductases, 2. transferases, 3. hydrolases, 4. lyases, 5. isomerases, 6. ligases), the next two further indicate, respectively, the reaction’s subclass and sub-subclass, and the fourth is the serial number of the enzyme in its sub-subclass (see http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html for more details). Unfortunately, the assignment of an EC-number, even when correct, can be imprecise. For instance, the EC number 2.7.1.69 corresponds to 60 genes in *Lactobacillus plantarum* [175]. Moreover, at current time, some 30 to 40% of the metabolic activities biochemically characterised, that is experimentally known to exist in nature and to which an EC number has been assigned, remain orphan in the sense that no gene responsible for the corresponding reactions has ever been identified in any organism [23, 106, 137].

Orthology approaches The classical way to annotate a whole-genome set of proteins is to perform sequence comparison between them and the proteins of other species whose genomes have

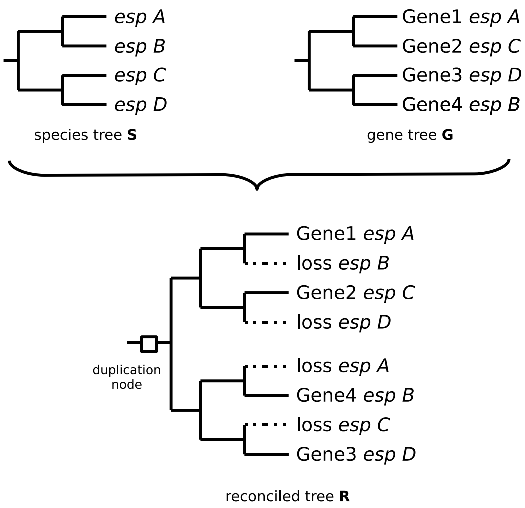


Figure 2: Reconciliation of the gene tree G and the species tree S into the tree R.

already been annotated in order to detect homology, more precisely orthology. The corresponding function assignment is indeed based on the assumption that orthologous enzymes have same activity. Orthologous proteins are homologs in different species that originated from a single ancestral gene after a speciation event [53]. One major difficulty with such methods is to distinguish orthologs from paralogs. The latter result from the intra-genomic duplication of an ancestral gene and are assumed to have different function(s). Two main approaches exist to address this problem.

The first is based on an all-against-all reciprocal sequence comparison of the proteins encoded in complete genomes using, for instance, Blast [6]. Two proteins are classified into the same group if they are more similar between them than to any other proteins from the same genomes, that is, if they are each other's Best-Reciprocal-Hits (BRH). BRH methods are well adapted for large-scale data and are at current time commonly used to annotate new sequenced genomes. COG ("Clusters of Orthologous Groups") [173] is the first database built in this way and is certainly the most famous one. The Kegg Orthology (KO) system [83] is based on the same concept and is also widely used, particularly for metabolic reconstruction. Neither COG nor KO are however able to take into account gene duplications which may occur after a speciation event and lead to the presence of co-orthologs in the same genome. The latter have more recently been considered by Remm *et al.* to improve the clustering of orthologous groups [146].

The second main approach to distinguish speciation events (evidence of an orthology link) from duplication events (evidence of a paralogy link) is based on a comparison of gene trees G with a reference species tree S. The result of the comparison is a "reconciled" tree R which is a variation of the species tree in which duplication nodes have been inserted in order to explain incongruences with a gene tree [41] (see Figure 2).

Automatic "tree reconciliation" methods were developed for genome-scale studies but require completely resolved gene and species trees. To circumvent this problem, Dufayard *et al.* proposed a method allowing for the presence of a number of unresolved nodes both in the gene and the species trees [40]. Several databases are built using a tree reconciliation method, among which are HOVERGEN [41] for vertebrates, INVHOGEN [132] for invertebrates and TreeFam [107] for animals. However, tree reconciliation methods present various drawbacks [102]: they are based on still poorly characterised models of gene duplication and loss, the species phylogeny used (in general this follows the taxonomy provided by the NCBI) contains a large number of unresolved nodes, and last, computing a reconciled tree is time-consuming.

Whatever specific method is chosen, all orthology-based approaches assume that orthologs have same function(s) while paralogs have different function(s). However, closely related paralogs may have more similar functions than distantly related orthologs. Moreover, it is not always the case that orthologs have same function(s). They may instead present rather different functionalities, even when the percentage of conserved bases among aligned orthologs is very high [58, 119].

Sequence signature approaches Since the function of a protein is essentially determined by one or several active domains, identifying sequence signatures may appear to be more appropriate for predicting protein function(s). Several databases, such as Interpro [118] and ProDom [160], propose an automated clustering of homologous domains that can then be used to establish the function(s) of a protein. More specifically in the context of metabolic genes, Claudel-Renard *et al.* introduced a method, called PRIAM, [26], that assigns enzymatic activities based on a classification into modules of the enzymes in the ENZYME database [11]. An enzyme module is defined as a longest homologous segment shared among an enzyme collection.

The increase in the number and diversity of the sequences available in databases make all homology-based methods error-prone, whether they are based on sequence comparison or on common sequence signature detection. As more proteins are automatically annotated, the propagation of such annotation errors may quickly escalate [58]. Indeed orphan genes may be associated to a sequence in another genome where no experimental evidence exists that would permit to link the orphans to that or to any other homologous sequence [137]. Moreover, some proteins with undetectable sequence similarity may have, in fact, the same function (they are called "analogs") [60].

Genomic context approaches Information provided by homology-based methods can be complemented by investigating protein co-evolution. The hypothesis is that functionally linked proteins evolve in a correlated fashion. If N genomes are considered, a phylogenetic profile is established for each protein that corresponds to a vector of bits of length N, each bit indicating the presence or absence of a homolog in a given genome. Proteins are then clustered, and function(s) determined according to the similarity of their phylogenetic profiles [133].

In prokaryotes where functionally related genes tend to be co-regulated and co-localised on the chromosome, groups of contiguous genes are in general preserved across different genomes, with conserved local order. This information can be used to infer the function(s) of the protein a gene encodes, even in the absence of any similarity with other sequences [149, 181]. Is has been shown also that genes linked by fusion events generally code for functionally related proteins [201, 44].

Other approaches Expression microarrays, protein-protein interaction networks and information on cellular localisation may provide further clues for the assignment of function to a protein. In this case, only a general function at the cellular level may be inferred, not a precise molecular function (such as a complete EC number). We refer the interested reader to [58] for an overview of these methods.

To avoid the drawbacks inherent to each of the above approaches, attempts have been made to combine them in an efficient way. The method of Chua *et al.* thus models into a same weighted graph information about sequence homology, protein-protein interactions, protein domains, gene expression data and literature [24]. Some "missing genes" approaches (see further details in Section 2.2.3) attempt also to integrate several kinds of information to identify genes for missing enzymes [126, 200].

Manual refinement After any automatic gene annotation, doing a manual refinement is essential. Several annotation platforms, such as GenDB [115], MaGe [181] or Iogma (http://www.genostar.org), provide powerful graphical interfaces to help experts to curate automatically generated annotations. Since it is very difficult to have an expert group to annotate all the genes in a single genome, Overbeek *et al.* propose an approach where all genes occurring in a "subsystem" (such as a metabolic pathway for instance) are analysed over a large collection of genomes by an expert in that subsystem [125].

Various currently available projects try to provide improved genome annotations by using curated and up-to-date genomic sequences. This is the case, for instance, of RefSeq [138], GenomeReviews [90] Ensembl [77] and Hamap (for "High-quality Automated and Manual Annotation of microbial Proteomes"). The latter uses both manual and semi-automatic curation to obtain the complete proteomes of sequenced bacteria [62]. The project Integr8 gathers data from both GenomeReviews and Hamap [90].

Despite the rapid growth-rate of new methods and data to determine protein function(s), the fractions of genes for which no function can be predicted is still high (around 40%), especially in eukaryotes, and remains a major problem [64, 137]. It is important to keep in mind this observation since, whatever the efficiency of the metabolic reconstruction methods, the metabolic capabilities of an organism will remain incomplete as long as the assignment of protein function(s) is incomplete or imprecise.

2.2.2 Defining the set of metabolic reactions and pathways

Once enzyme functions have been tentatively defined, links must be established between them and the corresponding biochemical reactions. The ENZYME database describes each type of characterised enzyme for which an EC number could be provided, and contains links to various metabolic databases and to Uniprot [11], which is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequences and functions created by joining the information contained in Swiss-Prot, TrEMBL, and PIR. The BRENDA database provides additional detail, when available, concerning the substrate specificity of the reactions across several organisms [154]. The BIOCYC [85] and KEGG [84] databases, currently the most widely used, store information on genomic data, proteins, reactions, compounds and metabolic pathways. Perhaps the two most currently used tools to automatically infer a set of biochemical reactions from genomic data are KAAS, the KEGG Automatic Annotation Server [116], and PathoLogic, the prediction tool in the Pathway-Tools used to build, query, and visualise BioCyc-like databases [87]. The two adopt a somewhat distinct strategy. In KEGG, the genes present in the databases are identified by their Kegg-Orthology number (KO number) which groups orthologs coding for the same enzymatic activity. A biochemical reaction is considered as possible in an organism if its genome contains a gene that is classified by sequence comparison into the KO group corresponding to the EC number of the reaction. PathoLogic does not do sequence comparison but assumes instead this has already been performed and the information stored, for instance, in GenBank, the comprehensive NIH-maintained genetic sequence database [13]. As a consequence, if an EC number has been assigned to a gene product (the EC number can be retrieved from the EC_NUMBER qualifier in GenBank), the corresponding reaction in a database called MetaCyc [22] used by PathoLogic is added to the list of reactions feasible in the organism. MetaCyc stores information on nonredundant, experimentally elucidated metabolic pathways for more than 900 organisms. If a gene has no EC number assigned, its assigned functions are matched against the list of names of functions present in MetaCyc. To a great extent, the strategy of KAAS is thus to re-do the assignment of a function for each gene while the strategy of PathoLogic is to use the pre-existing annotations. This means that the quality of the predictions in PathoLogic highly depends on the quality of the annotations. On the other hand, PathoLogic can also take into account information that does not come from sequence comparison. The main advantage of the KAAS metabolic reconstruction is that it can start its assignment from draft genomes.

Historically, metabolic pathways are sets of reactions corresponding to a metabolic function. For instance, glycolysis is a sequence of reactions involved in the degradation of glucose (a molecule with 6 carbon atoms) into pyruvate (a molecule with 3 carbon atoms) and other molecules needed for the biosynthesis of macromolecules constitutive of a cell. Glycolysis is one of the metabolic pathways (such as the TCA cycle and the pentose phosphate pathway for instance) which occurs in almost every organism. Reference metabolic pathways have thus been built to represent common and alternative organism-dependent reactions. By comparing the set of reactions expected to occur in an organism with the set of reactions in reference metabolic pathways, it is possible to infer the main metabolic functions of an organism. In KEGG, the reference metabolic pathways are organised into metabolic maps where all known variants are drawn together. From a list of reactions or even gene identifiers defined by KAAS, it is possible to highlight the corresponding reactions in each reference metabolic map. PathoLogic additionally attempts to predict which pathways are susceptible to occur in the input organism. The prediction is based on the proportion of metabolic reactions in the pathway for which there is an evidence [126] and on the presence of unique reactions. Indeed, reactions occurring in only one pathway provide a stronger evidence for the presence of a metabolic pathway than reactions occurring in several other pathways.

Neither KAAS nor PathoLogic can predict novel pathways. For this, the Pathway Hunter Tool (PHT) [143] can be used. Starting from a set of annotated EC numbers, and a source/destination metabolite pair selected by the user, the shortest metabolic pathways (k-shortest pathways) are computed by PHT. These provide alternative routes that may then be evaluated for biological significance.

A genome-based metabolic reconstruction enables to very quickly obtain a draft of the relations between genes, enzymes and reactions, and, for instance, to study the impact of some genomic events (such as duplications, transfers or deletions of genes) on a metabolic network. However, this draft reconstruction often contains errors or imprecisions that must in general be further painstakingly and expertly curated with help from the literature or from other more refined methods which are described next.

2.2.3 Refinements of a metabolic reconstruction

Missing genes and fragments in a metabolic reconstruction After any metabolic reconstruction, some reactions appear as “missing”. They correspond to gaps in the biochemical pathways that were nevertheless declared as being present in the reconstruction. We observe that there seems to be no established quantitative criterion in the literature to decide whether a pathway is present in a metabolic reconstruction. Any such missing reaction can be explained [29, 124] by:

1. a low sequence similarity of the corresponding gene with the reference ones coding for the missing enzyme in other organisms;

2. the products of the reaction can be obtained from alternative pathways or are provided by the environment;

3. the presence of multi-enzyme proteins;

4. the corresponding gene is really absent in the genome.

Various approaches exist to complete these gaps by trying to identify the genes encoding for a specific metabolic function. Like most methods used in this area, they are based on ad-hoc heuristics. Osterman *et al.* propose a very good review on the use of comparative genomics to search for missing genes [124]. Indeed, various types of genomic evidence can be combined to propose candidate genes for a missing reaction: genes coding for enzymes involved in the same pathway are often chromosomal neighbours (in prokaryotes, the case of eukaryotes is less clear) [66, 63, 93]; protein fusion events provide some evidence of potential functional coupling [93]; two proteins which appear in the same metabolic pathways tend to occur together or not in any specific organism [93]; genes functionally coupled are expected to be co-regulated [93]. Inference of the missing reactions may be performed using a supervised approach (such as Support Vector Machine) which requires partial knowledge of the network and good training sets [66, 200]. Candidate genes have their similarity to known genes measured according to the relative importance of each of the previously cited evidence. These methods are interesting for both metabolic reconstruction and gene annotation.

When the missing reactions represent a pathway fragment, another approach [21] may be used to infer it by comparison with a set of known pathways. The latter are previously modelled as enzyme graphs labelled with GO annotations. Data mining techniques are used to detect frequent pathway functionality patterns in the graphs and these are then matched onto the metabolic network of interest. The underlying idea is to focus on the biological processes carried by the various steps along a pathway rather than on the specific genomic or chemical features of each individual enzyme [21].

Reversibility of the reactions The direction of a reaction in certain physiological conditions is determined by the thermodynamic properties of the reaction, the kinetic properties of the enzyme and the concentration of its substrates and products. In a metabolic reconstruction, the direction of a reaction is often added *a posteriori* and manually, or the reaction is left as reversible. Methods to automatically assign a direction to the reactions in a genome-scale metabolic network reconstruction remain rare. Yang *et al.* [202] show how the direction of a reaction may be determined by analysing the stoichiometric matrix of the metabolic network. The *stoichiometric matrix* of a network gives the coefficients for all reactions in a metabolic network. The *stoichiometric coefficient* for a given metabolite in relation to a given reaction represents the degree to which the metabolite participates in the reaction. For instance, if the reaction R is $A + 2B \rightarrow C + 3D$ (reactant metabolites A and B are transformed into product metabolites C and D by reaction R), then the stoichiometric coefficients of A, B, C, and D are, respectively, 1, 2, 1, and 3. In general, stoichiometric coefficients are integers since elementary reactions always involve whole molecules. Obviously, the coefficient is zero for metabolites which do not participate in the reaction. Yang *et al.* [202] showed how, using the stoichiometric matrix, feasible reaction directions in a given system may be computed from those imposed by the reactions at the boundary of the system that transport metabolites in or out. However, the algorithm was tested on a reaction network of only 44 reactions. The algorithm proposed by Kümmel *et al.* [103] exploits instead all available experimental thermodynamic data, given by the derived Gibbs energies of formation, and metabolite concentrations to identify irreversible reactions. The standard Gibbs free energy of formation of a metabolite is the change of Gibbs free energy that accompanies the formation of 1 mole (the mole is a counting unit) of that substance from its component elements, at their standard states (the most stable form of the element at 25 degrees Celsius and 100 kilopascals). The Gibbs energy of a reaction can intuitively be thought of as the maximum amount of work obtainable from a reaction. Next, Kümmel *et al.*'s algorithm aims at identifying, on the basis of some heuristic rules, sets of reactions (subnetworks) whose simultaneous operation is thermodynamically infeasible. A thermodynamically infeasible subnetwork may, for instance, be a cyclic operation of a reaction set such as the example given in [103] of three reactions $A \leftrightarrow B$, $B \leftrightarrow C$ and $A \leftrightarrow C$. If A is converted to B, and B to C, then C must have a lower Gibbs energy of formation than A meaning that the reaction $C \rightarrow A$ is not feasible. The algorithm was tested on a genome-scale model of *Escherichia coli* (920 reactions) and assigned 130 of the 920 reactions as irreversible. Feist *et al.* also propose a genome-scale metabolic reconstruction for *Escherichia coli* K12 which includes thermodynamic information [47]. The directions are inferred from the values of standard Gibbs free energy change of formation for most metabolites and reactions, estimated from the structure of the metabolites [114].

Inferring the presence of reactions from metabolite data Large-scale techniques have recently been developed to determine the metabolome of an organism, *i.e.* the set of its detected metabolites. The techniques are grouped under the term "metabolomics". A good description of those currently used, with their advantages and limitations, has appeared already in several reviews [14, 69, 122]. The catalogue of metabolites thus produced provide additional information about which compounds are really present inside an organism. Indeed, because of incomplete knowledge on substrate specificity, reconstruction from sequence annotation gives only partial and approximate knowledge on the metabolites participating in a metabolic network.

Some methods suggest feasible biochemical reactions from sets of metabolites. For instance, Arita uses hypothetical links (16 basic types) between compounds to infer biochemical reactions even if they do not correspond to a known enzyme [8]. In the same way, Kotera proposes a method to assign partial EC numbers from a set of substrates and a set of products [101]. Breitling *et al.*'s approach on the other hand relies on the development of ultra-high resolution mass spectrometers and on the fact that only a limited repertoire of chemical transformations account for the majority of biochemical reactions within cells [17]. This enables to infer feasible biochemical reactions by computing accurate mass differences between compounds and then referring to a table which provides correspondences between mass difference and chemical reaction. The main advantage of such a method is that no genomic information is required. The disadvantage is that it does not provide any information about the relations between genes, enzymes and reactions.

Refining metabolic reconstruction using biological experimental evidence Mass spectrometry for the large-scale identification of the proteins in an organism enables to confirm the presence of enzymes predicted from genomic annotations [182, 191]. In the same way, several other high-throughput techniques currently exist to define the metabolome, that is, the set of metabolites present in an organism [14, 51, 57, 122]. High-throughput phenotyping and gene expression data may also be used together with the predictions of a computational model in order to refine a metabolic network [31]. At a smaller scale, physiological information may provide important additional clues to complete the set of reactions in a metabolic network. For instance, in the metabolic model of *Streptomyces coelicolor* built by Borodina *et al* [16], 89% of the reactions have an associated annotated gene while the remaining reactions were included based on physiological evidence. The purification of an enzyme and study of its catalytic activities following functional annotation or mass spectrometry may also further enable to precise its substrate specificity.

The analysis of admissible fluxes in a network, together with incorporated constraints that limit the network behaviour with respect to feasible steady state flux distributions, has also been used in order to predict the phenotypic behaviour of the network as a function of such constraints. At each iteration, the phenotypic predictions computed by such an analysis are compared with experimental observations. If the model predictions do not correlate with the experiments, hypotheses about the functions of an organism are generated to expand the model [16, 39].

2.3 Modelling metabolism

Metabolism can be studied from a structural or from a dynamic point of view. Distinct models are used in each case [166], the most common being graphs, so-called constraint-based models and differential equations. The latter is necessary for dynamic studies. Since this review is focused on the structural analysis of metabolism, we mostly cover the first two types, with an emphasis on graph models. The reader interested in dynamic aspects of networks and a representation by differential equations of the latter may refer to [170].

2.3.1 Graphs

Formally, a graph G is defined as a couple (V, E) with V a finite set of nodes and E a set of edges that is a subset of V^2. Modelling a metabolic network with a graph means choosing which biological entities will be associated to nodes and edges. In the context of metabolism, biological entities may be compounds, reactions or enzymes. We start by describing the graph models more frequently found in the literature and discuss in what context they may introduce ambiguities.

The question of which is the right graph model to choose depends mainly on the type of question asked, whether, for instance, the focus is on connectedness only, or if knowing the exact pattern of connection is necessary.

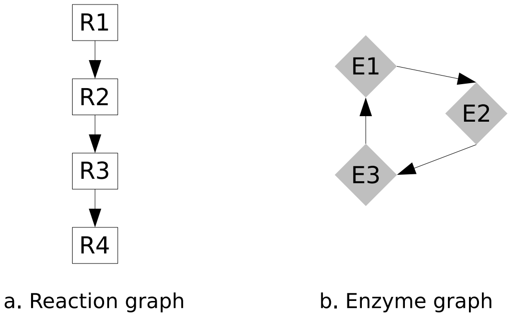


Figure 3: Reaction graph and enzyme graph for the sets of reactions: R1: $A + B \rightarrow C + D$, R2: $D + E \rightarrow F + G$, R3: $F + G \rightarrow H + I$, R4: $I \rightarrow J + K$. E1, E2 and E3 are enzymes: E1 catalyses R1 and R4, E2 catalyses R2 and E3 catalyses R3.

More commonly used graph models A compound graph is a model of the metabolic network where the nodes correspond to compounds and there is an edge between compounds A and B if there exists a reaction where one is substrate and the other is product. In a reaction graph, nodes correspond to reactions and there is an edge between two reactions if there exists a compound which is produced by one and consumed by the other. Enzyme graphs are also sometimes used. In this case, nodes correspond to enzymes and there is an edge between two enzymes if they catalyse reactions that share a compound. Using the enzyme graph may lead to confusion in the structure of the network, *i.e.* reactions catalysed by a same enzyme are “merged”, thereby creating shortcuts between distant compounds (see Figure 3).

Nevertheless, this model may still be adopted if the focus is really on enzymes and on the relations between them, as in [75]. One should call attention to the fact that enzyme and reaction graphs are often mistaken for one another although they are really different. Using a labelled graph (that is, a graph where nodes and edges may be labelled), enzymes can be introduced as reaction labels, either in the compound graph or in the reaction graph. This ensures that the structure of the network is preserved.

Compound and reactions graphs may sometimes be ambiguous [37]. For instance, the two following networks lead to the same compound graphs as indicated in Figure 4 a.

Network 1	**Network 2**
Reaction 1: $A \leftrightarrow C$	Reaction 1: $A + B \leftrightarrow C$
Reaction 2: $B \leftrightarrow C$	Reaction 2: $C \leftrightarrow D$
Reaction 3: $C \leftrightarrow D$	

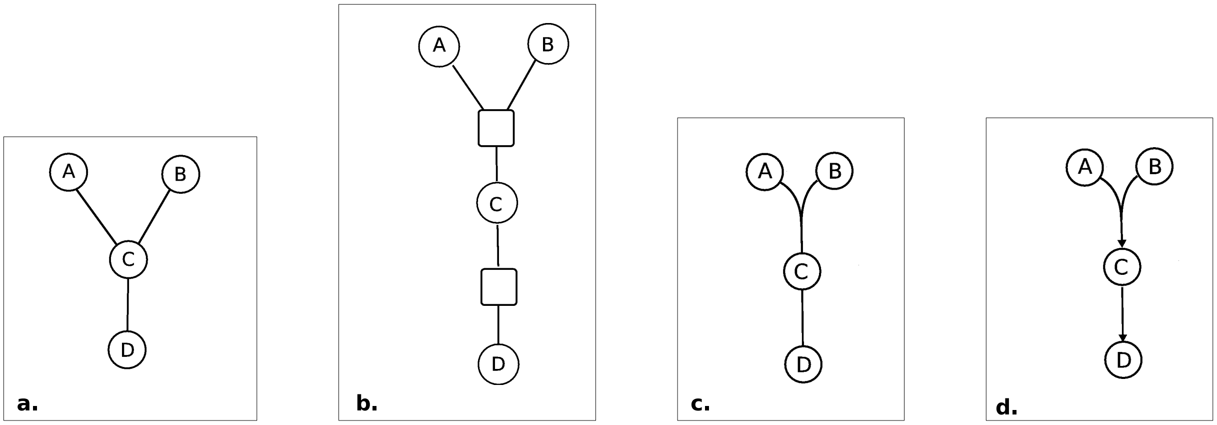


Figure 4: **a.** Compound graph for the sets of reactions: set 1. $A \leftrightarrow C$, $B \leftrightarrow C$, $C \leftrightarrow D$; and set 2. $A + B \leftrightarrow C$, $C \leftrightarrow D$. **b.** Bipartite graph for the set of reactions: $A + B \leftrightarrow C$, $C \leftrightarrow D$. **c.** Undirected hypergraph corresponding to the network: $A + B \leftrightarrow C$, $C \leftrightarrow D$. **d.** Directed hypergraph corresponding to the network: $A + B \rightarrow C$, $C \rightarrow D$.

One may resolve this ambiguity by adding reactions as edge labels. Another method consists in using a more expressive graph model: either a bipartite graph or an hypergraph.

Most expressive models Formally, a bipartite graph is a graph whose set of nodes V can be divided into two disjoint subsets V_1 and V_2 such that each edge has a node in V_1 and the other in V_2. In the context of a metabolic network, one of the sets corresponds to compounds and the other to reactions. In the case of the example of Networks 1 and 2 above, the ambiguity would be avoided as shown in Figure 4 b.

Another way of solving the ambiguity of simple graphs is to use a hypergraph. Intuitively, a hypergraph is a graph where the edges may link more than two nodes. Formally, a hypergraph H is a pair (V, E), where $V = \{v_1, v_2, ..., v_n\}$ is the set of nodes and $E = \{e_1, e_2, ..., e_m\}$, with $E_i \subseteq V$, for $i = 1, ..., m$, is the set of hyperedges. Clearly, if $|E_i| = 2$, for all $i = 1, ..., m$, then the hypergraph is a simple graph. In order to model a metabolic network with a hypergraph, nodes are usually associated to compounds and hyperedges to reactions (see Figure 4 c).

The models mentioned previously were undirected. A directed graph is a graph where each directed edge, that is arc in mathematical notation, is an ordered pair of nodes, one of which is the source node and the other the target node. In the case of a hypergraph, a directed hyperedge, or hyperarc, may have several sources and several targets (see Figure 4 d.).

Orientation of edges generally enables to model the direction of a reaction. In the case where all reactions are irreversible (resp. reversible), the network will be modelled as a directed (resp. undirected) graph. In the case where only some reactions are reversible, one may use a mixed graph (some edges are directed and some are not). In several papers [50], the assumption is made that all reactions are reversible. Indeed, we can argue that in presence of an excess of product, even reactions which have a favoured direction may occur in the opposite direction.

Finally, one should notice that modelling a reversible reaction with undirected edges may still be ambiguous (several networks may have the same representation), even when using bipartite graphs. If we go back to the example given in Figure 4 b., the same graph could have been obtained using the set of reactions: $A \leftrightarrow B + C$, $C \leftrightarrow D$. Indeed, there is no indication in the graph on which compounds are on which side of the reaction; no difference is made between right compounds and left compounds. One way of solving this problem is to use edge labels. Such edge labels may also help in avoiding to compute biologically meaningless paths which link compounds that are on the same side of a reaction (a substrate to another substrate for example). We may observe that labelled bipartite graphs and hypergraphs are strictly equivalent in terms of the quantity of modelled information. One may easily transform one into the other.

Possible model simplifications In all models presented above, it is commonly assumed that all reactions and all compounds are equivalent. In practice, for a given reaction, some compounds may be considered as primary and others as auxiliary. For instance, ATP and NADH are involved in many reactions as, respectively, energy source and reducing agent. Considering them as regular compounds to build a graph model would lead to artefactually link a very large number of reactions.

A classical way of dealing with this problem is to remove from the network all compounds that participate in a large number of reactions [80, 110, 109]. This method has several caveats: 1. one needs to heuristically define a threshold to determine how many metabolites to remove, 2. some highly connected metabolites like pyruvate or fructose will be removed whereas it is commonly agreed that they belong to the backbone of metabolism, and 3. even compounds like ATP that could be safely removed from most reactions, should not be eliminated from those that participate in their own synthesis.

Another option is to remove compounds only in the reactions where they are involved as side compounds, *i.e.* remove edges and not nodes of the graph [104]. The idea is to maintain the backbone of metabolism and to get rid of shortcuts. The distinction between side compounds and main compounds for each reaction is however not always available. It has been automatically generated in BioCyc, initially for drawing purposes [86]. KEGG maps do not represent all compounds either.

Applying no treatment to such highly connected compounds may lead to misleading conclusions when computing network measures as we shall see in Section 3.1.1.

2.3.2 Constraint-based modelling

The term “constraint-based modelling” was coined, and subsequently adopted by the bioinformatics community, following two papers by the Palsson group [32, 127]. In a constraint-based framework, the network is modelled by its stoichiometric matrix (see Section 2.2.3). This actually corresponds to a directed edge-labelled bipartite graph (or equivalently, hypergraph). In this case, the labels are the stoichiometric coefficients of the compounds in the reactions, with signs to indicate if they are produced or consumed. Reversibility of the reactions is not handled within the matrix representation but provided as a separate information.

In constraint-based modelling, the focus is on the distribution of mass fluxes through the reactions under some constraints. The principal motivation is to analyse the metabolic capabilities of an organism. The main constraints that have been considered are: 1. steady state (every metabolite that is produced has to be consumed); and 2. thermodynamic constraints (irreversible reactions can only be taken in the appropriate direction).

In the following, we introduce the concept of a flux vector, denoted by v. A flux vector (or flux distribution) is an m-vector of the space of reactions $\Re^m$, where the element v_i describes the flux through reaction i and $\Re = Rev \cup Irrev$ with Rev (resp. $Irrev$) the set of reactions that are reversible (resp. irreversible). In this context, the term flux is equivalent to the net rate of the reaction, that is to the difference between the rate of the direct reaction and the rate of the reverse reaction.

If S is the stoichiometric matrix for the network, the two constraints previously mentioned (steady state and thermodynamic) can be expressed as follows:

1. $Sv = 0$
2. $v_i > 0, \forall i \in Irrev.$

They define a portion of the flux space represented by a convex polyhedral cone containing all admissible flux vectors. A flux vector is also called a *mode*.

Given this framework, two main issues have been addressed. The first, called Flux Balance Analysis (FBA), is concerned with finding an admissible flux vector which optimises an objective function. The objective functions that have been considered in the literature include biomass or ATP production. FBA has many applications and has been shown to have good phenotypic predictive power [43]. FBA may also be used to measure the phenotypic effects of complete or partial metabolic gene deletions and other types of perturbations on a system. Gene deletion studies are in general performed by constraining the reaction flux(es) corresponding to the gene(s) (and associated proteins(s)) to zero and applying FBA on the network. This approach assumes that the corresponding mutant system will display an optimal metabolic state, but as indicated in [158], knockouts probably do not possess a mechanism for immediate regulation of fluxes toward the optimal growth configuration. Another flux-based analysis method, called MoMA [158], was therefore developed that is similar to FBA and based on the same stoichiometric constraints, but where the optimal growth flux for mutants is relaxed. Instead, MoMA provides an approximate solution for a sub-optimal growth flux state, which is nearest in flux distribution to the unperturbed state. It therefore involves a different optimisation problem than FBA, namely distance minimisation in flux space.

When all admissible vectors are of interest, one may wish to find a set of vectors that can generate all of them. This problem has been addressed several times and in different areas in the past [28]. Several similar concepts now co-exist. The most widely used in the context of metabolic networks are the concepts of elementary modes [156], extreme pathways [152] and minimal T-invariants [185]. The three enable to investigate the space of all physiological states that are meaningful [167].

Intuitively, an elementary mode is a special mode that has the property of not containing any other mode. More formally, an elementary mode is a flux vector v that satisfies conditions 1 and 2 above plus the following condition:

3. There is no non-trivial admissible flux vector v' such that: $R(v') \subset R(v)$

where $R(v) = \{j \mid v_j \neq 0\}$, *i.e.*, $R(v)$ is the set of reactions participating (with non-zero flux) in v; $R(v)$ is called the *support* of v.

Elementary modes have been said to represent a formalised definition of a biological pathway. Indeed, a biological interpretation can be given to such flux vectors: a mode is a set of enzymes that operate together at steady state [155] and a mode is elementary when the removal of one enzyme causes it to fail. Extreme pathways were introduced in the field by Schilling *et al.* [151] and are actually a subset of elementary modes. Both notions coincide in the case where all exchange reactions (reactions connecting a metabolite with the outside of a metabolic system) are irreversible. For a detailed comparison of the two, we refer the reader to [98]. As outlined in [157], the concept of minimal T-invariants used in Petri Nets is also closely related to elementary modes, as is the notion of extreme currents defined by B. Clarke [25]. Petri nets may be seen as directed bipartite graphs with two types of nodes called *places* and *transitions*. Places may contain tokens which are passed to other places through transitions according to some local rules. Unlike extreme pathways and elementary modes, minimal T-invariants and extreme currents have only been defined in the case of a network of irreversible reactions.

Clearly, there are links between the algorithms for enumerating elementary modes and the ones for obtaining minimal T-invariants. Interested readers may refer to [28] as concerns minimal T-invariants and to [59, 161, 180] for elementary modes. More generally, the usefulness of Petri-Net approaches to the study of metabolic pathways is presented in [185]. As discussed in [2], counting and enumerating such objects are hard algorithmic problems. In this case again, applications are numerous and range from model validation [73] to prediction of gene expression patterns [167].

An elementary mode may be seen as a set of reactions that, when used together, performs a given task. One may be interested instead in determining a set of reactions one needs to inhibit to prevent a given task, usually called target reaction, from being performed. This leads to the concept of a minimal (reaction) cut set that was recently introduced by Klamt and Gilles [96]. As mentioned in a latter paper [95], the task to be silenced can also be a combination of reactions. There is a strong link between reaction cut sets and elementary modes since the first have been operationally defined as corresponding to a set of reactions whose deletion from the network stops each elementary mode that contains the target reaction(s).

Readers interested in constraint-based modelling and wishing to have more details and examples on these concepts are invited to read [98] and the references therein indicated.

3 Topological analyses

3.1 Synthetic graph measures

Once a metabolic network is modelled as a graph, classical measures from graph theory may be used to check whether they provide any further insight into the structure and general characteristics of the network. Several such measures have been applied to metabolic as well as to other types of biological networks. In the "network analysis toolkit", one may thus find a number of measures such as: degree distribution, average inter-node distance (average shortest path length), closeness centrality, diameter, clustering and assortativity coefficients, node- and edge-betweenness centrality, and synthetic accessibility.

The *degree* of a node i, denoted by k_i, is the number of edges linking it to other nodes of the graph. The *distance* between two nodes i and j, is the length of a shortest path between these two nodes (the path may be not unique). Average inter-node distance of a given network, that is *average shortest path length*, is the mean of the distance for all pairs of nodes. *Closeness centrality* is the mean distance between a given node i and all other nodes reachable from it. The *diameter* of a network is the maximum distance between any pair of nodes in the network. The *clustering coefficient* of a node i is the proportion of existing edges over all those that are possible between the neighbours of this node. It quantifies how close from a clique is the subgraph induced by i and its adjacent nodes, where a clique is a maximal complete subgraph (each pair of nodes is connected by an edge). The node-betweenness centrality, or *betweenness-centrality* for short, of a node i, is the proportion, over all shortest paths between every pair of nodes in the network, of those that contain i. Edge-betweenness centrality is similarly defined by replacing node i by edge e. Both are a measure of the centrality of a node in a graph. The *assortativity coefficient*, denoted by ac, is defined by $ac = \frac{\sum_k (e_{kk} - a_k b_l)}{1 - \sum_k a_k b_l}$ where e_{kl} is the fraction of edges from a node of degree k to a node of degree l, $a_k = \sum_l e_{kl}$, and $b_l = \sum_k e_{kl}$. It tells in a concise fashion how nodes of different degrees are preferentially connected among themselves.

All the above measures may be applied to either a reaction or compound graph representation of a metabolic network. A further measure, called *synthetic accessibility* [198], is more specific of metabolites. The synthetic accessibility S_m of an output metabolite m is the minimal number of metabolic reactions needed to produce m from a set of compounds defined as the inputs of a network. Synthetic accessibility defined this way is a generalisation of graph diameter for directed branching chemical reactions in an input-output transport network [198]. It has been used to predict the viability of knockout strains with an accuracy that appears to be comparable to flux balance analysis on large, unbiased mutant data sets [198].

We now comment on two main works where some of the above synthetic structural measures, namely degree distribution and diameter, were applied to metabolic networks in an attempt to infer biological meaning. The corresponding papers have met with an open success in the bioinformatics community mainly because they propose to relate complex processes to simple measures. We try to discuss the limitations of these approaches. One general word of caution is called for already as concerns the use of synthetic measures: as stated in [68], such measures are truly informative only when the network lacks a modular structure, or when it is modular but all modules are homogeneous in terms of the mechanisms that originated them and their properties. If neither of these two conditions is fulfilled, then any theory proposed may need to take into account the modular structure of the network. Modularity is discussed in Section 3.2.

3.1.1 Small-world networks

The first work on the use of synthetic measures to analyse biological networks we comment on is based on the concept of small-worlds. Historically, the small-world phenomenon was first described by Stanley Milgram in the 1960s in a series of experiments where he chose individuals in cities across the US and asked them to send a letter to another unknown person on the opposite coast through a chain of people they knew on a first name basis [179]. The results of these experiments enabled him to estimate the average path length of the considered social network. Surprisingly, this value was very small (around 6) which later led to the famous expression “6 degrees of separation”.

The concept of small-world network was further formalised by Watts and Strogatz [195] who gave a construction method for such networks and provided evidence that many types of networks indeed fulfill the properties of a small-world, that is, their diameter increases logarithmically with the number of nodes. The construction proceeds as follows. The starting point is a regular network with n nodes and k edges per node. Each edge is then randomly rewired with probability p. The construction method allows to adjust between regularity ($p = 0$) and disorder ($p = 1$). The authors showed that it is the addition of edges between distant nodes (shortcuts) that causes a disruption in the diameter of the graph and therefore yields the small-world property (low diameter, high clustering coefficient). It is worth observing in passing that this construction explains how to obtain a network where there exists, as in Milgram's experiment, short chains of acquaintances linking together arbitrary pairs of strangers, but it does not explain why arbitrary pairs of strangers should be able to find short chains of acquaintances that link them together. A more operational definition that provides a better explanation for this was proposed by Kleinberg [99], who showed that, in addition to having short paths, a network should contain latent structural cues that can be used to guide a message towards a target. Not all small-world networks possess such cues and Kleinberg therefore captured another component of Milgram's experiment that was not evidenced by Watts and Strogatz's work.

In the case of metabolism, Fell and Wagner [50, 188] proposed soon after Watts and Strogatz' paper appeared that the metabolic network of the bacterium *Escherichia coli* satisfies the properties of a small-world network. The authors further argued that this type of architecture enables to minimise the transition times between metabolic states and thus contains clues as to the evolutionary history of metabolism. In 2004, Arita [9] suggested however that the graph model used by Fell and Wagner was not realistic enough. In Arita's model, the compounds themselves are represented as graphs where nodes are atoms and edges are chemical bonds. The paths are then calculated as atom trajectories between carbon atoms instead of complete metabolites (which enables to more accurately represent mass transfer). With this model, Arita showed that the network diameter is much larger than estimated previously, thereby questioning the biological interpretation of this network measure. A same argument was advanced by Alm and Arkin [4] who pointed to the fact that the concept of small-world networks tends to overlook the stoichiometry inherent to biochemical reactions. Indeed, the edges that enable to reduce the path length in metabolic networks can often be explained by co-factors that connect seemingly unrelated reactions. Yet another way of dealing with paths in metabolic networks is to assign weights to the nodes according to their degree (see Section 3.1) and to look for lightest paths [33]. This heuristic was shown to bring better results (at least in terms of prediction of known metabolic pathways) compared to the situation where frequent compounds are removed.

Clearly however, the notion of path in not very well suited to metabolic networks and one should prefer the notion of balanced hyperpath as defined in constraint-based modelling (Section 2.3.2).

3.1.2 Scale-free networks

Another notion has become even more popular in the biological network literature than the small-world property. This is the notion of scale-freeness. The term was introduced in 1999 by Barabasi and Albert [12] to qualify a large diversity of networks whose degree distribution is believed to deviate from the classical Poisson distribution that is expected in random graphs, and is instead better approximated by a power-law distribution. The types of networks studied by Barabasi and Albert in their 1999 Science paper included genetic, nervous and social networks as well as the web. It was further proposed that, because of this property, such networks would be error tolerant and robust to random attacks [142]. In the following years, other biological networks, including metabolic [80], were suggested by Barabasi and colleagues to belong to this very general class of networks. This same idea was then taken up and the property repeatedly exhibited by a number of other authors. The reader may consult Khanin and Wit for further references [92].

In order to explain more formally what a "scale-free" network is, let us first consider the function $p(k)$, which gives the probability for a randomly chosen node to have k edges connected to it. A network is said to be "scale-free" if for all k_1, k_2, the ratio $p(k_1)/p(k_2)$ is invariant by multiplication of k_1 and k_2:

$$\frac{p(k_1)}{p(k_2)} = \frac{p(\alpha k_1)}{p(\alpha k_2)} = F(\frac{k_1}{k_2})$$

where α is a positive constant and F is the rescaling function. One can demonstrate that this property is true if and only if the probability function $p(k)$ follows a power law, *i.e.* $p(k) \propto k^{-\gamma}$ where γ is the power-law exponent [139].

One of the characteristics of scale-free networks (which differentiates them from Erdös-Rényi graphs for which the degrees of the nodes follow a Poisson distribution) is that a few nodes have many connections while a large number of nodes have very few connections. Highly connected nodes are sometimes called "hubs" and are thought to play a particular role in the network. For instance, in [81], the authors argued that, in the context of protein interaction networks, hubs correspond to essential proteins in the sense that mutants lacking this protein have lethal phenotype. Scale-free networks satisfy also the small-world property.

Recently, several works started to question the results of Barabasi and colleagues. This questioning may be classified into four main categories.

The first concerns the quality of the data modelled into the networks. It has been argued that the high rate of errors in the data used to build the networks create artefactual links that invalidate or at least weaken some of the results obtained. This concerns either the scale-free property or the existence of hubs and their enrichment with essential nodes [5, 30]. This observation has been so far applied more to protein-protein interaction networks (PPIs) than to metabolic networks. Incompleteness of data is however a problem that concerns metabolic as much as other types of biological networks. Stumpf *et al.* [168] thus noticed that the networks considered in the literature, including for modelling metabolism, usually correspond to partial data reflecting our incomplete knowledge of the studied processes. They therefore asked the legitimate question: if the networks we observe are scale-free, does it imply that this is also true for the larger networks from which they are extracted? Their answer to this question is no.

The second category of questioning is methodological. Two main works [92, 168] have thus contested the proposition that the biological networks available and commonly used in the literature are scale-free. Indeed, the test that is classically applied to show that the degrees of a graph are drawn from a power-law distribution consists in fitting a straight line on a log-log plot. However, as stated in [92], "fitting a straight line doesn't necessarily make the points follow it". More formal methods are required to ascertain that the scale-free distribution is indeed the best model for the data observed. Khanin and Wit [92] showed that several networks considered as scale-free using the elementary fit-to-a-line test were no longer considered as such when applying a maximum-likelihood method and chi-squared goodness of fit tests. They suggested other distributions have similar qualitative properties as observed in biological networks, namely a few nodes with a high number of connections and many with few connections. These distributions include truncated power-law (which however they indicated exhibits also a poor fit to the data), generalised Pareto law, stretched exponential distribution, geometric random graph, geometric distribution, or combinations of the above [92]. By applying statistical model selection methods using maximum likelihood inference, composite likelihood methods, the Akaike information criterion and goodness-of-fit tests, Stumpf and colleagues [168] obtained that the degree distribution of present-day metabolic networks appear to be better described by a log-normal model (if Y is a random variable with a normal distribution, then $X = exp(Y)$ has a log-normal distribution). Like the scale-free distribution, the log-normal as well as others mentioned by Khanin and Wit or tested by Stumpf *et al.*, are fat-tailed, but they are not scale-free.

The third category of questioning of the scale-freeness, or of other general properties of metabolic networks are more deeply rooted in biology. Indeed, as observed by Alm and Arkin [4], one important characteristic of metabolism is forgotten by all purely topological studies such as the node degree distribution: this is that each node has a specific identity, which furthermore is usually distinct from the identity of other nodes. The node degree distribution captures therefore only a very small part of the real characteristics of biological networks. As advanced by Alm and Arkin, “whereas the Internet might function similarly if individual nodes were rewired while keeping the same overall topology, metabolic reactions are highly specific and edges cannot, in general, be swapped because of additional constraints such as conservation of mass”. They further mention a comment by J. Doyle that biological networks might perhaps more accurately be considered as “scale rich” because they are composed of many nodes of different types organised into modular and hierarchical structures.

Finally, at a more conceptual level, it can be argued, as Keller did [88], that even if a network is indeed proven to be scale-free, knowing this would not be very informative since this class is possibly too general. Therefore, it seems that there is a need for finer measures to reach a more relevant biological interpretation.

The works on small-world and scale-free networks illustrate the initial enthusiasm in the field for general results on the structure of biological networks. Both have now been shown to have limited impact on our understanding of metabolism. It seems that either the measures or the interpretation we make of these measures have to be changed or further elaborated to be useful when trying to make sense out of the structure of metabolism. The comments of Alm and Arkin [4] point out also to crucial problems for testing any of the hypotheses that have been advanced concerning the global structural properties considered in this section, but also any of the issues that we shall discuss next: what is a good null graph-model, and therefore, what are the properties that we expect by chance? Despite an abundant literature, this is a question that seems to remain widely open.

3.2 Modularity

Wagner *et al.* informally define modularity as “an abstract concept that seeks to capture the various levels and types of heterogeneity found in organisms” [190]. The authors argue that different kinds of modules may be distinguished. All have however in common the fact that they are integrated with respect to a certain kind of process (such as, for instance, natural variation, function, development and so on) and should be relatively autonomous from other processes or parts of the organisms to which they belong. All ideas of modularity appear therefore to refer to a pattern of connectedness in which elements are grouped into highly connected subsets, the modules, which are more loosely connected to other such groups. The notion of connection should not be understood here in the strict sense of physical interaction but rather as, elements are considered as connected when they belong to a common process. The independence of modules may be spatial or temporal, chemical or genetic, structural or dynamic [197].

Modularity has been perceived both as a fundamental notion (deciding if a network is modular is informative *per se*) and as an operational notion (decoupling a system into independent modules may be computationally and formally very efficient). No really clear formal definition of this concept has been reached.

Most biologists do however agree on the fact that modularity is present in biology. This is obvious already at a high level [170]. Organ transplants are an example of this. In the same way as organs can be considered as the modules of an organism, an organism may also be seen as the unit of a population. At a lower scale, organs may be decomposed into cells formed of molecules which in turn may be further decomposed into modules such as protein domains. At the intermediary level between cells and molecules, the picture becomes however less clear. Nevertheless, several authors have argued in favour of modularity at all levels, including the cellular level [197]. Several examples of cellular modules can be given [71] – DNA replication, glycolysis, protein synthesis – some of which have been successfully reconstructed *in vitro*, which in itself represents an excellent validating criterion for modularity. From a different perspective, the very fact that phenotypic evolution can be studied on a character by character basis lead Wagner to plead in favour of the existence of genetically independent modules [189], for which the term evolutionary module was later coined. How modules originated however, for instance did they arise through the action of natural selection or because of biased mutational mechanisms, remain largely open questions.

In the context of metabolism, several methods have been applied to identify modules, using different operational definitions.

3.2.1 Top-down identification

Pathways A natural way to define modules in the context of metabolic networks is to build on pre-existing concepts like the one of metabolic pathways. Indeed, glycolysis has been shown to be reconstituted *in vitro*, thereby outlining its independence with respect to the rest of metabolism. One limitation of this approach is that the concept of metabolic pathway has itself never been defined formally. Instead, the partition of a network into pathways is partly due to historical reasons, related to the way metabolism was discovered, and therefore reflects one subjective view of metabolism.

Attempts at formalising the notion of a metabolic pathway have however been made. Yamada *et al.* [199] for instance define a "pathway module" to be comprised of enzymes that have the same phylogenetic profiles and are also close to each other in the metabolic network. The phylogenetic profile of an enzymatic gene is represented by a string that encodes its presence or absence in every available fully sequenced genome.

The concept of elementary mode has also been proposed as a formalised definition of a metabolic pathway [155] and is therefore another good candidate for defining the concept of module. Indeed, an elementary mode can be interpreted as a minimal set of enzymes that operate together at steady state. Unfortunately, the exploding number of elementary modes (more than half a million for a network of around one hundred reactions [97]) strongly limits their usage as an operational definition. This framework may still be useful if employed, for instance, together with the notion of coupled reactions [19], also named co-sets [128], *i.e.* reactions which participate in the same elementary modes. More formally, coupled reactions are defined as reaction pairs such that the fluxes passing through them, respectively v_1 and v_2, share any of the following types of coupling [19]:

1. Directional coupling ($v_1 \rightarrow v_2$), if a non-zero flux for v_1 implies a non-zero flux for v_2 but not necessarily the reverse.
2. Partial coupling ($v_1 \leftrightarrow v_2$), if a non-zero flux for v_1 implies a non-zero, though variable, flux for v_2 and vice versa.
3. Full coupling ($v_1 \Leftrightarrow v_2$), if a non-zero flux for v_1 implies not only a non-zero but also a fixed flux for v_2 and vice versa.

Connectivity-based definitions Besides the notion of a pathway, various other methods were developed to identify structures in networks that would be densely connected within the module and loosely connected to the rest of the network. Many are inspired by long standing mathematical work on modular graphs and graph decomposition. Great care must be taken however as the terms module and decomposition in particular are often employed in a different and sometimes looser sense in the bioinformatics community. For ease of reference, we chose to follow the terms as used in the bioinformatics community but do call attention to the potential confusion this may create. It is an interesting issue whether more attention should not be paid to even seemingly unrelated mathematical theory on modularity and decomposition when trying to define and identify modules in biological networks.

In the bioinformatics effort to formalise the notion of modules in graphs, one may distinguish between two main situations: all nodes are classified in at least one module; some nodes may remain unclassified. The first case leads to the so-called network-decomposition approaches, the second to methods for module detection.

Concerning module detection, Spirin *et al* [165] proposed a method to detect modules in protein interaction networks (nodes represent proteins and edges represent interactions between proteins) that could be applied also to a metabolic reaction or compound graph. A module is defined by the authors as a dense subgraph. The density of a subgraph is given by the function $Q(m, n) = 2m/(n(n-1))$, where m is the number of interactions between the n nodes of the subgraph. A statistical criterion then enables to decide if the value taken by Q is exceptional. The null model considered is a random graph model where the degree sequence is the same as in the studied graph. One may observe that in order to be able to deal with large subgraphs, the authors use heuristics (local search, simulated annealing) for the counting as well as approximations and simulations for the statistical part. Indeed, the problem of searching for the heaviest induced subgraph has been shown to be NP-hard [145, 100]. A problem is said to be NP (for Nondeterministic Polynomial time) if given a solution, this can be checked for correctedness in polynomial time. Informally speaking, a problem is said to be NP-hard if an algorithm for solving it can be translated (in polynomial time) into one for solving any other NP-problem. In other words, a problem is NP-hard if it is as at least as hard as any NP-problem. A problem which is both NP and NP-hard is called an NP-complete problem.

Most network decomposition methods yield non-overlapping modules, *i.e.* modules constitute a partition of the network. Elementary modes are one notable exception to this. Graph partitioning is a method widely used, for instance in parallel computing (for a review, see [54]). In its simplest formulation, graph partitioning consists in separating the nodes of a graph into p disjoint subsets of similar size, while minimising the number of edges between subsets. This problem is NP-complete even if $p = 2$ [61]. A limitation for the applicability of graph partitioning to other fields is that the number of modules has to be known in advance which is not the case when considering metabolism.

Other methods initially developed for sociology are commonly used when the number of modules is not known [194]. The problem still consists in separating the nodes of a graph into disjoint subsets but the criterion to minimise is not anymore the number of inter-group edges. Indeed, in this case the optimal solution would be a unique module containing the whole network. Instead, the criterion to optimise, termed modularity function, is the sum for all modules of the difference between the number of edges intra-module and the number of intra-module edges expected under a null model [120]. The random graph model generally used is again one that preserves the degree sequence of the nodes. In practice, the number of observed edges is given by the adjacency matrix of the graph, and the number of expected edges is given by $\frac{k_i \times k_j}{2m}$, where k_i (resp. k_j) is the degree of node i (resp. j) and m is the number of edges in the graph. Finding the partition which optimises this criterion is a difficult problem. Several heuristics have been proposed. Guimera *et al.* [67] use a simulated annealing approach which they then apply to the metabolic network of *Escherichia coli.* The modules identified (the method finds 19) are then compared to the metabolic pathways as defined in KEGG. In some cases, modules can be given a general function. More recently, the same type of approach was used by Parter *et al.* [130] to analyse the relationship between modularity in the metabolic network of bacteria and variability in their environment. On the methodological side, Newman [121] introduced a matricial formulation of the graph partition problem which enables to use more efficient techniques from spectral algorithmics while Daudin *et al.* cluster nodes using a method based on mixture degree distributions to obtain modules differently defined from those of Newman and Guimera [36].

Two modular decomposition methods based on information provided by the stoichiometric matrix have been more recently proposed [136, 204]. In [204], Yoon *et al.* attempted to identify related modules by applying an algorithm for top-down partitioning of directed graphs with non-uniform edge weights [204]. The weights are determined by the metabolic flux distribution and require to perform FBA on the network. In [136], the idea was to establish a distance between any two pairs of reactions i, j. The distance chosen is a Pearson's correlation coefficient between the fluxes carried by i, j for all possible steady states of the system. Obtaining such correlation coefficients is done by computing the null-space of the stoichiometric matrix of the metabolic network. Using such distances, a hierarchy is then constructed where the metabolic system is represented by a tree whose root node corresponds to the whole system, leaf nodes to the reactions and intermediate nodes to unique subsystems of reactions. Cutting the tree at any given level produces modules of a certain granularity.

Finally, more ad-hoc methods with no clearly specified methodology have suggested that metabolic networks are organised in a hierarchical modular [144] or bow-tie nested [205] way. The perception in this case is that metabolic networks present a highly modular core-periphery structure, in which the core modules are tightly linked together and perform basic metabolic functions, whereas the periphery modules interact with only a few other modules and accomplish relatively independent and specialised functions. This reflects the "giant strong component" idea earlier described by Ma and Zeng [110, 111]

It seems clear that further work in the field of module identification should include a more thorough discussion of the modularity function and especially of the null model considered, suggestions for improving the optimisation procedure to avoid local optima, and propositions of alternate ways for module validation. More generally, all methods presented here have been defined for simple graphs, new methods which would apply to bipartite graphs or hypergraphs probably need to be developed.

Motifs Close to the concept of module is the one of motif. Both are thought to be building blocks of a network. Mostly for computational reasons, only small motifs have been studied but the difference between the concepts of module and motif is not so much a question of size as of definition. A motif is thus generally not required to be autonomous but rather repeated (reused).

Motifs have first been introduced in the context of gene regulatory networks where they have been defined as patterns of interconnections, more formally isomorphic subgraphs that appear unexpectedly often in a network [162]. Once again, the expectation has to be defined and good random graph models need to be discussed. Motifs were later proposed for metabolism with the same [45] or with a different definition. In the latter case, the notion of a coloured motif was introduced [104]. A coloured motif is defined as a collection of node labels that induces a connected subgraph. The focus is on node labels (reaction mechanisms in this case) and not on the topology of the subgraph although connectedness is required.

3.2.2 Bottom-up identification

So far, we mainly covered so-called top-down approaches to discover structural modules. One limit of such approaches is that they propose structures of interest in a graph but the modules still need to be biologically validated. In contrast, bottom-up approaches such as described in [184] start from a small number of entities and add elements one by one, experimentally verifying autonomy of the module at every step. Such a process clearly cannot be applied in general since it implies a thorough preliminary knowledge of the studied system in order to minimise the number of experiments. Instead, methods that try to validate structural modules evaluate whether the module is either functionally coherent (functional module) or evolutionarily conserved (evolutionary module).

Snel and Huynen [163] addressed the question of the overlap between functional modules and evolutionary modules in the case of metabolic networks (where functional modules are identified as metabolic pathways) and of protein interaction networks (where functional modules are considered as protein complexes). The issue is complex. Should for instance co-regulation be considered besides membership into a same pathway or protein complex, how is the fact that functional differences may exist within a same orthologous group taken into account, and how is evolutionary modularity quantified? For example, how modular is the evolution of a functional module when the "same" module in another species is partly composed of different proteins and/or of fewer proteins? As the authors indicated, with this as with many other issues, trends only can be identified, that are in this case scored based on the observed level of evolutionary modularity relative to the level expected for a random set of proteins. The general conclusion at which the authors arrive is that there are substantial differences in the evolutionary modularity between individual functional modules (for instance, biosynthetic pathways are more conserved as units than catabolic pathways). The same general conclusion was reached by Spirin *et al.* [164] who showed that modules of high genomic association and metabolic proximity do not necessarily match traditional metabolic pathways. Such modules, rather than the traditional pathways, should therefore be thought of as evolutionary and regulatory units. They also observed that genomic associations favour linear metabolic pathways and break apart at branching points. The authors then suggested that linear pathways are regulated and inherited as a single "building block" of the metabolic network. Although enzymatic subunits are strongly associated, indicating persistent co-regulation and co-evolution, they also proposed, based on the results obtained, that regulation and evolution of isoenzymes depend on their role in providing alternative specificity or differential expression.

Structural modules (defined using topology), functional modules (defined using annotations) and evolutionary modules (defined using orthology) therefore seem to be three different concepts reflecting three different points of view on metabolism.

4 Evolution of metabolism

It is questionable whether the evolution of metabolism should be treated separately from the rest since structural studies (connectivity, motifs, modules) already must take into account evolution to be sound. We can however argue in favour of this choice by observing that the work described next is exclusively dedicated to answering to the question: how did metabolism evolve?

More specifically, we attempt to provide an overview of the different theories for the evolution of metabolic networks. A discussion on pre-enzymatic evolution is beyond the scope of this paper. The interested reader may consult (this is a non-exhaustive list) Wächtershäuser [186, 187], Maden [112], Lazcano and Miller [105], Morowitz *et al.* [117], and Caetano-Anollés *et al.* [20]. It is also beyond the scope of this paper to discuss a mirror question to the evolution of networks which is to understand how a metabolic network may shape the evolution of its enzymes. The interested reader may start with the paper of Vitkup *et al.* [183] for a nice introduction to the topic.

After reviewing different models that have been suggested for the evolution of metabolism, we comment on several works where the comparison of metabolic pathways or networks was used to reconstruct evolutionary scenarii across different species.

4.1 Models for the evolution of metabolic networks

4.1.1 Biological models

Various biological models were proposed to explain the evolution of metabolic networks. Schmidt *et al.* [153] reviewed five and provided examples for each one of them. The five models include *de novo* invention, retro-evolution, specialisation of multifunctional enzymes, pathway duplication and patchwork recruitment. Clearly all these models are not exclusive and each one may help to explain different parts of the evolution of metabolism. Moreover, one may observe that at least two different questions are treated by the models recalled in [153]. These are: 1. in which order are enzymes recruited, and 2. where do they come from? We propose to try and decouple the two questions although this may seem an artificial exercise since the order could depend on what is available in earlier metabolism, and therefore on the "wherefrom". Proceeding in this way may however help in getting at a better understanding of the issues at play.

Essentially two models give an answer to the first question, in which order are enzymes recruited? The first was proposed in 1945 by Horowitz [76]. According to this model, by so-called retrograde evolution, selective pressure on a metabolic pathway mainly targets the successful production of its end-product. The formation of the required end-product from an intermediate metabolite therefore increases the fitness of the organism. The same reasoning can be applied recursively, selective pressure then acting on the production of the intermediate metabolite. The formation of this metabolite from another metabolite again gives a selective advantage and the pathway thus evolves backwards (see Figure 5).

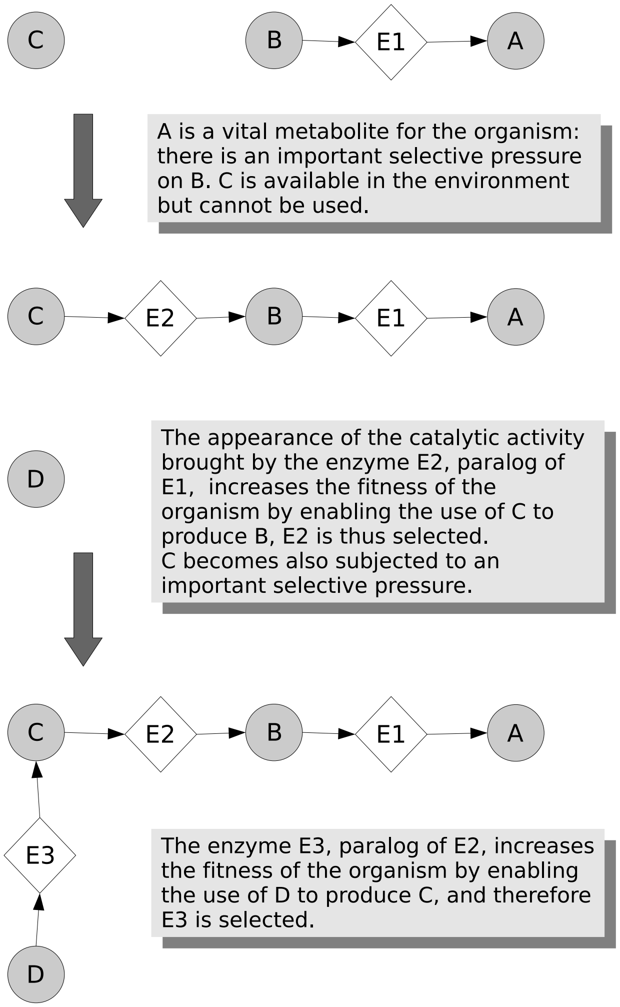
Some thirty years later, a different model was advanced by Jensen [79] following earlier work [192, 203]. This model is based on the central concept of substrate ambiguity observed for many enzymes. In this model, that came to be known as the "patchwork-evolution" or "recruitment" model, initial metabolism is carried out by a few substrate-ambiguous unregulated enzymes. In this context, "any fortuitously formed compound that happened to be useful would have conferred a selective advantage, thereby providing a basis for increased and more specific production of that compound (by gene duplication and specialization via mutation)" [79] (see Figure 6). Jensen further proposed that enzyme recruitment may be done "en bloc", several enzymes being recruited in a single step (see Figure 7).

Possible criticisms to either model bring to the fore the second type of question, where do the enzymes come from? As suggested by Jensen, gene duplication followed by specialisation via mutation appears to be one of the main sources of metabolic versatility whatever model is considered.

As pointed out by Díaz-Mejía *et al.* in [42], the two models, retrograde and patchwork, present two main differences in relation to gene duplication. In the retrograde model, gene duplication provides an enzyme that can supply an exhausted substrate. This would often give rise to homologous enzymes that catalyse consecutive reactions.

In the patchwork model, duplication of genes encoding for enzymes capable of catalysing one or more reactions allows each descendant enzyme to specialise in relation to one of the ancestors. Enzymes generated by patchwork evolution may be expected to catalyse reactions a greater distance apart in the pathway than those originated by retrograde evolution.

This is the first main difference. The second is that the retrograde model invokes consecutive reactions and can therefore originate enzymes catalysing chemically dissimilar reactions that however preserve substrate specificity while in the patchwork model promiscuous enzymes tend to catalyse chemically similar reactions even while acting on different types of substrates. By studying the reconstructed metabolic networks of the bacterium *Escherichia coli* and of a number of other organisms, Díaz-Mejía *et al.* evaluated the influence of both chemical similarity and distance between



A is a vital metabolite for the organism: there is an important selective pressure on B. C is available in the environment but cannot be used.

The appearance of the catalytic activity brought by the enzyme E2, paralog of E1, increases the fitness of the organism by enabling the use of C to produce B, E2 is thus selected.
C becomes also subjected to an important selective pressure.

The enzyme E3, paralog of E2, increases the fitness of the organism by enabling the use of D to produce C, and therefore E3 is selected.

Figure 5: Model of retrograde evolution.

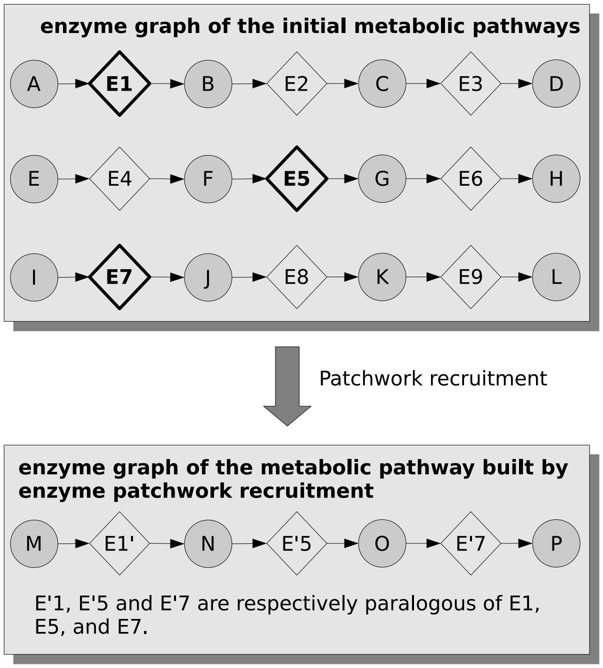

enzyme graph of the initial metabolic pathways
A
E1
B
E2
C
E3
D
E
E4
F
E5
G
E6
H
I
E7
J
E8
K
E9
L
Patchwork recruitment
enzyme graph of the metabolic pathway built by enzyme patchwork recruitment
M
E1'
N
E'5
O
E'7
P
E'1, E'5 and E'7 are respectively paralogous of E1, E5, and E7.

Figure 6: Model of patchwork evolution.

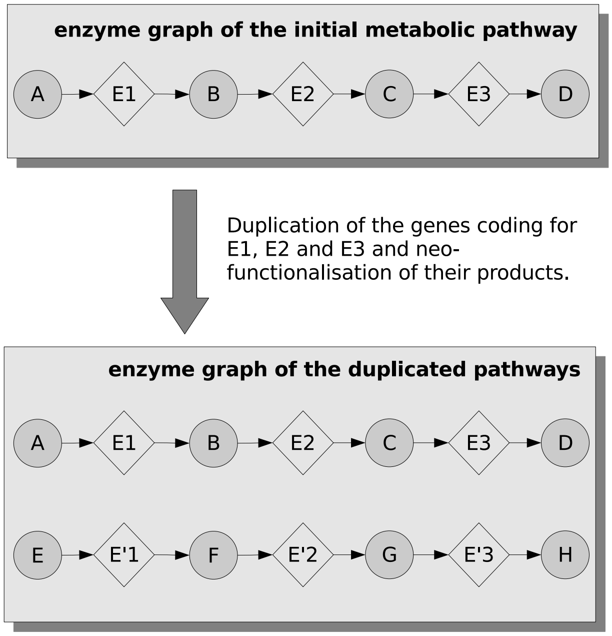
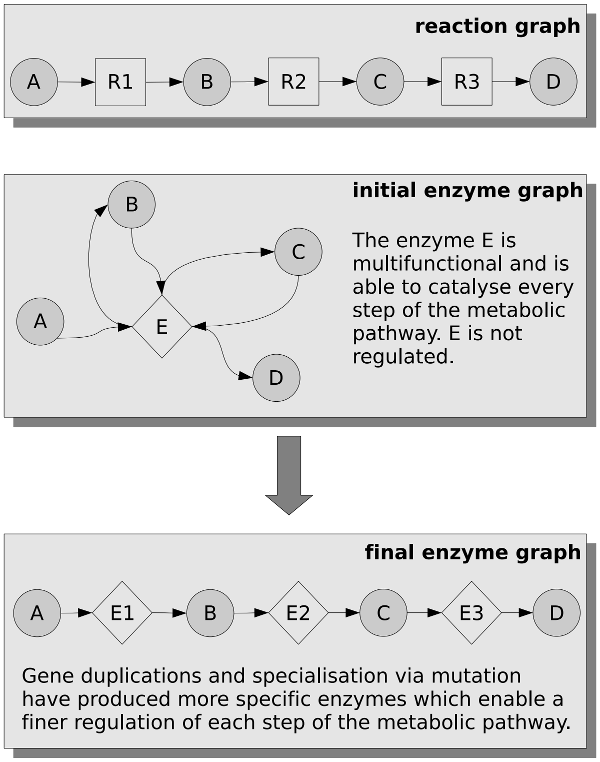

enzyme graph of the initial metabolic pathway
A
E1
B
E2
C
E3
D
Duplication of the genes coding for E1, E2 and E3 and neo-functionalisation of their products.
enzyme graph of the duplicated pathways
A
E1
B
E2
C
E3
D
E
E'1
F
E'2
G
E'3
H

Figure 7: Duplication of genes and neo-functionalisation of enzymes.



The enzyme E is multifunctional and is able to catalyse every step of the metabolic pathway. E is not regulated.

Gene duplications and specialisation via mutation have produced more specific enzymes which enable a finer regulation of each step of the metabolic pathway.

Figure 8: Duplication of genes and increase of the specificity of the enzymes.

reactions (computed as the number of reactions that separate them) on the rate of retention of duplicates. They uncovered an increased retention of duplicates for enzymes catalysing consecutive reactions. This was observed also in enzyme-enzyme interaction networks, but not in interaction networks of non-enzymatic proteins or in gene transcriptional regulatory networks. It suggests that retention of duplicates results from the biochemical rules governing substrate-enzyme-product relationships. The authors also confirmed a high retention of duplicates between chemically similar reactions. The retention of duplicates between chemically dissimilar reactions is, however, also greater than expected by chance as evaluated against null models. Altogether, the results seem to indicate the influence on the retention of duplicates of both distance apart in the network (as expected under the retrograde model) and chemical similarity of reactions (as expected under the patchwork model). This suggests that the retrograde and patchwork models may not be independent of each other. Rison *et al.* and Teichmann *et al.* [147, 174] had proposed the same in conclusion of their papers but their study gave much stronger weight to the influence of patchwork recruitment as compared to local recruitment in the evolution of pathways.

In [148], Rison and Thornton pointed also to another problem with the retrograde model of Horowitz which is that it fails to explain how the model would work in an environment that is poor in intermediate metabolites. Indeed, one strong assumption of this model is that each intermediate of the backwardly evolving pathway was readily available in the primitive environment [79]. The model also does not explain the development of pathways that include unstable metabolites which could not accumulate in the environment long enough for retrograde recruitment to take place. Roy [150] had however previously proposed a possible answer to this second problem which would call for evolution to have taken place by jumps, *i.e.*, by the local recruitment of a multi-functional enzyme capable of catalysing several steps at a time, albeit inefficiently. Roy further speculated that in some cases, one primordial multi-enzyme might have catalysed the whole sequence of reactions of a biosynthetic pathway (see Figure 8). The pathway would then have evolved by a single leap. In their study, Díaz-Mejía *et al.* [42] also pointed to a significant retention of duplicates as groups instead of single pairs. In parallel to this, Mahadevan and Lovley [113] showed that the role of gene duplications to boost enzymatic flux rather than provide metabolic resilience, as was advanced by Papp *et al.* in [129] for prokaryotes, may not be universal. Indeed, in eukaryotes, redundancy in metabolic networks appear to provide significantly more genetic buffering than do even gene families [94].

In parallel to the controversy between a retrograde and/or patchwork model of metabolic pathway evolution, other studies have argued that new enzymes plug preferentially at the periphery of bacterial networks [141].

With large-scale datasets now at hand, all these models may be tested or refined and new models can be proposed. As indicated in [42], massive gene duplications, in particular of whole genomes, as well as consideration of other potential sources of metabolic versatility such as horizontal gene transfer [109, 171], gain and loss of enzymatic genes [171, 65, 172] and gene fusions [46] could enhance our understanding of the evolution of metabolism.

4.1.2 Computational models

In contrast with biological models, computational ones do not necessarily have biologically realistic mechanistic foundations. Those presented in this section operate by the application of a set of evolutionary rules and in general are said to be valid if the created graph presents characteristics that match well with what is observed in real biological networks. In some cases, it is not direct fit that is checked but instead fit to some properties the data is believed to satisfy.

In the context of their work on network structure, Barabasi and Albert [12] thus proposed a model of network evolution that could explain the estimated power-law distribution of observed node degrees. The two rules implemented in the model were: 1. the network grows continuously by addition of new nodes (network growth), and 2. new nodes are linked preferentially to existing nodes that are well connected (preferential attachment). The latter rule is put into effect by assuming that the probability that a new node will be connected to a node i depends on the degree k_i of that node. This model was indeed later formally proven to yield networks with power-law degree distributions in [38] although it is not explained how to take probabilities proportional to the degrees when these are all zero [15]. Following Khanin and Wit [92] and Keller [88], one may however question the validity of a mathematical model for evolution that adjusts the latter's underlying principles and assumptions so that the model may faithfully reproduce a topological property of the network, namely here scale-freeness, that has not been conclusively proven.

Pfeiffer *et al.* [134] followed a logically sounder approach than Barabasi and Albert [12] which consisted in studying whether hubs emerge when simulating for the evolution of a network according to a scenario proposed by Kacser and Beeby [82], that itself is closely related to the model of patchwork evolution advanced by Jensen [79]. Kacser and Beeby's scenario is based on the assumption that, because of the low coding capacity of early genomes, and because it is unlikely that a large number of specialised enzymes emerged *de novo*, it is plausible to assume that ancestral cells had only a few enzymes with broad specificities. This allowed for the catalysis of all essential reactions at the cost of a low turnover for any single biochemical reaction. The ancestral cells are then assumed to have been selected for growth rate and to have evolved by mutations affecting the kinetic properties of the enzymes and occasional gene duplications. The simulations indicated that duplications and specialisations lead to the loss of biochemical reactions and of intermediary metabolites, and that complex features of metabolic networks such as the presence of hubs may indeed result.

Two other pieces of work proposed to algorithmically explore the evolution of a metabolic network by applying two different types of operations. The first starts from a given compound or sets of compounds and determines what is the scope of this compound [70]. The scope of a compound corresponds to the set of compounds that can be generated starting from it using the reactions that are available in an organism. This type of approach enables to address questions such as, which compounds are essential, and may help to infer an order of appearance of compounds during evolution. The second work addressed the question of the evolution of metabolism by iteratively eliminating “unnecessary reactions” from a network [140]. The eliminating process is the following: starting from a metabolic network, a reaction is picked at random and is withdrawn from the network if its removal does not affect significantly the production of biomass (which is evaluated using a flux balance analysis framework). This process is repeated until no such reaction can be found anymore. The authors then related this process to the adaptation of bacterial metabolism to new environmental niches. A major problem with the approach is that during the whole process, the growing media and the biomass reaction stay constant, which lacks realism when modelling adaptation. Moreover, one should notice that this process is greedy and outputs a minimal network (many other minimal networks may exist) but not the minimum network.

4.2 Using comparative analysis to infer evolutionary scenarii

Various authors have started to reconstruct evolutionary scenarii for metabolic pathways and networks across different species. A parallel can be made here with sequence comparison. In order to study the evolution of sequences, sequence similarity can be used to deduce common ancestry (*i.e.* homology). Phylogenetic reconstruction methods may then allow to infer an evolutionary history between homologous sequences. This history is generally represented as a tree. The same ideas have been applied to metabolism.

We sketch the few methods currently available for inferring evolutionary scenarii after a discussion of algorithms for computing a distance between pathways or networks by aligning them since this is at the base of some of those methods.

4.2.1 Pathway and network alignment

From a methodological point of view, several works have been concerned with the alignment of metabolic pathways, and more recently of whole networks. An alignment of metabolic pathways differs from an alignment of sequences in two main aspects. First, the units to align are not nucleotides or amino-acids but reactions, and second, the structure is not linear anymore but is represented instead as a graph or hypergraph.

In order to align reactions, it is necessary to define a distance measure between them. When aligning nucleotides, such distance is generally chosen to be proportional to the probability of mutation from one nucleotide to another (substitution cost). Indeed, in this context, a substitution corresponds to a precise biological mechanism: a mutation fixed by natural selection. When comparing reactions, no such parallel can be made between substitution and mutation. Instead, distances between reactions are usually functional and do not necessarily reflect an evolutionary relationship. This remark has no consequence if one is interested in the structural comparison of metabolic pathways but may be crucial if one is concerned with establishing their evolutionary history. Tohsato *et al.* [178] introduced a simple method to compare reactions based on the EC numbers (see Section 2.1). Since the Enzymatic Classification forms a hierarchical structure which can be represented as a tree, the similarity score between two EC numbers can be defined as a function of the distance between them in the tree.

The authors use this distance to compute alignments between linear metabolic pathways and outline the structural similarities in the different amino acid biosynthesis pathways within the same species and between different species [178]. A generalisation of the work of [178] was proposed by Pinter *et al.* [135] in the case of branched metabolic pathways. The underlying computational problem is linked to labelled subgraph isomorphism which can be solved in polynomial time. This work does not allow to deal with general graphs (the authors heuristically break cycles) which somehow restricts its applicability. The main motivation for restricting to trees is that subgraph isomorphism is NP-hard. Nevertheless, efficient algorithms can still be designed even in this case, provided that they take into account some specificities of the instances, such as their so-called local diversity which reflects the fact that neighbouring nodes usually have distinct labels [196]. In the context of protein interaction networks, Kelley *et al.* [89] proposed to align graphs using their decomposition into paths, a technique that could also be used for metabolic networks. However, the results of path alignments still have to be gathered *a posteriori* which constitutes the bottleneck of the algorithm. Other rougher distance measures between two metabolic networks have been adopted, for instance by Tun *et al.* who opted for the Hamming distance between the adjacency matrices for the compound graphs representing the networks. Passing from pathway to a whole network alignment represents a change of scale that clearly requires further methodological developments.

The choice can also be made to ignore altogether topology and to identify the metabolic networks to be compared with bags of enzymes and metabolites [27], or with various structural indices such as the clustering coefficient, betweeness centrality, average path length, diameter, concentration of subgraphs [206, 193, 131].

These can provide only very rough comparative measures and one should remember that graphs themselves are already poor models for biomolecular reactions. Effort should therefore be placed instead into reaching more realism, for instance by dealing with bipartite graphs or hypergraphs. The question of how to align metabolites (or simply ignore metabolites in the alignment) should then be addressed. More complex distance measures between reactions may also have to be considered that take into account either the sequence or structure of the enzyme which enables the reaction, or identify each reaction with the chemical transformation it represents. In the latter case, a distance between two reactions could be modelled as a distance between two transforms. There are two issues involved with such an approach. One is concerned with finding an optimal atom mapping between the substrate(s) and the product(s) of an enzymatic reaction that at the same time is reliable and can be efficiently computed. The second issue requires defining a similarity measure between mappings that would accurately reflect the chemical and possibly evolutionary similarity between chemical transformations. Some work has been done in this direction but it is partial [84, 8, 101, 3, 49, 48, 72]. Clearly, this remains an open question, whose discussion may be informatively fueled by reading Dandekar *et al.*'s paper [35] who compare glycolysis among different organisms by using three methods: biochemical, by means of elementary modes and through a comparative analysis.

4.2.2 Inferring evolutionary scenarii

The first method for inferring evolutionary scenarii between pathways we present was proposed by Cunchillos *et al.* [34]. Its motivation was to predict the order of appearance of metabolic pathways in the course of evolution. A metabolic pathway was simply modelled as a sequence of presence/absence of enzymes, and its structure was not taken into account. Reconstruction was performed by a method of maximum of parsimony. Interestingly, this work proposed a coherent model of enzyme evolution in the sense that it allowed to discriminate between changes in substrate specificity, cofactor binding, etc. Liao *et al.* [108] worked at the larger level of full networks but with a coarser measure based on distances between profiles that recorded this time the presence and absence of whole metabolic pathways. Heymans *et al.* [74] worked as Cunchillos *el al.* at the level of pathways but modelled each one as a set of enzymes. A phylogenetic tree was built using a distance-based method where the distance reflects the similarity of sequences between enzymes as well as the sequence similarity of the neighbouring enzymes in the pathway. In this way, the structure of the pathway was indirectly taken into account. Forst and Schulten [56] proposed first to detect homologs between the enzymes of different pathways using sequence alignment, and then to calculate a distance between the pathways introducing gap penalties when no homolog is found. A distance between pathways is therefore a combination of sequence alignment scores and gaps. Later, the same first author and others [55] extended the distance measure to networks and based it this time on simple operations on hypergraphs, such as union, intersection and difference of two hypergraphs. Whole networks are also considered by Oh *et al.* [123] who used a kernel-based method to compute the similarity between metabolic networks in polynomial time. The features fed into the kernel-based algorithm are the connection information between any pair of enzymes and the phylogenetic trees were reconstructed using hierarchical clustering.

The main conclusion that can be drawn on such scenarii is that there remain serious limitations in the use of pathway or network comparison methods to infer evolutionary histories. The principal one is the adoption of functional distances between pathways/networks. If the tree is intended to reflect ancestry, then distances between pathways should integrate a realistic model for metabolism evolution. In general, there is a lack of discussion on the possible impact of the chosen distances on the reconstruction. It seems also that no available method satisfactorily considers the full structure of the networks. Clearly, this topic is still in its early stages.

4.3 Conclusion

We hope this tour through the literature on the structural analysis of metabolic networks will provide to researchers interested in the field, whether debutant or experienced, a good reference guide. On reaching the end of the tour, it appears clear that the structure of metabolic networks does behold functional and evolutionary information, even if not always. Basic measures such as the degree sequence or the network diameter thus seem to provide only a limited amount of information, if any. More advanced notions like modules and motifs on the other hand may help in getting a better grip on the biology of those networks, but they sometimes lack biological validation. The central notion of a realistic random graph model for metabolism is still open. Among the currently available ones, few integrate the notion of time, which would enable to model evolutionary processes. Random graphs appear notably in the context of hypothesis-testing as we saw repeatedly mentioned across the paper. In most cases, an observed property must be compared to its expected value under a null hypothesis in order to decide if the property is exceptional (*i.e.* unexpected) or, instead, can be explained by chance alone. Random graph models come into play at this point and are crucial to model what is expected under a null hypothesis. Unfortunately, null hypotheses are usually rarely discussed in the context of biological networks, partly due to the lack of good corresponding random graph models. Future directions in the field therefore include a thorough discussion of possible such models and, perhaps, the development of hypergraph analysis methods to more finely deal with the structure of metabolism.

The tour, long though it was, is also far from being complete. A lot more could be said on the topic, and many more interesting references could have been given and discussed. And this was concerned with “metabolism only”, and just one aspect of it! Understanding the relation between structure and dynamics (the way the structure actually defines or constrains the dynamics) is yet another fascinating topic for which Segrè *et al.* [158], for instance, provide one very nice example. Much work remains however to be done in this area, and a lot more on attempting to jointly model several levels of organisation such as genomic, transcriptomic and metabolic, even though the literature is already quite vast on both topics, as it is for each level of organisation considered separately. Almost none of this literature is cited in this paper. Presenting it would require another (very) long tour, but we must now "put the subject by, 'the rest next time-' " (Alice in Wonderland, Lewis Carroll). At some point also, as Gryphon answered to the Mock Turtle who was asking him to "Explain all that", "No, no! The adventures first, [...] explanations take such a dreadful time", one should stop reading what others have done and take a plunge into the topic, and the dirty data, oneself. Our final hope is to have provided any "just curious" reader of this paper the motivation for throwing him/herself happily into the deep water.

5 Acknowledgements

This work was partially supported by the ACI Nouvelles Interfaces des Mathématiques of the French Ministry of Research (Π-Vert Project), the Actions Concertées Coopératives of the INRIA (IBN Project), the ANR (REGLIS project NT05-3_45205), the ANR and the BBSRC (MetNet4SysBio project ANR-07-BSYS 003 02).

References

[1] *Webster's Unabridge Dictionary.* Random House.

[2] V. Acua, F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M.-F. Sagot, and L. Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. *BioSystems*, accepted, 2008.

[3] T. Akutsu. Efficient extraction of mapping rules of atoms from enzymatic reaction data. *J Comput Biol*, 11(2-3):449–462, 2004.

[4] E. Alm and A. P. Arkin. Biological networks. *Curr Opin Struct Biol*, 13(2):193–202, 2003.

[5] P. Aloy and R. B. Russell. Potential artefacts in protein-interaction networks. *FEBS Lett*, 530(1-3):253–254, 2002.

[6] S. F. Altschul et al. Basic local alignment search tool. *J Mol Biol*, 215:403–410, 1990.

[7] K. F. Aoki-Kinoshita and M. Kanehisa. Gene Annotation and Pathway Mapping in KEGG. *Methods Mol Biol*, 396:71–92, 2007.

[8] M. Arita. Metabolic reconstruction using shortest paths. *Simulation Practice and Theory*, 9:109–125, 2000.

[9] M. Arita. The metabolic world of Escherichia coli is not small. *Proc Natl Acad Sci U S A*, 101(6):1543–1547, 2004.

[10] M. Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[11] A. Bairoch. The ENZYME database in 2000. *Nucleic Acids Res*, 28(1):304–305, 2000.

[12] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[13] D. A. Benson et al. Genbank. *Nucleic Acids Res*, 35:21–25, 2007.

[14] R. J. Bino et al. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci*, 9(9):418–425, 2004.

[15] B. Bollobás, C. Borgs, J. Chayes, and O. Riordan. Directed scale-free graphs. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 132–139, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[16] I. Borodina, P. Krabben, and J. Nielsen. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res*, 15(6):820–829, 2005.

[17] R. Breitling et al. Ab initio prediction of metabolic networks using fourier transform mass spectrometry data. *Metabolomics*, V2(3):155–164, 2006.

[18] C. Bro, B. Regenberg, L Förster, and J. Nielsen. In silico aided metabolic engineering of saccharomyces cerevisiae for improved bioethanol production. *Metab Eng*, 8:102–111, 2005.

[19] A. P. Burgard, E. V. Nikolaev, C. H. Schilling, and C. D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res*, 14(2):301–312, 2004.

[20] G. Caetano-Anolls, H. Shin Kim, and J. E. Mittenthal. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci U S A*, 104(22):9358–9363, 2007.

[21] A. Cakmak and G. Ozsoyoglu. Mining biological networks for unknown pathways. *Bioinformatics*, 23(20):2775–2783, 2007.

[22] R. Caspi et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34(Database issue):D511–D516, 2006.

[23] L. Chen and D. Vitkup. Distribution of orphan metabolic activities. *Trends Biotechnol*, 25(8):343–348, 2007.

[24] H. N. Chua, W-K. Sung, and L. Wong. An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, 23(24):3364–3373, 2007.

[25] B. L. Clarke. Complete set of steady states for the general stoichiometric dynamical system. *J. Chem. Phys*, 75:4970–4979, 1981.

[26] C. Claudel-Renard, C. Chevalet, T. Faraut, and D. Kahn. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res*, 31(22):6633–6639, 2003.

[27] J. C. Clemente, K. Satou, and G. Valiente. Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. *Genome Inform*, 16(2):45–55, 2005.

[28] J. M. Colom and M. Silva. Convex geometry and semiflows in p/t nets. a comparative study of algorithms for computation of minimal p-semiflows. In *Proceedings of the 10th International Conference on Applications and Theory of Petri Nets*, pages 79–112. Springer-Verlag, 1981.

[29] S. J. Cordwell. Microbial genomes and "missing" enzymes: redefining biochemical pathways. *Arch Microbiol*, 172(5):269–279, 1999.

[30] S. Coulomb, M. Bauer, D. Bernard, and M-C. Marsolier-Kergoat. Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci*, 272(1573):1721–1725, 2005.

[31] M. W. Covert et al. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, 2004.

[32] M. W. Covert and B. O. Palsson. Constraints-based models: Regulation of gene expression reduces the steadystate solution space. *J. Theor. Biol*, 221, 2003.

[33] D. Croes, F. Couche, S. J. Wodak, and J. van Helden. Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol*, 356(1):222–236, 2006.

[34] C. Cunchillos and G. Lecointre. Evolution of amino acid metabolism inferred through cladistic analysis. *J Biol Chem*, 278(48):47960–47970, 2003.

[35] T. Dandekar et al. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J*, 343(Pt 1):115–124, 1999.

[36] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statis Comput*, 2008.

[37] Y. Deville, D. Gilbert, J. van Helden, and S. J. Wodak. An overview of data models for the analysis of biochemical pathways. *Brief Bioinformatics*, 4(3):246–59, 2003.

[38] S. N. Dorogovtsev, J. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Phys Rev Lett*, 85(21):4633–4636, 2000.

[39] N. C. Duarte et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 2007.

[40] J-F. Dufayard et al. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–2603, 2005.

[41] L. Duret, D. Mouchiroud, and M. Gouy. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res*, 22(12):2360–2365, 1994.

[42] J. J. Daz-Meja, E. Prez-Rueda, and L. Segovia. A network perspective on the evolution of metabolism by gene duplication. *Genome Biol*, 8(2):R26, 2007.

[43] J. S. Edwards and B. O. Palsson. Robustness analysis of the Escherichia coli metabolic network. *Biotechnol Prog*, 16(6):927–939, 2000.

[44] A. J. Enright and C. A. Ouzounis. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*, 2(9):RESEARCH0034, 2001.

[45] Y-H. Eom, S. Lee, and H. Jeong. Exploring local structural organization of metabolic networks using subgraph patterns. *J Theor Biol*, 241(4):823–829, 2006.

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

[47] A. M. Feist et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007.

[48] L. Félix and G. Valiente. Validation of metabolic pathway databases based on chemical substructure search. *Biomol Eng*, 2006. In press.

[49] L. Félix, G. Valiente, and F. Rossello. Optimal artificial chemistries and metabolic pathways. In *Proc. 6th Mexican Int. Conf. Computer Science, IEEE Computer Science Press*, pages 298–305. IEEE Computer Society, 2005.

[50] D. A. Fell and A. Wagner. The small world of metabolism. *Nat Biotechnol*, 18(11):1121–1122, 2000.

[51] O. Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Mol Biol*, 48(1-2):155–171, 2002.

[52] I. Fishtik, C.A. Callaghan, and R. Datta. Reaction route graphs. i. theory and algorithm. *Journal of Physical Chemistry B*, 108(18):5671–5682, 2004.

[53] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113, 1970.

[54] P. Fjallstrom. Algorithms for graph partitioning: A survey. *Computer and Information Science*, 3, 1998.

[55] C. V. Forst, C. Flamm, I. L. Hofacker, and P. F. Stadler. Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, 7:67, 2006.

[56] C. V. Forst and K. Schulten. Phylogenetic analysis of metabolic pathways. *J Mol Evol*, 52(6):471–489, 2001.

[57] E. Fridman and E. Pichersky. Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. *Curr Opin Plant Biol*, 8(3):242–248, 2005.

[58] I. Friedberg. Automated protein function prediction–the genomic challenge. *Brief Bioinform*, 7(3):225–242, 2006.

[59] J. Gagneur and S. Klamt. Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5:175, 2004.

[60] M. Y. Galperin, D. R. Walker, and E. V. Koonin. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res*, 8(8):779–790, 1998.

[61] M. R. Garey and D. S. Johnson. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. Freeman, 1979.

[62] A. Gattiker et al. Automated annotation of microbial proteomes in SWISS-PROT. *Comput Biol Chem*, 27(1):49–58, 2003.

[63] J. A. Gerlt. How to find ”missing” genes. *Chem Biol*, 10(12):1141–1142, 2003.

[64] J. A. Gerlt and P. C. Babbitt. Can sequence determine function? *Genome Biol*, 1(5):REVIEWS0005, 2000.

[65] G. V. Glazko and A. R. Mushegian. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol*, 5(5):R32, 2004.

[66] M. L. Green and P. D. Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5:76, 2004.

[67] R. Guimerà and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[68] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. A network-based method for target selection in metabolic networks. *Bioinformatics*, 23:1616–1622, 2007.

[69] R. Hall, M. Beale, O. Fiehn, N. Hardy, L. Sumner, and R. Bino. Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell*, 14(7):1437–1440, 2002.

[70] T. Handorf, O. Ebenhh, and R. Heinrich. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J Mol Evol*, 61(4):498–512, 2005.

[71] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52, 1999.

[72] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc*, 125(39):11853–11865, 2003.

[73] M. Heiner, I. Koch, and J. Will. Model validation of biological pathways using Petri nets–demonstrated for apoptosis. *BioSystems*, 75(1-3):15–28, 2004.

[74] M. Heymans and A. K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19 Suppl 1:i138–i146, 2003.

[75] A. B. Horne, T. C. Hodgman, H. D. Spence, and A. R. Dalby. Constructing an enzyme-centric view of metabolism. *Bioinformatics*, 20(13):2050–2055, 2004.

[76] N. H. Horowitz. On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci U S A*, 31(6):153–157, 1945.

[77] T. J. P. Hubbard et al. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–D617, 2007.

[78] C. J. Jeffery. Moonlighting proteins: old proteins learning new tricks. *Trends Genet*, 19(8):415–417, 2003.

[79] R. A. Jensen. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, 30:409–425, 1976.

[80] H. Jeong et al. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[81] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[82] H. Kacser and R. Beeby. Evolution of catalytic proteins or on the origin of enzyme species by means of natural selection. *J Mol Evol*, 20(1):38–51, 1984.

[83] M. Kanehisa et al. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280, 2004.

[84] M. Kanehisa et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):D354–D357, 2006.

[85] P. D. Karp et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*, 33(19):6083–9, 2005.

[86] P. D. Karp and S. Paley. Automated drawing of metabolic pathways. In *Proceedings of the Third International Conference on Bioinformatics and Genome Research*, 1994.

[87] P. D. Karp, S. Paley, and P. Romero. The Pathway Tools software. *Bioinformatics*, 18 Suppl1:225–238, 2002.

[88] E. F. Keller. Revisiting "scale-free" networks. *Bioessays*, 27(10):1060–1068, 2005.

[89] B. P. Kelley et al. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32(Web Server issue):W83–W88, 2004.

[90] P. Kersey et al. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res*, 33(Database issue):D297–D302, 2005.

[91] I. M. Keseler et al. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res*, 33(Database issue):D334–D337, 2005.

[92] R. Khanin and E. Wit. How scale-free are biological networks. *J Comput Biol*, 13(3):810–818, 2006.

[93] P. Kharchenko et al. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, 7:177, 2006.

[94] T. Kitami and J. H. Nadeau. Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nat Genet*, 32(1):191–194, 2002.

[95] S. Klamt. Generalized concept of minimal cut sets in biochemical networks. *BioSystems*, 83(2-3):233–247, 2006.

[96] S. Klamt and E. D. Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234, 2004.

[97] S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2):233–236, 2002.

[98] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends Biotechnol*, 21(2):64–69, 2003.

[99] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, New York, NY, USA, 2000. ACM.

[100] G. Kortsarz and D. Peleg. On choosing a dense subgraph. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium*, 1993.

[101] M. Kotera et al. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J Am Chem Soc*, 126(50):16487–16498, 2004.

[102] E. V. Kriventseva, N. Rahman, O. Espinosa, and E. M. Zdobnov. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res*, 2007.

[103] A. Kmmel, S. Panke, and M. Heinemann. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*, 7:512, 2006.

[104] V. Lacroix, C. G. Fernandes, and M-F. Sagot. Motif search in graphs: application to metabolic networks. *IEEE/ACM Trans Comput Biol Bioinform*, 3(4):360–368, 2006.

[105] A. Lazcano and S. L. Miller. On the origin of metabolic pathways. *J Mol Evol*, 49(4):424–431, 1999.

[106] O. Lespinet and B. Labedan. Puzzling over orphan enzymes. *Cell Mol Life Sci*, 63(5):517–523, 2006.

[107] H. Li et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34(Database issue):D572–D580, 2006.

[108] L. Liao, S. Kim, and J.F. Tomb. Genome comparisons based on profiles of metabolic pathways. In *Sixth international conference on knowledge-based intelligent informationand engineering systems, Crema, Italy*, pages 469–476, Crema, Italy, 2002.

[109] S. Light, P. Kraulis, and A. Elofsson. Preferential attachment in the evolution of metabolic networks. *BMC Genomics*, 6:159, 2005.

[110] H. Ma and A-P. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277, 2003.

[111] H-W. Ma, X-M. Zhao, Y-J. Yuan, and A-P. Zeng. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, 20(12):1870–1876, 2004.

[112] B. E. Maden. No soup for starters? Autotrophy and the origins of metabolism. *Trends Biochem Sci*, 20(9):337–341, 1995.

[113] R. Mahadevan and D. R. Lovley. The degree of redundancy in metabolic genes is linked to mode of metabolism. *Biophys J*, 2007. in press.

[114] M. L. Mavrovouniotis. Estimation of standard Gibbs energy changes of biotransformations. *J Biol Chem*, 266(22):14440–14445, 1991.

[115] F. Meyer et al. GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*, 31(8):2187–2195, 2003.

[116] Y. Moriya et al. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, 35(Web Server issue):W182–W185, 2007.

[117] H. J. Morowitz, J. D. Kostelnik, J. Yang, and G. D. Cody. The origin of intermediary metabolism. *Proc Natl Acad Sci U S A*, 97(14):7704–7708, 2000.

[118] N. Mulder and R. Apweiler. InterPro and InterProScan: Tools for Protein Sequence Classification and Comparison. *Methods Mol Biol*, 396:59–70, 2007.

[119] D. G. Naumoff, Y. Xu, N. Glansdorff, and B. Labedan. Retrieving sequences of enzymes experimentally characterized but erroneously annotated : the case of the putrescine carbamoyltransferase. *BMC Genomics*, 5(1):52, 2004.

[120] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[121] M. E. J. Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, 2006.

[122] I. Nobeli and J. M. Thornton. A bioinformatician's view of the metabolome. *Bioessays*, 28(5):534–545, 2006.

[123] S. J. Oh, J.-G. Joung, J.-H. Chang, and B.-T. Zhang. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics*, 7:284, 2007.

[124] A. Osterman and R. Overbeek. Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol*, 7(2):238–251, 2003.

[125] R. Overbeek et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–5702, 2005.

[126] S. M. Paley and P. D. Karp. Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics*, 18(5):715–724, 2002.

[127] B. O. Palsson. The challenges of *in silico* biology. *Nat. Biotechnol*, 18:1147–1150, 2000.

[128] J. A. Papin, J. L. Reed, and B. O. Palsson. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem Sci*, 29(12):641–647, 2004.

[129] B. Papp, C. Pàl, and L. D. Hurst. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992):661–664, 2004.

[130] M. Parter, N. Kashtan, and U. Alon. Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol*, 7:169, 2007.

[131] M. Parter, N. Kashtan, and U. Alon. Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol*, 7:169, 2007.

[132] I. Paulsen and A. von Haeseler. INVHOGEN: a database of homologous invertebrate genes. *Nucleic Acids Res*, 34(Database issue):D349–D353, 2006.

[133] M. Pellegrini et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, 1999.

[134] T. Pfeiffer, O. S. Soyer, and S. Bonhoeffer. The evolution of connectivity in metabolic networks. *PLoS Biol*, 3(7):e228, 2005.

[135] R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408, 2005.

[136] M. G. Poolman, C. Sebu, M. K. Pidcock, and D. A. Fell. Modular decomposition of metabolic systems via null-space analysis. *J Theor Biol*, 249(4):691–705, 2007.

[137] Y. Pouliot and P. D. Karp. A survey of orphan enzymatic activities. *BMC Bioinformatics*, 8:244, 2007.

[138] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61–D65, 2007.

[139] T. M. Przytycka and Y-K. Yu. Scale-free networks versus evolutionary drift. *Comput Biol Chem*, 28(4):257–264, 2004.

[140] C. Pl et al. Chance and necessity in the evolution of minimal metabolic networks. *Nature*, 440(7084):667–670, 2006.

[141] C. Pl, B. Papp, and M. J. Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*, 37(12):1372–1375, 2005.

[142] H. Jeong R. Albert and A-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[143] S. A. Rahman et al. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, 21(7):1189–1193, 2005.

[144] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[145] R. Ravi, A. Agrawal, and P. Klein. Ordering problems approximated: single-processor scheduling and interval graph completion. In *Proceedings of the 18th international colloquium on Automata, languages and programming*, pages 751–762, New York, NY, USA, 1991. Springer-Verlag New York, Inc.

[146] M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052, 2001.

[147] S. C. G. Rison, S. A. Teichmann, and J. M. Thornton. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J Mol Biol*, 318(3):911–932, 2002.

[148] S. C. G. Rison and J. M. Thornton. Pathway evolution, structurally speaking. *Curr Opin Struct Biolo*, 12(3):469–473, 2002.

[149] I. B. Rogozin et al. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*, 30(10):2212–2223, 2002.

[150] S. Roy. Multifunctional enzymes and evolution of biosynthetic pathways: Retro-evolution by jumps. *Proteins: Structure, Functions, and Genetics*, 37:303–309, 1999.

[151] C. H. Schilling, D. Letscher, and B. O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, 203(3):229–248, 2000.

[152] C. H. Schilling, S. Schuster, B. O. Palsson, and R. Heinrich. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog*, 15(3):296–303, 1999.

[153] S. Schmidt, S. Sunyaev, P. Bork, and T. Dandekar. Metabolites: a helping hand for pathway evolution? *Trends Biochem Sci*, 28(6):336–341, 2003.

[154] I. Schomburg et al. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*, 32(Database issue):D431–D433, 2004.

[155] S. Schuster, D. A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol*, 18(3):326–332, 2000.

[156] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *J Biol Systems*, 2:165–182, 1994.

[157] S. Schuster, C. Hilgetag, J. H. Woods, and D. A. Fell. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J Math Biol*, 45(2):153–181, 2002.

[158] D. Segrè, D. Vitkup, and G. M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA*, 99(23):15112–15117, 2002.

[159] D. Segura, R Mahadevan, K Juárez, and D.R. Lovley. Computational and experimental analysis of redundancy in the central metabolism of *geobacter sulfurreducens*. *PLoS Comput Biol*, 4(2):e36, 2005.

[160] F. Servant et al. ProDom: automated clustering of homologous domains. *Brief Bioinform*, 3(3):246–251, 2002.

[161] P. D. Seymour. The matroids with the max-flow min-cut property. *J. Comb. Theory Ser. B*, 23(7):189–222, 1977.

[162] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 31(1):64–68, 2002.

[163] B. Snel and M. A. Huynen. Quantifying modularity in the evolution of biomolecular systems. *Genome Res*, 14(3):391–397, 2004.

[164] V. Spirin, M. S. Gelfand, A. A. Mironov, and L. A. Mirny. A metabolic network in the evolutionary context: multiscale structure and modularity. *Proc Natl Acad Sci U S A*, 103(23):8774–8779, 2006.

[165] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21):12123–12128, 2003.

[166] J. Stelling. Mathematical models in microbial systems biology. *Curr Opin Microbiol*, 7(5):513–518, 2004.

[167] J. Stelling et al. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 2002.

[168] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A*, 102(12):4221–4224, 2005.

[169] P.F. Suthers, A.P. Burgard, M.S. Dasika, F. Nowroozi, S. Van Dien, J.D. Keasling, and C.D. Maranas. Metabolic flux elucidation for large-scale models using 13c labeled isotopes. *Metab Eng*, 9(5-6):387–405, 2007.

[170] Z. Szallasi, J. Stelling, and V. Periwal. *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts*. The MIT Press, 2006.

[171] T. Tanaka, K. Ikeo, and T. Gojobori. Evolution of metabolic networks by gain and loss of enzymatic reaction in eukaryotes. *Gene*, 365:88–94, 2006.

[172] T. Tanaka, K. Ikeo, and G. Takashi. Evolution of metabolic networks by gain and loss of enzymatic reaction in eukaryotes. *Gene*, 365:88–94, 2006.

[173] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 28(1):33–36, 2000.

[174] S. A. Teichmann et al. The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. *J Mol Biol*, 311(4):693–708, 2001.

[175] B. Teusink et al. In silico reconstruction of the metabolic pathways of Lactobacillus plantarum: comparing predictions of nutrient requirements with those from growth experiments. *Appl Environ Microbiol*, 71(11):7253–7262, 2005.

[176] I. Thiele, T.D. Vo, N.D. Price, and B.O. Palsson. Expanded metabolic reconstruction of helicobacter pylori (iit341 gsm/gpr): an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol*, 187:5818–5830, 2005.

[177] K. F. Tipton. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. *Eur J Biochem*, 223(1):1–5, 1994.

[178] Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc Int Conf Intell Syst Mol Biol*, 8:376–383, 2000.

[179] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[180] R. Urbanczik and C. Wagner. An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, 21(7):1203–1210, 2005.

[181] D. Vallenet et al. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res*, 34(1):53–65, 2006.

[182] N. C. VerBerkmoes, H. M. Connelly, C. Pan, and R. L. Hettich. Mass spectrometric approaches for characterizing bacterial proteomes. *Expert Rev Proteomics*, 1(4):433–447, 2004.

[183] D. Vitkup, P. Kharchenko, and A. Wagner. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol*, 7(5):R39, 2006.

[184] G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406(6792):188–192, 2000.

[185] K. Voss, M. Heiner, and I. Koch. Steady state analysis of metabolic pathways using petri nets. *In Silico Biol*, 3(3):367–387, 2003.

[186] G. Wächtershäuser. Evolution of the first metabolic cycles. *Proc Natl Acad Sci USA*, 87(1):200–204, 1990.

[187] G. Wächtershäuser. On the chemistry and evolution of the pioneer organism. *Chem Biodivers*, 4(4):584–602, 2007.

[188] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478):1803–1810, 2001.

[189] G. P. Wagner. Homologues, natural kinds and the evolution of modularity. *Integrative and Comparative Biology*, 36(1):36–43, 1996.

[190] G. P. Wagner, M. Pavlicev, and J. M. Cheverud. The road to modularity. *Nat Rev Genet*, 8(12):921–931, 2007.

[191] M. A. Wagner et al. Global analysis of the Brucella melitensis proteome: Identification of proteins expressed in laboratory-grown culture. *Proteomics*, 2(8):1047–1060, 2002.

[192] S. G. Waley. Some aspects of the evolution of metabolic pathways. *Comp Biochem Physiol*, 30(1):1–11, 1969.

[193] Z. Wang et al. Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria. *BMC Genomics*, 7(1):100, 2006.

[194] S. Wasserman and K. Faust. *Social Network Analysis Methods and Applications.* Cambridge University Press, 1994.

[195] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[196] S. Wernicke and F. Rasche. Simple and fast alignment of metabolic pathways by exploiting local diversity. *Bioinformatics*, 23(15):1978–1985, 2007.

[197] D. M. Wolf and A. P. Arkin. Motifs, modules and games in bacteria. *Curr Opin Microbiol*, 6(2):125–134, 2003.

[198] Z. Wunderlich and L. A. Mirny. Using topology of the metabolic network to predict viability of mutant strains. *Biophysics J*, 91:2304–2311, 2006.

[199] T. Yamada, S. Goto, and M. Kanehisa. Extraction of phylogenetic network modules from prokayrote metabolic pathways. *Genome Inform*, 15(1):249–258, 2004.

[200] Y. Yamanishi et al. Prediction of missing enzyme genes in a bacterial metabolic network. Reconstruction of the lysine-degradation pathway of Pseudomonas aeruginosa. *FEBS J*, 274(9):2262–2273, 2007.

[201] I. Yanai, A. Derti, and C. DeLisi. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U S A*, 98(14):7940–7945, 2001.

[202] F. Yang, H. Qian, and D. A. Beard. Ab initio prediction of thermodynamically feasible reaction directions from biochemical network stoichiometry. *Metab Eng*, 7(4):251–259, 2005.

[203] M. Ycas. On earlier states of the biochemical system. *J Theor Biol*, 44(1):145–160, 1974.

[204] J. Yoon, Y. Si, R. Nolan, and K. Lee. Modular Decomposition of Metabolic Reaction Networks based on Flux Analysis and Pathway Projection. *Bioinformatics*, 23(18):2433–244, 2007.

[205] J. Zhao et al. Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics*, 7:386, 2006.

[206] D. Zhu and Z. S. Qin. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, 6:8, 2005.