RICE UNIVERSITY

Graph-based Modeling and Evolutionary Analysis of Microbial Metabolism

by

Wanding Zhou

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Luay K. Nakhleh, *Director*
Associate Professor of Computer Science

Jianpeng Ma, *Chair*
Professor of Bioengineering

Oleg Igoshin
Assoicate Professor of Bioengineering

George N. Bennett
Professor of Biochemistry and Cell Biology

HOUSTON, TEXAS
APRIL 2013

Abstract

Graph-based Modeling and Evolutionary Analysis of Microbial Metabolism

by

Wanding Zhou

Microbial organisms are responsible for most of the metabolic innovations on Earth. Understanding microbial metabolism helps shed the light on questions that are central to biology, biomedicine, energy and the environment. Graph-based modeling is a powerful tool that has been used extensively for elucidating the organising principles of microbial metabolism and the underlying evolutionary forces that act upon it. Nevertheless, various graph-theoretic representations and techniques have been applied to metabolic networks, rendering the modeling aspect *ad hoc* and highlighting the conflicting conclusions based on the different representations.

The contribution of this dissertation is two-fold. In the first half, I revisit the modeling aspect of metabolic networks, and present novel techniques for their representation and analysis. In particular, I explore the limitations of standard graph representations and the utility of the more appropriate model—hypergraphs—for capturing metabolic network properties. Further, I address the task of metabolic pathway inference and the necessity of accounting for chemical symmetries and alternative tracings in this crucial task.

In the second part of the dissertation, I focus on two evolutionary questions. First, I investigate the evolutionary underpinnings of the formation

of communities in metabolic networks—a phenomenon that has been reported in the literature and implicated in an organism's adaptation to its environment. I find that the metabolome size better explains the observed community structures. Second, I correlate evolution at the genome level with emergent properties at the metabolic network level. In particular, I quantify the various evolutionary events (e.g., gene duplication, loss, transfer, fusion, and fission) in a group of proteobacteria, and analyze their role in shaping metabolic networks and determining organismal fitness.

As metabolism gains an increasingly prominent role in biomedical, energy, and environmental research, understanding how to model this process and how it came about during evolution becomes more crucial. My dissertation provides important insights on both of the two issues.

Acknowledgements

My deepest gratitude goes to the following individuals:

To Luay Nakhleh, my academic advisor, for all of the scientific, professional and financial support he provided in helping me obtain the Ph.D degree. All of the research discussed in this dissertation is an outcome of our close collaboration since the beginning of 2009, a time when I, as an international student, was in a very difficult situation. Were it not for Luay's trust and subsequent rescue, I may not have obtained a degree at Rice. Joining his group simply displaced my dejection and revealed to me the promise of future academic development. Luay is an outstanding mentor. Given the number of student he has, I don't think I would have received an equal amount of attention had I worked anywhere else, let alone as a "refugee" from a different department. I still remember the summer when Luay took me out for lunch almost every other day and we chatted about science, academia and life—all candid, no holds barred. Under his mentoring, I received full training as an independent scholar in the making. I had the chance of being the submitting author and the correspondence author of most of the papers I authored. I luxuriated in the freedom of exploring any idea, stupid and whimsical as they might be. Luay is always patient with me even, when some research projects turned out unrewarding and when life was hard. As a student, I could not have bennefitted more. I sincerely thank Luay for all that he did for me.

To Jianpeng Ma, who, kindly serves as the chair of my committee

and lent me generous assistance in my career development. I had several thought-provoking discussions with Jianpeng at various points in my graduate study. Jianpeng is a distinguished scientist in the field of biophysics and one of Chinese descent. His advice for me was alarmingly critical and valuable.

To Deepak Nagrath, who collaborated with me, guided me in problems on convex optimization and provided generous assistance in my career development. To my research, Deepak gave precious opinions from an experimental biologist's perspective.

To Oleg Igoshin and George Bennett, my committee members, for instrumental criticisms upon reviewing my dissertation. Oleg and George are the most knowledgable systems biologist and biochemist I knew on campus. Their work always humbles me. Oleg's class tremendously widened my horizon and deepened my understanding of systems biology. Discussions with them were always very helpful.

To Weijia Li, my roomate, for taking care of me when I was sick. His support and company made my time as a dissertating student bearable.

To Alexander Bufetov, Michael Boshernitzan, Fathi Ghorbel and Ramon Gonzalez. They are excellent instructors from whose classes I learnt a lot and was greatly inspired.

To Kattie Basnett from the Department of English who helped edit the language of this dissertation.

To my friends and colleagues who shared with me their expertise, availed me of profound ideas and opinions, accompanied and tided me through the difficult times: Ioan Sucan, Troy Ruths, Liheng Zheng, Hyun Jung Park, Nikola Ristic, Natalie Berestovsky, Ji Qi, Wei Zhou, Yun Yu,

Yun Zhu, Matt Barnett, Kevin Liu, Jianrong Dong, Yong Cui, Zhechao Ruan, Ning Ruan, Ye Yuan and Hang Wen.

To other faculty members, staff and fellow graduate students, in both the Department of Computer Science and the Department of Bioengineering at Rice University, for their understanding and kindness.

Last but foremost, to my parents, for their unconditional love and care and for bringing me into this world. My father taught me a lot about how to face this life—what to fight for and what to let go, while my mother constantly reminded me to take care of myself in tough situations. When nights were cold and my heart was forlorn, they related to me, from the other side of the ocean, with interesting news and joyful stories which aroused in me the fondest memories. I know I should have spent more time with them. I can never thank them enough for their understanding. Little I might have achieved, I must attribute to my parents' love and support.

To pioneers in life sciences.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and background

Microorganisms, or microbes, affect our lives in many ways. Under healthy conditions, they contribute to the homeostasis of the human body — they help us digest food, produce essential vitamins and shield us from invading pathogens. Under pathological conditions, they themselves are the intruding pathogens. They are major players in circulating carbohydrates and water in the ecosystem we live. They are utilized to produce medicine (such as antibiotics) and food (such as bread and alcoholic products). Nowadays, microbes are also given much attention due to their potential to be engineered to produce biofuels — a promising alternative energy solution to the consumption of petroleum.

Understanding the diverse metabolism of microbes is one of the key steps to understanding the role microbes play in performing the aforementioned functions. For example, knowledge of metabolic enzymes and metabolic reactions is crucial for discovering novel biochemical pathways, a process computationally referred to as pathway inference. Better understanding of microbial metabolism also leads to the rational engineering of microbes for the purpose of optimizing the yield of desirable chemical compounds.

Studying the metabolism of microorganisms also helps biochemists understand the

metabolism of more complex, less primitive multi-cellular organisms. To start with studying the metabolism of microorganisms has several advantages. For instance, the relatively simple forms and smaller genomes of microbes makes them preferred targets for studying metabolism in both experiments and computational modeling. Model organisms such as *E.coli* and *S.cerevisiae* are easy to manipulate in the lab. Sequencing the whole genome is easier and can be applied to a larger number of taxa to make phylogenetic, or even population level studies feasible. Despite great distinction in the organization of the metabolism, a limited but substantial amount of conservative homologous enzymes share function between single cellular organisms (such as yeast) and multi-cellular organisms (such as plants).

Over centuries, a large amount of knowledge has been collected on the metabolism of microbes. This knowledge has centered around the metabolic reactions, the enzymes that catalyze those reactions and the genes that encode those enzymes. Generations of biochemists have left us a treasure of information on the properties of these reactions and enzymes. Our understanding of the metabolism of certain model organisms, such as *E.coli* and *S. cerevisiae*, almost covers their entire genomes. This expansion of knowledge promotes the study of the metabolic systems of up to thousands of metabolites and reactions — instead of systems of fewer than 10 components, whose study had gone almost hand-in-hand with the early experimental discoveries in microbial metabolism.

The development of computing technology that has revolutionized life has also fundamentally altered the way we research microbial metabolism. The annotation and aggregation of metabolic information—a process usually referred to as model integration or reconstruction—can now be automated using computers. Distributed computing technology enables the comparative study of the metabolism of thousands of different species. *In-silico* modeling is now a standard approach for analyzing the

dynamic properties and evolutionary histories of microbial metabolism.

Towards the goal of understanding the microbial metabolism, my PhD work focuses on the graph-based modeling of metabolic networks and how the metabolic networks evolve. Modeling metabolism in the graph theoretical fashion is widely used to elucidate the so-called organising principles of metabolism. More accurate representation of metabolism leverages more detailed knowledge of the chemical structures and reaction mechanisms. The fact that several chemical structures might be equally important, a phenomenon called symmetry, is a great obstacle to concatenating atom mappings of different reactions for the purpose of tracing atoms in metabolic networks. I formulate the problem of symmetry as a graph automorphism problem and investigate the extent to which symmetries give rise to alternative mappings.

To further study the organizing principle of metabolism, I consider it in the light of evolution. Previous researchers have found a relationship between how modular metabolic networks are as is quantified by a measure called modularity and the variability of the microbes' living habitats. Such relationship is claimed to be an outcome of evolvability. Studying the evolution of metabolic genes also helps to understand the relationship between their biological functions and the evolutionary events they went through, such as the horizontal gene transfer, as well as the logic of selection that gives rise to such relationships.

In the following I will discuss these two topics in detail.

1.1 Metabolic modeling

Metabolism is a biological process where a set of small chemical compounds, called *metabolites*, are consecutively transformed into each other via metabolic *reactions* controlled by a set of *enzymes* to produce energy or biomass so that the organisms can grow and reproduce. The interaction pattern of these metabolites, reactions and

enzymes can be represented in a *graph*—a mathematical structure $\mathcal{G} = (V, E)$ where V denotes the set of vertices and E the set of edges that connect elements in V. Such representation provides a platform for one to study metabolic interactions, to analyze how metabolites, enzymes and reactions are organized, and to explain evolutionary principles underlying such organization.

Graph-based modeling, or the topological study of the metabolic networks, together with constraint-based modeling and the dynamic simulation form the three major modeling platforms are used for analyzing microbial metabolism.

Graph-based modeling aims at describing, explaining the organization of metabolic networks and looking for topological features that highlight the evolutionary principles, neutral or adaptive alike, that govern the evolution of microbial metabolic networks. In contrast, constraints-based modeling, including flux balance analysis [1, 2, 3, 4] (see [5] for a review), elementary flux mode [6, 7, 6, 8] (see [9] for a review) and extreme pathway analysis [10] incorporates the stoichiometry and estimates the relative magnitude of reaction fluxes by assuming the steady states of the system—there is no net gain or loss in the metabolites' concentration over time. Ordinary differential equations (ODE) (e.g., [11, 12, 13], see [14] for a review) and other executable models, such as Petri-nets [15, 16, 17, 18] (See [19] for a review), belong to the third major class of metabolic network simulation. They have the strength of explicitly capturing the dynamics of the system. However, these models suffer from inadequate parameterization and hence, are limited to small or simplified metabolic systems.

1.1.1 Graph and hypergraph-based modeling

Two major questions that graph-based modeling, or topological modeling of metabolic networks, attempts to address are: what are the evolutionary principles that are be-

hind certain topological features of the metabolic network, and how to characterize the function of a metabolic subsystem. These two topics are elaborated in the following.

1.1.1.1 Organising principles as outcomes of adaptive/neutral evolution

When the network is large in size and complex in connection, to characterize the network structure poses a challenge. Conventional graph-theoretic concepts such as geodesic distance, pairwise clustering coefficient, node accessibility, betweenness, and cyclomatic number are still used to describe the local structure of the network. Huge efforts have been made to characterize the global topology of the graph either by bringing local characteristics into a statistical context (e.g. diameter, radius, average betweenness, degree distribution, global clustering coefficient, characteristic path length) or by designing new tools (community structure and modularity).

Based on these graph-theoretic concepts, certain topological features of metabolic networks have been identified and debated. For example, metabolic networks are shown to have a small-world structure [20, 21], scale-free [22] or, on the contrary, self-dissimilar and scale-rich [23, 24], and have a bow-tie pattern [25, 26] which is nested as a sub-network isolated from the giant component of network [27].

There is a growing trend in recent years to look for the topological features of metabolic networks that can be associated with the classification of the organism based on the characterization of the macroscopic phenotypes (such as the temperature preference, oxygen tolerance and the ability to adapt to different environments). For example, in prokaryotic species, Takemoto shows that higher temperature of the living environment would lead to fewer short-cuts in metabolic networks [28]. Modularity of metabolic networks correlates to the variability of a species' habitat [29], the size of the network, obligability of the living environment, and niche specializa-

tion [30]. A "seed set" of each metabolic network is defined and studied to describe the biochemical interface of an organism with the external environment and can be used to infer the adaptation of species' metabolism to a changing environment during evolution [31]. Using the network expansion method [32], Raymond et al. [33] showed that metabolic networks in aerobic, facultative and anaerobic species have different network expansion in the presence of O_2 as well as other common metabolites. They demonstrated how O_2 availability can change the architecture of the metabolic network.

Many biological questions regarding metabolic networks have been tackled in the light of evolution. For example, Zhao et al. [34] showed that genes inside a network module are more likely to co-evolve. Core modules evolve slower and peripheral modules evolve faster. Liu et al. [35] looked at enzymes in terms of their phylogenetic profiles (i.e. the number of species that contain the enzyme) from a set of prokaryotic species. They found that enzymes with higher phylogenetic profiles have higher betweenness. Diaz-Mejia et al. [36] showed that, within a functional module of a metabolic network, duplicates are more likely to be retented. Evolutionary modules where genes co-evolve are shown to match structural modules in metabolic networks on three different levels of inspection [37]. In a study that investigated a *Drosophila* dataset, focusing on metabolism, enzymes with substrates sharing with many other enzymes (or at the branch points of different pathways) are shown to be under stronger selection constraint [38].

1.1.1.2 Functional characterization

Another category of problem that can be approached using graph-based modeling is the functional characterization of practical applications, which usually requires unique reconstruction of the metabolic network or a special means of characterizing

the network topology that encodes functional information. For example, diseases connected with mutated enzymes that catalyze adjacent metabolic reactions form a metabolic disease network (MDN) which can be used to infer disease co-morbidity using the degree information of the network [39]. Bottleneck nodes in a metabolic network, signaled by a higher betweenness centrality, are shown to be more likely to be essential proteins [40]. Hubs in a metabolic network are shown to possess a hub status in gene co-expression networks and interaction networks as well [41]. Networks with nodes weighted on the number of reactions a metabolite participates in can be used to infer biochemically meaningful pathways by tracking the path with the lowest accumulation of weight [42, 43]. Metabolic networks can be used to infer operons [44] given the fact that enzymes close in networks are, by a higher chance, close on chromosomes [45]. Combined with linear physical graphs, Ogata *et al.* [46] inferred related enzyme clusters (FREC) which help build ortholog relationships. Network resilience (or robustness) refers to the resistance of network integrity or node accessibility under the removal of a random or particular choice of node or edge. It was shown that a metabolic network is extremely robust compared to the null model (random graph preserving the degree distribution) [47].

1.1.1.3 Standard graph representation

Standard graph representations are assembled based on metabolic reconstruction—a collection of knowledge of all of the metabolic reactions, enzymes and genes relevant to the organism's metabolism. To assemble the standard graph representation, one needs to decide on the node semantics (what nodes in the graph represent) and edge semantics (by what the nodes are linked by edges). In addition, different *pruning* methods can be applied to remove biochemically irrelevant edges before the metabolic graph is analyzed. Graph-theoretic characterizing functions are computed on these metabolic

graphs. Results are compared to assess how different treatment in representing the network can affect the topological features, as indicated by these characterizing functions. It is crucial to make sure that results are not sensitive to the choice of specific graph representations.

1.1.1.4 Graph-theoretic terminology and methods

Diameter The *diameter* is defined as the maximum geodesic distance over all pairs of vertices in the graph.

Clustering coefficient (global) The *clustering coefficient* C (also called *transitivity*) is defined as,

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{1.1}$$

Betweenness (geodesic) The *vertex (edge) betweenness* is defined as the number of geodesics passing though the vertex (edge).

Betweenness (random walk) The *vertex (edge) betweenness* is defined as the average of the number of times that a random walk between a pair of vertices will pass a particular vertex (edge) over all pairs of vertices.

Clique A clique is a completely connected graph with edges between every pair of nodes.

k-section Given an integer $k > 0$, the *k-section* of hypergraph $H_{(k)} = (X, \mathcal{E}_{(k)})$ formed by X and the set

$$\mathcal{E}_{(k)} = \{F \subset X \mid 1 \leq |F| \leq k; F \subset E_i \text{ for some } E_i \in \mathcal{E}\} \tag{1.2}$$

In other words, it is a hypergraph with hyperedges of all sets with cardinality less than k and is included in at least one hyperedge of the original graph.

1.1.1.5 Hypergraph representation

To capture more biological features than standard graph representation can code, generalizations of the standard graph have been extensively used. Nodes and edges of the graph can carry biologically relevant attributes. Edges can be weighted to quantify the strength of the linkages between nodes. The characterizing functions can also be adapted to these generalized graph structures. A certain generalization of graph called hypergraph proves particularly suitable for the metabolic networks which are composed of reactions involving an arbitrary number of metabolites.

As an extension of common graph formalism, the hypergraph can overcome the limitations of standard graphs, where one edge connects exactly two nodes. Intuitively, a hypergraph $\mathcal{H} = (X, \mathcal{E})$ is a graph for which an edge (called *hyperedge*) $E \in \mathcal{E}$ may subsume more than two nodes in X (see formal definition below). The number of nodes connected by a hyperedge is called the *cardinality* of the hyperedge. When all hyperedges have cardinality 2, the hypergraph reduces to a standard graph.

Or, more formally, given a finite set X and a family of subsets $\mathcal{E}$ on X. $\mathcal{E}$ is called a hypergraph if

1. $E \neq \emptyset \quad (E \in \mathcal{E})$

2. $\bigcup_{E \in \mathcal{E}} E = X$

$|X|$ is called the *order of the hypergraph*. X is called the *node set*. $\mathcal{E}$ is called the *hyperedge set*. The hypergraph is called *simple* if edges are all distinct. If $|E| = 2$ for all $E \in \mathcal{E}$, a simple hypergraph reduces to a *standard graph*.

Given an integer $k > 0$, the *k-section* of hypergraph $H_{(k)} = (X, \mathcal{E}_{(k)})$ formed by X

and the set

$$\mathcal{E}_{(k)} = \{F \subset X \mid 1 \leq |F| \leq k; F \subset E_i \text{ for some } E_i \in \mathcal{E}\} \tag{1.3}$$

In other words, it is a hypergraph with hyperedges of all sets with cardinality less than k and is included in at least one hyperedge of the original graph.

A hyperedge captures the essence of a reaction by subsuming all of the reactants in one instead of representing them as a clique where the intrinsic integrity of the reaction entity is obscured (see Klamt et al. [48] for an illustration of the problem). The relative significance of the interactions among different pairs of reactants are systematically understated. To further tailor to the nature of the metabolic reaction, a hyperedge can be bipartitioned to reactants on opposite sides of a reaction by considering hyperedges as a 2-tuple $E = (T, H)$ where T is the *tail set* and H is the *head set*. The hyperedge can be naturally directed by distinguishing between T and H and directing from T to H, giving rise to a *directed hyperedge* or *hyperarc* (see [49] for an early review). Towards a more accurate description, a hyperarc can have two extra functions, $c_T : T \to \mathbb{N}$ and $c_H : H \to \mathbb{N}$, mapping each metabolite to its stoichiometric coefficient in the reaction. This is equivalent to a special stoichiometric matrix where we represent the stoichiometric coefficients of substrates and products in different signs and treat every reversible reaction as two separate reactions with opposite directions.

In some circumstances, the graphical behavior is emphasized by considering a hyperedge as a generalized edge represented in drawings by a loop if it has cardinality 1, a curve joining the two vertices if it has cardinality 2 and a closed circle enclosing all its nodes if it has cardinality higher than 2 (see Figure 1.1 for an example). A hypergraph with no hyperedge being a subset of another hyperedge is called a *Sperner hypergraph*. For example, the hypergraph in Figure 1.1 is not Sperner as hyperedge

E_6 is included in E_7. Without E_6 the hypergraph becomes Sperner.

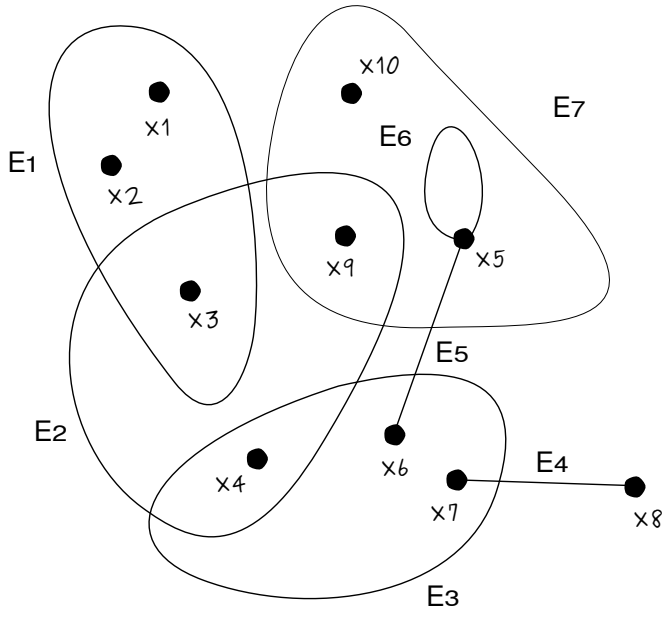

Figure 1.1: **Illustration of a hypergraph**

In others, its resemblance to matrix behavior is more amenable to analysis. Theoretically, each hypergraph corresponds to a boolean matrix (matrix with entry of only 0 and 1) as long as no row or columns of it contains only zero, which violates the two conditions in the definition of a hypergraph. Such a matrix is called an *incidence matrix* I of the hypergraph and can be constructed by considering its m rows for m hyperedges and n columns for n nodes such that I_i^j is equal to 1 if node i is contained in hyperedge j and 0 otherwise. The incidence matrix for the hypergraph in Figure

1.1 is,

$$I = \begin{array}{c} \\ \mathrm{x}_1 \\ \mathrm{x}_2 \\ \mathrm{x}_3 \\ \mathrm{x}_4 \\ \mathrm{x}_5 \\ \mathrm{x}_6 \\ \mathrm{x}_7 \\ \mathrm{x}_8 \\ \mathrm{x}_9 \end{array} \begin{array}{c} \begin{array}{ccccccc} \mathrm{E}_1 & \mathrm{E}_2 & \mathrm{E}_3 & \mathrm{E}_4 & \mathrm{E}_5 & \mathrm{E}_6 & \mathrm{E}_7 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{1.4}$$

For directed hypergraphs, the incidence matrix $(I)_{ij}$ can be defined as,

$$I_{ij} = \begin{cases} -1 & \text{if } x_i \in T(E_j) \\ 1 & \text{if } x_i \in H(E_j) \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

A stoichiometric matrix S widely used in modeling metabolic systems can be considered a generalization of an incidence matrix where reactions are hyperedges and metabolites are nodes, and where if metabolite i participates reaction j, then S_i^j is equal to the stoichiometric coefficient of that metabolite [50]. Please refer to [51, 52] for a detailed introduction of hypergraph properties.

There are several concepts that transform a hypergraph to standard graphs or other hypergraphs. Some of them are already tacitly used in reconstruction of simple metabolic networks. Among the most important ones, a *dual hypergraph* $\mathcal{H}^*$ switches the roles of nodes and hyperedges in a hypergraph $\mathcal{H}$ by taking the set of hyperedges

in $\mathcal{H}$ as its node set and defining new hyperedge X_j^* as,

$$X_j^* = \{E_i \in \mathcal{E} \mid E_i \ni x_j \text{ in } \mathcal{H}\} \tag{1.6}$$

Conceived in the matrix formalism, the operation of taking dual is equivalent to taking the transpose of the original incidence matrix. Unlike standard graphs where extra constraints on the edge cardinality prevent the dual graph from being simple, hypergraphs are closed on the operation of taking duals (i.e. the dual of a hypergraph must be a hypergraph). This is obvious as no row or column of a transposed matrix contains all zero if the original matrix does not. Besides, the idea of the *primal graph* of a hypergraph transforms it into a simple one by representing a hypergraph in a clique of elements in simple edges. More general transformations are possible by considering the k-section of a hypergraph where taking primal can be regarded as a special case of taking a 2-section without loops.

Many existing standard graph models can be brought into the general framework of a hypergraph representation. Take as an example, a substance graph model where metabolites are treated as nodes and all metabolites participating a reaction are connected. This can be considered the primal graph of the hypergraph model, where reactions stand for hyperedges and metabolites participating the reaction represent nodes that are included in that hyperedge. A reaction network where reactions are connected as long as they share common reactants can be considered equivalent to taking the dual of the hypergraph model considered above and then taking the primal. Similarly, an enzyme graph model where the reactions catalyzed by two enzymes couple with each other and give rise to a connection between the enzymes can be considered as merging and splitting nodes in a reaction hypergraph and then taking the primal.

1.1.1.6 Hypergraph characteristics adaptation

Many standard graph characteristics can be adapted into hypergraphs [48]. A comprehensive adaptation of graph-theoretical tools into both undirected and directed hypergraphs will be carried out. A metabolic network will be reconstructed using both standard graph representation and hypergraph representation. Hypergraph specific features will be analyzed, such as hyperedge cardinality distribution. Then graph/hypergraph characteristics will be applied to these representations. Finally, we will check whether previously obtained results can be reproduced. The graph-theoretic characteristics include: degree, diameter, radius, geodesics, characteristic path length, clustering coefficient, betweenness, modularity, robustness etc.

Some straight forward adaptations are:

degree The *degree* of a hypergraph node is defined as the number of hyperedges containing the node.

path A *path of length* q is defined as a sequence $(x_1, E_1, x_2, E_2, \ldots, E_q, x_{q+1})$ such that,

1. $x_1, x_2, \ldots, x_q$ are all distinct.
2. $E_1, E_2, \ldots, E_q$ are all distinct.
3. $x_k, x_{k+1} \in E_k$ for $k = 1, 2, \ldots, q$.

A path is called a *cycle* if $x_1 = x_{q+1}$ and $q > 1$.

geodesics (shortest path) The *geodesics*, or shortest path between any two pairs of nodes (x, y) in a hypergraph, is defined as the path with minimal length connecting x and y. The minimal length is called *geodesic length*.

diameter The *diameter* of a hypergraph is defined as the maximum geodesic length among all pairs of nodes.

radius The *radius* of a hypergraph is defined as the *shortest longest path* length among all pairs of nodes.

clustering coefficient (local) The hypergraph *clustering coefficient* for a pair of vertices (u, v) is defined as (see [48]),

$$cc(u, v) = \frac{|N(u) \bigcap N(v)|}{|N(u) \bigcup N(v)|} \tag{1.7}$$

where $N(x)$ denotes the set of neighboring nodes.

clustering coefficient (global) The average of local clustering coefficient over all pairs of vertices.

With characterizing functions adapted to hypergraphs, many graph-theoretic questions can be revisited in light of the hypergraph nature of metabolism. For example: Is the hypergraph degree distribution still a power law? Is the hypergraph scale-free or scale-rich [22, 23]? Will a bow-tie pattern still exist [25, 26, 27]? Are bottleneck proteins (with higher betweenness) still essential [40]? Do enzymes with higher phylogenetic profiles have a higher betweenness centrality [35]? Do metabolic hypergraphs still share hub sets with co-expression networks and interaction networks [41]? Are enzymes close on the metabolic hypergraph still close on the chromosome? [45]? Are metabolites with higher hypergraph degree subject to stronger purifying selection [38]? Is the metabolic hypergraph robust [47]? Do modularities remain correlated with the variability of the organism's living habitat [29, 30]? Are metabolites close in metabolic hypergraph also close in functionality [53]? Revisiting these question informs us whether certain evolutionary principles reported are outcomes of neutral factors, such as a representation artifact and genetic drift, or really a consequence of the organism's adaptation to the changing environment.

1.1.2 Constraint-based modeling

Another modeling framework that can be conducted on the genome scale, constraint-based modeling of metabolic network has gained popularity due to its little requirement of parameterization. Constraint-based modeling makes use of stoichiometries and thermodynamic information on metabolic reactions and assumes the steady state of the metabolites' concentration. Just like graph-based modeling, the problem formulations rely on metabolic reconstruction. Typically, the modeling seeks a solution in the space of all feasible flux distributions through convex optimization (see Varma and Palsson [54] and Feist et al. [55] for a review). The optimization goal can vary, from the maximization of a biomass accumulation formula, to the distance to a known reference flux distribution. As its name implies, the space of feasible flux distribution investigated in the modeling procedure is delineated by the constraints which stem from the steady state assumption,

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0} \tag{1.8}$$

where $\mathbf{S}$ is the stoichiometric matrix and $\mathbf{v}$ is the vector of reaction fluxes, as well as other constraints of the reaction fluxes imposed by thermodynamic properties of the reaction, educated empirical setting as well as the environmental availability of external nutrients.

Two important concepts link the structure of the space of feasible fluxes to the metabolic pathways—*elementary flux modes* (EFM) and *extreme pathways* (EP) which are subsets of the former. They are properties defined by the inherent network connectivity as well as the thermodynamic constraints. Schilling and Palsson [56] use extreme pathways to identify the minimal substrate requirement of influenza. They also suggest using the same method to detect reactions that need to be reconciled and

re-annotated. Gene co-regulation or co-expression pattern were predicted. Edwards and Palsson [57] reconstructed the stoichiometric matrix for the central metabolic network of *E.coli*. With additional capacity constraints, they showed that their model was capable of predicting the growth potential of mutated strains. With the same model but optimized for growth rate, Edwards et al. [58] showed that an experiment-consistent relationship can be generated among carbon source uptake rate, oxygen uptake rate and cellular growth rate. On the same model organism (*E.coli*), Stelling et al. [59] use EFM to infer network functionality, robustness and gene regulation. Combined with comparative genomics, Pal et al. [2] showed that metabolic networks grow by horizontal gene transfer by getting new genes and enzymes integrated at the periphery. As a quantitative measure of thermodynamic benefit, free energy change of the reaction is used to weigh flux on the corresponding edge (Holzhutter [4]). By minimizing flux in the above setting, Holzhutter re-discovered, without detailed kinetics, results from kinetic modelling and experiments.

The terminologies mentioned above are summarized below:

Feasible flux Given a stochiometric matrix $\boldsymbol{S}_{n\times m}$ of a metabolic network with thermodynamic constraints on each arc (directed edge), a *feasible flux* is a vector $\boldsymbol{e} = (e_1, e_2, \ldots, e_m)$ such that,

1. $\boldsymbol{S} \cdot \boldsymbol{e} = 0$ (mass conservation)
2. $e_i \leq 0$ if reaction i is irreversible.

Elementary flux mode An *elementary flux mode* is a feasible flux $\boldsymbol{v}$ such that no other feasible flux where the set of non-zero components is a proper subset of the set of non-zero component of $\boldsymbol{v}$.

Extreme pathway An *extreme pathway* is defined as the minimal set of EFMs that can describe all feasible fluxes by linear superposition. Every set of extreme

pathways form a *convex basis* of the set of feasible flux.

Accompanied with the development of constraint-based modeling techniques is an outburst of genome sequences, annotations and new metabolic reconstructions and models on these species mushroomed. Using genome-wide information, Reed et al. [60] reconstructed a more comprehensive metabolic network of *E.coli*. Based on the metabolic reconstruction of *S.cerevisiae* by Forster et al. [61], Famili et al. [62] built a flux-balance model which generates experiment-consistent results on growth and metabolic by-product secretion. Upon the same model, Segre et al. [3] built the epistasis interaction network, which exhibits biological modularity. So far organisms with comprehensive metabolic reconstruction as well as modeling on it include: *H. influenzae* [63], *E.coli K12 MG1655* [57, 60, 64], *H.pylori* [65, 66], *S.cerevisiae* [61, 67, 68] (see [69] for a consensus), *H.sapiens* [70, 71], *P.falciparum* [72], *H.sapiens* (mitochondria) [73, 74], *M.jannaschii* [75], *S.aureus N315* [76, 77], *L.lactis* [78], *M.musculus* [79], *S.coelicolor* [80], *M.barkeri* [1], *G.sulferreducens* [81], *L.plantarum* [82], *N.meningitidis* [83], *M.tuberculosis* [84, 85], *M.succiniciproducens* [86], *R.etli* [87], *H.salinarum* [88], *P.putida* [89], *T.Leishmania* [90], *P.aeruginosa* [91], *C.glutamicum* and [92].

1.1.3 Kinetics and dynamics

For kinetic analysis, two major classes of models exist: the ordinary differential equation (ODE) model and executable models, including Petri-nets and Boolean networks.

- ODE model

  Most of the biological implications arising from ODE models involve the control coefficient (or sensitivity coefficient), either for flux or for concentration. Using phenomenological Michaelis-Menten kinetics, Chance et al. showed inhibition mechanisms that can affect the steady state level in respiratory chain [93] and

glycolysis [94]. Similar simulation has been done by Garfinkel and Hess [95] (See [96] for a review). The dependence of the glycolytic flux on the deterioration of one enzyme is studied by looking at the flux control coefficient [97]. As an abstraction of the ODE model, qualitative reasoning has been employed in modeling and identifying of metabolic pathways by King et al. [98] and Heidtke and Schulze-Kremer [99].

- Executable models

  As one of the most studied executable models for metabolic networks, the first Petri-net representation of a metabolic pathway was introduced by V.N.Reddy [100] in 1993, when a successful application of the Petri-net onto a combined glycolytic and pentose phosphate pathway was presented. Over a decade later, Heiner and Koch [19] presented a Petri-net model validation for two other metabolic pathways, the sucrose breakdown pathway in the potato tuber (see [18] for a detailed description) and glycolysis-pentosphosphate pathway. Using a decomposition algorithm, Schuster et al. [6] showed that the concepts in Petri-net theory can be related to terms in conventional pathway structure, say, for example, the equivalence between T-invariants and elementary flux mode. By applying a colored Petri-net onto a discriminated set of metabolites in the same glycolysis-pentosephosphate pathway, Voss et al. [15] obtained further insights, in addition to qualitative analysis, by just looking at the invariants. It is worth mentioning that stochastic Petri-nets (SPN), though widely used in gene regulatory networks, have not, at least to the author's knowledge, been adapted into the field of metabolic networks.

1.1.4 Integration, control and crosstalks

In microbial organisms and other species, metabolism does not function in isolation from the other cellular networks. Transcriptional regulation and signal transduction exert crucial control over the metabolic network. The outcome of the metabolic reactions are constantly fed back to the two other cellular networks as a sensor of the energy budget of the cell. Certain metabolites are important cofactors that regulate the secondary messenger in the signaling transduction. Products of catabolism, such as nucleotides and amino acids, are fundamental building blocks for synthesizing proteins and DNA in the transcriptional processes and, hence, affect the transcriptional regulation. Due to the intimate coorperation and intensive crosstalks between metabolism and non-metabolic processes, it is highly desirable to have integrated modeling platforms which simulate metabolic and non-metabolic processes simultaneously.

Current literature has already describes a few successful integrations between metabolic network modeling and the modeling of other cellular networks. Covert *et al.* first studied the input control of metabolic networks by transcriptional regulation under the flux balance anlaysis framework by encoding the regulatory effect as time-dependent constraints [101]. Later, the same group developed the regulatory FBA, or rFBA, where regulatory constraints are more systematically represented in a Boolean network[102]. With gene expression constraints imposed by transcriptional regulation, they are able to show how the solution space of feasible flux distributions can be reduced resulting in a so-called "second generation flux balance analysis" [103]. Along the same line, Shlomi *et al.* predict the metabolic flux distribution in perturbed systems (gene knockout) by describing gene transcription in Boolean dynamics [104]. Later on, the same group published SR-FBA—a full integration of metabolism and regulation with metabolism modeled in FBA and regulation modeled in Boolean

equations [105]. Lee *et al.* published the first model, idFBA, which integrated all the three major cellular networks—metabolism, signal transduction and transcriptional regulation [106]. Recently, Chandrasekaran and Price published another regulatory-metabolism model PROM which accounts for stochasticity from the inference of the regulatory network [107].

1.2 Evolution of metabolic networks

Understanding the past is always the first step towards the prediction of the future. No matter whether the goal is to react to the acquisition and spread of antibiotic resistance in pathogens [108, 109], to rationally engineer high-yield fuel-producing bacteria [110] or to introduce certain bacteria for waste management purposes [111], one needs to know how and why metabolic evolution happens.

Apart from these practical purposes, knowledge about the evolutionary history of microbes has significant scientific value for understanding the evolution of higher organisms [112] and the progression of ecosystems [113]. The evolution of microbes bridges the pre-biotic era and the emergence of eukaryotic organisms, including human beings. Microbes occupy the widest range of habitats on earth. They contribute to both physical and chemical circulation of mass and energy, both in quantity and in quality. Studying their evolution is a critical component in understanding how the ecosystem is shaped.

The study of genome evolution is being carried out on various scales, from single nucleotide mutation to large scale genome rearrangement, and is centered around the driving forces and outcomes of these mutational events. To differentiate whether a mutational event has an adaptive role, or if it is merely the random outcome of a neutral force, a direct approach is to reconstruct these mutational events and try to reason these events from a functional point of view. In the case of metabolism, the

role of a gene in the metabolic pathway, or more generally, the metabolic network is investigated to determine from its reconstructed history whether the existence and fixation of certain mutational events is an outcome of natural selection.

In the following subsections, I want to first of all introduce the topic of using genome-scale phylogenetic methods, dubbed phylogenomics, to reconstruct the ancestral states and, hence, evolutionary history of microbial genes (not necessarily restricted to metabolic ones). I also want to outline the typical methods used in investigating the adaptive role of mutational events in metabolism.

1.2.1 Phylogenomics

Phylogenetics is the study of the tree-like genealogy of species or, more specifically, the ancestor-descendant relationship among a given set of taxa and their ancestors [114]. Phylogenomics is the extension of phylogenetics to genome-scale data [115], originally defined for the purpose of predicting functionally uncharacterized proteins based on sequence similarity [116]. Later on, the subject became more centered around the reconciliation of the species phylogeny (the genome tree) and each gene phylogeny (the gene tree) [117, 118].

It has long been appreciated that the gene tree does not necessarily agree with the species tree due to mutational events that operate on a large chunk of DNA instead of performing single nucleotide editing [117]. These mutational events include duplication [119, 120, 121, 122], loss [123, 124, 125, 126], horizontal transfer [127, 128] among individuals and other population-level phenomena, such as incomplete lineage sorting (deep coalescence). They confound the conventional tree-like concept of evolutionary history, which assumes the vertical inheritance of genetic content from ancestors to descendants.

Although many methods for reconstructing the sequence evolution history take

into consideration only single nucleotide point mutations [129, 114], many attempts have also been made towards the accurate recovering of the evolutionary history by accounting for some, if not all, types of medium-to-large scale mutational events.

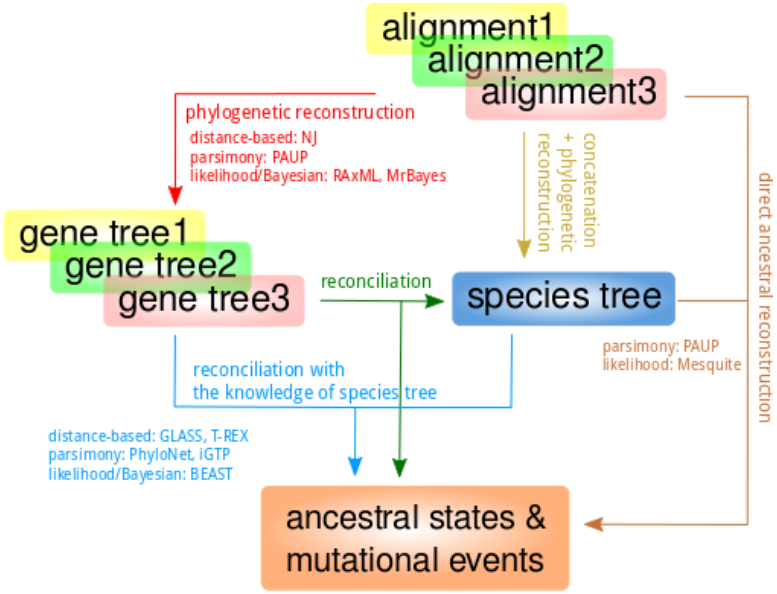

Figure 1.2: **Flowchart of phylogenomic practices.**

Fig. 1.2 is a flow chart of common practices in phylogenomic inferences. Many of the steps employ conventional phylogenetics methods, such as the gene tree inference.

There are two major classes of methods for the reconstruction of medium-to-large-scale mutational events—the direct approach and the tree-reconciliation-based approach. In the direct approach, ancestral states are reconstructed without first seeking the phylogeny of the DNA segment of interest (such as the gene trees). The method that reconstructs the ancestral states directly from the gene family composition are algorithmically no different from the methods for reconstructing ancestral sequences. The only difference is in the alphabet used for coding objects such as genes and proteins.

In the tree-reconciliation-based approach, mutational events are recovered from reconciling gene trees built from sequence alignments using conventional phylogenetic

reconstruction methods. The reconciliation is performed with or without the knowledge of the species phylogeny. The species phylogeny can also be reconstructed by reconciling these gene trees. Just as point mutations can be reconstructed under different frameworks—such as the distance-based method, parsimony-based method and likelihood/Bayesian-based method—gene tree reconciliation can also be implemented under these frameworks. Table 1.1 tabulates some of the recent tools for tree-reconciliation-based phylogenomic inference.

framework	ILS	Duplication / Loss	HGT
distance	iGLASS [130] GLASS [131, 132] STAR / STEAC [133]		T-REX [134]
parsimony	PhyloNet [135]	MPR [136] Notung [137] RAP [138] DupTree [139] iGTP [140]	RIATA-HGT [141]
likelihood	COAL [142] STELLS [143] STEM [144]	[145]	[146]
Bayesian	BEST [147] BEAST [148] BUCKy [149]		

Table 1.1: **Phylogenomic methods for species tree reconstruction and gene tree reconciliation.** ILS: Incomplete Lineage Sorting; HGT: Horizontal Gene Transfer; Other methods include: majority consensus, democratic vote and concatenation.

1.2.2 Adaptive role of mutational events

Two major classes of explanations, i.e., the neutral theory and adaptive theory, exist to interpret the reconstructed evolutionary history based on observed genetic composition of extant taxa. These explanations interpret the genetic configuration in extant taxa as a result of (1) neutral evolution implemented using these mutational events; (2) the adaptive evolution, mainly in the form of natural selection; or (3) the interac-

tion of the two. In the field of microbiology horizontal gene transfer, duplication and loss are the medium-to-large-scale mutational events most focused on. Understanding the proportion to which a certain mutational event confers a neutral, beneficial or deleterious consequence helps us understand the mechanism for the spread of genetic content, as well as phenotypes, such as pathogenicity, in the microbial community.

The neutral theory states that certain observations, such as the gene frequency distribution in the pan-genome [150], are results of neutral processes instead of an outcome of adaptive evolution. For example, according to a nearly neutral theory of horizontal gene transfer [151], transferred genes are neutral or nearly neutral to the recipient instead of carrying a beneficial trait as is traditionally perceived. Daubin and Ochman showed that newly acquired genes, or ORFans, are under (at most) weak purifying selection [152]. This supports the nearly-neutral evolution of horizontal gene transfer. Other attempts explain the mutational events from the perspective of the living environment, but still from a neutral point of view. For example, Abby et al. [153] found that bacteria living in a protected niche have less chance of receiving or donating exogenous genes. The rate of horizontal gene transfer is closely associated with the microbe's living habitat.

Other researchers seek adaptive explanations to mutational events found by investigating whether the mutation is an outcome of purifying or directional selection. For example, the *Complexity Hypothesis* states that genes that are part of a complex system are more resistant to transfer [154] because their transfer can barely confer any beneficial traits without the co-transfer of the other components of the system. In other words, genes that are not part of any complex systems have a better chance of being harbored and fixed in a population. Evidence has been reported supporting this hypothesis [153]. The *Selfish Operon Theory* states that physical proximity of genes organized in an operon facilitates the co-transfer of all genes required for a

weakly selectable trait, which might be responsible for the spread of pathogenicity [155]. Contradicting evidence on the operon structure of essential gene, which are under strong purifying selection, has also been reported [156]. Bacteria leading a symbiotic lifestyle have experienced tremendous amount of gene loss because unnecessary genes which only incur overhead of maintenance were lost during evolutionary history [123, 157, 158, 124].

Since the two classes of explanation are not mutually exclusive, there are models that apply both. A famous example is Ohno's *Neofunctionalization* model for gene duplication [119], where the initial fixation of the duplicated copy is entirely neutral, but new copies may bring beneficial traits due to their being relieved from purifying selection (and thus could circumvent local maxima on the fitness landscape). Genes with multiple functions could have higher tendency of being duplicated in the light of separately specializing the duplicates for each of the functions (*Subfunctionalization*) [159].

1.3 Contributions of the thesis

Standard graphs, where each edge links two nodes, have been extensively used to represent the connectivity of metabolic networks. It is based on this representation that properties of metabolic networks, such as hierarchical and small-world structures, have been elucidated and null models have been proposed to derive biological organization hypotheses. However, these graphs provide a simplistic model of a metabolic network's connectivity map, since metabolic reactions often involve more than two reactants. In other words, this map is better represented as a hypergraph. Consequently, a question that naturally arises in this context is whether these properties truly reflect biological organization or are merely an artifact of the representation.

In this dissertation, I address this question by re-analyzing topological proper-

ties of the metabolic network of *E. coli* under a hypergraph representation, as well as standard graph abstractions. I find that when clustering is properly defined for hypergraphs and subsequently used to analyze metabolic networks, the scaling of clustering, and thus the hierarchical structure hypothesis in metabolic networks, become unsupported. Moreover, I find that incorporating the distribution of reaction sizes into the null model further weakens support for the scaling patterns. These results combined suggest that the reported scaling of the clustering coefficients in the metabolic graph and its specific power coefficient may be an artifact of the graph representation, and may not be supported when biochemical reactions are atomically treated as hyperedges. This study highlights the implications of the way a biological system is represented and the null model employed on the elucidated properties, along with their support, of the system.

A common practice of assembling a metabolic graph is to prune (i.e., remove nodes and edges) the metabolic graph prior to any analysis in order to eliminate confounding signals from the representation. Currently, this pruning process is carried out in an *ad hoc* fashion, resulting in discrepancies and ambiguities across studies. I propose a biochemically informative criterion, the *strength of chemical linkage* (SCL), for a systematic pruning of metabolic graphs. By analyzing the metabolic graph of *E. coli*, I show that thresholding SCL is powerful for selecting the conventional pathways' connectivity out of the raw network connectivity when the network is restricted to the reactions collected from these pathways. Further, I argue that the root of ambiguity in pruning metabolic graphs is in the continuity of the amount of chemical content that can be conserved in reaction transformation patterns. Finally, I demonstrate how biochemical pathways can be inferred efficiently if the search procedure is guided by SCL.

In order to calculate SCL and perform other fine analysis on the atom scale,

we need the knowledge of atom mapping for each metabolic reaction. Atom tracing provides valuable information in many analyses of metabolic networks including pathway inference and flux estimation. Symmetries—mapping operations that produce atom equivalencies—introduce alternative tracings when multiple atom mappings are aggregated. Although several attempts have been made to consider symmetry while curating atom mappings, a definition of the symmetry amenable to automated computation and a systematic quantification of the extent of symmetries in both compounds and reactions is still lacking. Moreover, the impact of symmetries on the calculation of the atom economy of pathways and the simulation of isotopomer distribution is yet to be assessed. In this study, I formulate the symmetries of both compounds and reactions as automorphic mappings of the corresponding graph representations. I investigate the extent of both compound and reaction symmetries in several metabolic systems. I find, through random walking in the metabolic network of *E.coli*, that alternative tracings originated from symmetries could give rise to a considerable amount of differential conservation of atoms and distinct transition patterns of the isotopomer distribution.

It has been reported that the modularity of bacterial metabolic networks of is closely related to the variability of their living habitats. However, given the dependency of the modularity score on the community structure, it remains unknown whether organisms achieve certain modularity via similar or different community structures.

In this dissertation, I study the relationship between similarities in modularity scores and in community structures of the metabolic networks of 1021 species. Both similarities are then compared against the genetic distances. I revisit the association between modularity and variability of the microbial living environments and extend the analysis to other aspects of their life style, such as temperature and oxygen re-

quirements. I also test both topological and biological intuition of the community structures identified and investigat the extent of their conservation with respect to the taxonomy. I find that similar modularities are realized by different community structures and that the convergent evolution of modularity is closely associated with the number of (distinct) enzymes in the organism's metabolome, a consequence of different life styles of the species. I find that the order of modularity is the same as the order of the number of the enzymes under the classification based on the temperature preference but not on the oxygen requirement. Additionally, inspection of modularity-based communities reveals that these communities are graph-theoretically meaningful, yet not reflective of specific biological functions. From an evolutionary perspective, I find that the community structures are conserved only at the level of kingdoms. My results call for more investigation into the interplay between evolution and modularity: how evolution shapes modularity, and how modularity affects evolution (mainly in terms of fitness and evolvability). Further, my results call for exploring new measures of modularity and network communities that better correspond to functional categorizations.

The role of evolutionary events, neutral and adaptive alike, in shaping microbial metabolic pathways and networks has been extensively studied. In particular, as horizontal gene transfer (HGT) seems to be ubiquitous in bacteria, its role in metabolic innovation and distinctness from other mutational events has been investigated. These existing studies are mostly based on the analysis of *gene families*, which are groups of gene orthologs and paralogs collected from a set of species. That is, these studies assume that (protein-coding) genes are the basic unit of evolution. However, recent studies have highlighted the large extent of gene fission and fusion events in bacterial genomes, thus challenging the concept of a *gene tree* and bringing into question the notion of a gene family and how to infer it. A phylogenomic study

that targets mutational events which act at the sub-gene level and their implications on microbial metabolic genes and pathways is lacking. In particular, it is not known what, if any, effects sub-gene mutational events have on metabolic networks and their innovation. In this dissertation, I conduct an extensive phylogenomic analysis of a proteobacterial data set, catalog all of the mutational events that take place, including HGT and fission/fusion, and study their effect on metabolic networks and organismal fitness via metabolic flux analyses. I introduce the notion of *modules*—the longest gene fragments which are conserved in all species under investigation—as a more appropriate alternative to gene families. These modules are detected and clustered into families based on sequence similarity. Further, I define the *mosaicity* of genes and pathways and quantify them for the proteobacterial data set. I find that DNA-binding, ion-binding proteins, as well as those related to transposon, are more mosaic. I conduct a systematic investigation of the mosaicity of metabolic pathways, defined as the number of modules from all the genes involved in the pathway. I find mosaicity to be closely related to the living habitat of an organism and that metabolic genes are more mosaic than non-metabolic ones. My study emphasizes the significance of ancestral reconstruction of mutational events at the sub-gene scale and provides important insights into the emergence of mosaic genes and metabolic pathway evolution.

Chapter 2

Graph and hypergraph modeling of metabolism

2.1 Introduction

Graphs have been used extensively to model the connectivity of cellular processes [160], including metabolic networks [161]. Once represented as a graph, a wide array of tools can be applied to visualize and analyze the graph to elucidate properties of the corresponding cellular network [162, 161]. Analyses of metabolic networks based on the graph representation have revealed a wide range of significant properties of the network connectivity, including a short mean path length [20], a scale-free degree distribution [21] and a bow-tie structure [26]. The statistical significance of such findings, and whether these graph features have been subject to adaptive evolution, are often assessed by comparing biological networks to networks generated under null models. In this context, null models produce random (standard) graphs that are constrained to satisfy one or more requirements, such as an expected degree distribution. However, in metabolic networks, a reaction often involves more than two reactants, rendering standard graphs too simplistic and consequently requiring a certain abstraction. For example, one commonly used technique for enabling a graph representation of a metabolic network's connectivity map is to model each reaction by a complete

subgraph, where each pair of reactants on both sides of the reaction are linked by an edge. Analyses based on different representations of the metabolic network of *E. coli* have revealed conflicting patterns related to its small-worldness [22, 21, 163]. It is therefore natural to ask whether these properties, that are elucidated based on a standard graph representation and a null model, truly reflect biological organization or are merely an artifact of the representation.

To investigate this question, we analyze metabolic network connectivity maps from a *hypergraph* perspective. Given that metabolic reactions may involve more than two reactants, hypergraphs—where an edge connects any finite number of nodes—provide a more realistic model of the connectivity of a metabolic network. Indeed, Klamt et al.[48] recently argued that any metabolic (standard) graph representation fails to describe the dependence of a metabolite on others that participate in the same reaction. They illustrated that even a bipartite graph, with metabolites and reactions being the two node types, fails to remedy the problem [48] as links in bipartite graphs still remain independent. Further, Lacroix et al. [161] suggested that each reaction has to be taken as a whole (yet did not specify how to analyze such data). To properly represent reactions that involve more than two entities, hypergraphs (see [51, 52] for introductory texts on hypergraphs) are the natural representation of metabolic networks' connectivity maps (e.g., see [48]). A generalization of standard graphs, a hypergraph allows any subset of two or more nodes to form an edge, called a *hyperedge*. Further, to distinguish between the metabolites on different sides of a metabolic reaction, and to allow for the designation of the reaction direction, the set of nodes connected by a hyperedge can be bipartitioned into the *head set* and the *tail set*. Standard graph representation of a metabolic network connectivity is in fact a transformation of the underlying hypergraph. The *substance model* (every pair of substances/metabolites participating in the same reaction are connected by an edge),

substrate-substrate model (every pair of metabolites on the same side of a reaction are connected by an edge), and *substrate-product* model (every pair of metabolites on opposite sides of a reaction are connected by an edge), discussed in [164], correspond to the *primal*, *cis-primal*, and *trans-primal*, respectively, of the underlying hypergraph. These transformations on hypergraphs are formally defined in the Methods section below, and are illustrated in Fig. 2.1.

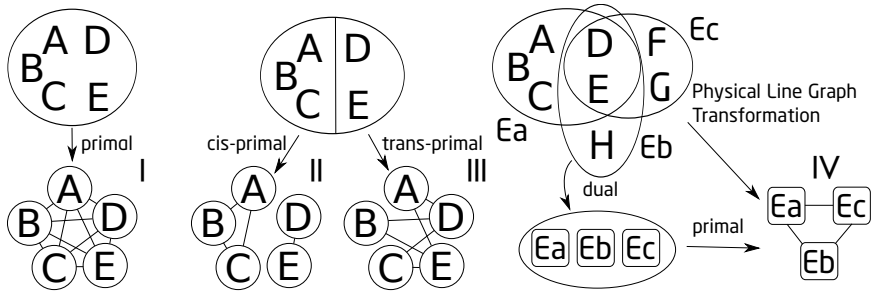

Figure 2.1: **Illustration of the hypergraph transformations and abstractions.** Left: a hyperedge is turned into a complete graph linking every pair of nodes to obtain the primal graph (I). Middle: the *cis-* (II) and *trans*-primal (III) graphs are obtained by connecting either nodes in the same side of the hyperedge partition or on different sides, respectively. Right: The physical line graph transformation (IV) can be obtained by taking the primal of the dual of the hypergraph; that is, it is the composition of two transformations.

Some work on metabolic connectivity hypergraphs already exists. For example, Forst et al. [165] used algebraic operations to compare metabolic hypergraphs across multiple species for phylogenetic reconstruction. A directed hypergraph-based tool, Rahnuma, has been developed recently for metabolic pathway analysis [166]. An algorithm for computing the *minimal cutting set* on hypergraphs was proposed [167].

Further, it is worth pointing out that the hypergraph property of the dependence among metabolites participating in the same reaction has already been widely, though implicitly, captured in other modeling techniques, such as network expansion [168], reachability analysis [169], constraint-based modeling [170] and Petri-net modeling [16]. For example, the stoichiometric matrix used in the constraint-based

modeling is essentially a weighted incidence matrix of the underlying hypergraph (where each column corresponds to a hyperedge). This again reflects the rather natural view that metabolic network connectivity maps are inherently hypergraph-like. Nonetheless, with the exception of these very few studies, most analyses of metabolic networks' connectivity maps in the literature are based on (standard) graph representations. This lack of adoption of hypergraphs may owe to a host of factors. One of them is the inherent difficulty in visualization [171]. Obtaining an informative hypergraph layout is much more involved than obtaining a standard graph layout (see [172] for a typical algorithm for drawing hypergraphs under the *subset standard*). Besides, many problems that can be solved efficiently on standard graphs become NP-hard on hypergraphs (e.g., the problem of finding the shortest-path in a hypergraph with hyperedges weighted by their cardinalities [49]). Finally, the lack of well-defined hypergraph counterparts to the common standard graph characteristics, such as clustering coefficients, may have made their use less appealing.

In this chapter, we address the aforementioned question by conducting three tasks on the metabolic network connectivity map of *Escherichia coli*. First, we analyze the scaling of degree distributions [20, 22] and average local clustering coefficients [173] on various standard graph abstractions. While a host of topological properties can be analyzed, we focus on these properties since they are central to the two aforementioned hypotheses about metabolic networks. Then, we show how these analyses are affected when the null model incorporates the reaction size (hyperedge cardinality)—a quantity that, to the best of our knowledge, is ignored in existing studies. Finally, we devise measures of local and global clustering coefficients that apply directly to hypergraphs and differ from those of Estrada and Rodríguez-Velázquez [174] in their satisfaction of desired properties. Based on these three tasks we find that a null model that incorporates the hyperedge cardinalities changes the analysis results significantly

compared to the previously used null models.

Further, when clustering is analyzed directly on the hypergraph representation, the scaling property, which has been reported in the literature, becomes poorly supported. These results combined suggest that the reported scaling of the clustering coefficients in the metabolic graphs and its specific power coefficient may be an artifact of the graph structure produced by the abstraction process and may not be supported when biochemical reactions are atomically treated as hyperedges. This study highlights the implications of the systems representation and null model employed in an analysis on the hypotheses derived for that system. Further, these results have implications beyond metabolic networks since, for example, signal transduction networks contain many enzymatic and complexing reactions that form hyperedges. The weakening of statistical support of reported properties of biological networks when the new null model is considered calls into question claims that adaptive evolution is the (only) explanation for the emergence of complex, or non-intuitive, network features. More generally, this study further emphasizes the issue that the use of proper representations and null models is fundamental to understanding the biology underlying the abstract model.

2.2 Hypergraph representation of metabolism

2.2.1 Formal definition of metabolic hypergraphs

An *undirected hypergraph* H is an ordered pair $(V, \mathcal{E})$, where V is the set of nodes and $\mathcal{E}$ is the set of hyperedges. Each hyperedge $E \in \mathcal{E}$ connects, or corresponds to, a subset $V' \subseteq V$, where $|V'| \geq 2$. Hypergraphs are a natural model of the connectivity of metabolic networks. For example, to model the metabolic reaction $A + B \rightleftharpoons C + D$ as an undirected hypergraph, we take $V = \{A, B, C, D\}$, and $\mathcal{E} = \{E\}$, where $E = V$.

To distinguish between the two sets of metabolites on opposite sides of a reaction, a hyperedge E can be further bipartitioned into two subsets E_t, the *tail set*, and E_h, the *head set*. In this case, we write E as the ordered pair (E_t, E_h), and the direction of the edge is, by convention, from the tail set to the head set. Using this notation of directed hyperedges, a *directed hypergraph* is defined. For example, the hyperedge corresponding to the irreversible reaction $A + B \rightarrow C + D$ is the ordered pair $E = (\{A, B\}, \{C, D\})$.

The *degree* $d(v)$ of a node $v \in V$ in a hypergraph is defined as the cardinality of the set $\{E \in \mathcal{E} \mid v \in E\}$. The *neighborhood* of a node v, denoted by $N(v)$, in a hypergraph is defined as the node v itself together with the set of all nodes connected to it by a hyperedge. More formally,

$$\begin{aligned} N(v) &= \{v\} \cup \{u \in V \mid \{u, v\} \subseteq E \\ &\qquad \text{for some } E \in \mathcal{E}\}. \end{aligned} \tag{2.1}$$

The neighborhood of a set of nodes, U, is defined as the union of the neighborhoods of all nodes in U, or $\bigcup_{v \in U} N(v)$. Further, we denote by $\mathcal{M}(v)$ the set of hyperedges of which v is an element, that is, $\mathcal{M}(v) = \{E \in \mathcal{E} \mid v \in E\}$.

2.2.2 From hypergraphs to standard graphs

A variety of transformations can be applied to a hypergraph to obtain standard graph representations. We now define transformations that are applicable (and have been applied) in the context of representing metabolic networks. Let $H = (V, \mathcal{E})$ be a hypergraph. The *primal* of H is a (standard) graph $G_p = (V, E_p)$, where every two nodes in V that are connected by a hyperedge in H are connected by an edge in G.

In other words,

$$E_p = \{\{u, v\} \mid \{u, v\} \subseteq E \text{ for some } E \in \mathcal{E}\}.$$

The primal of a metabolic hypergraph is also called the *substance model* [164], since every pair of substances (metabolites) participating in the same reaction are connected by an edge (i.e., form a *clique*). For directed hypergraphs, primal graphs can be defined in two ways. The *cis*-primal is obtained by connecting with an edge every pair of nodes within the same partition of the hyperedge (both nodes from the head set or both from the tail set). In other words, the *cis*-primal of H is a graph $G_{cp} = (V, E_{cp})$, where

$$E_{cp} = \{\{u, v\} \mid \{u, v\} \subseteq E_t \text{ or } \{u, v\} \subseteq E_h \text{ for some } (E_t, E_h) \in \mathcal{E}\}. \tag{2.2}$$

This corresponds to the *substrate-substrate model* [164], where metabolites on the same side of a reaction are connected. The *trans*-primal is obtained by connecting with an edge every pair of nodes that belong to two different parts of a hyperedge (one from head set and the other from the tail set). In other words, the *trans*-primal of H is a graph $G_{tp} = (V, E_{tp})$, where

$$E_{tp} = \{\{u, v\} \mid u \in E_t \text{ and } v \in E_h \text{ for some } (E_t, E_h) \in \mathcal{E}\}. \tag{2.3}$$

This corresponds to the *substrate-product model* [164], where metabolites on opposite sides of a reaction are connected. Fig. 2.1 illustrates these three transformations.

Every undirected hypergraph can be completely described by a binary matrix M, called the *incidence matrix*, where columns correspond to hyperedges and rows to nodes. An entry $M[i, j] = 1$ denotes that node i is an element of hyperedge j

while an entry $M[i, j] = 0$ denotes otherwise (Notice that a stoichiometric matrix is a weighted incidence matrix of a metabolic network's connectivity map.). A binary matrix is a valid incidence matrix if and only if every row and column contains at least one 1. Thus, the transpose of the incidence matrix of any hypergraph is also a valid incidence matrix. The transpose of the incidence matrix of a hypergraph H corresponds to the *dual hypergraph* H'. The common practice of creating a reaction graph by connecting two reactions if they share a reactant [164] (also known as the *physical line graph transformation* [175], or PLGT for short hereafter) amounts to first computing the dual of the original metabolic hypergraph, and then taking the primal of the resulting hypergraph (see Fig. 2.1).

Finally, common set operations, such as union and intersection, can also be introduced into the hypergraph transformation. One of the widely, yet implicitly, used case is the generation of enzyme/gene hypergraphs from the underlying reaction hypergraph [161]. Each hyperedge in the transformed hypergraph is the union of all hyperedges corresponding to reactions that are catalyzed by some particular enzymes/genes. This process is equivalent to resampling a number of subsets of the set of all hyperedges. Note that unlike reaction hyperedges, these hyperedges may substantially overlap or even coincide with each other (when multiple enzymes/genes catalyze a same set of reactions).

2.2.2.1 Data

We assembled the metabolic hypergraph of *Escherichia coli* using the KEGG database [176]. The presence of a reaction was inferred based on whether there is a gene that is annotated to generate any enzyme that catalyzes the reaction. Reaction formulas, enzyme identities and gene annotations were downloaded from KEGG. We recognize that the metabolic networks thus constructed may not provide a complete coverage

of the entire metabolic system in *E. coli*. However, this is a common way of constructing metabolic networks in existing studies. Further, since our study is aimed at the differences in properties elucidated from different representations of the same system, a complete coverage, while desirable, is not a necessary prerequisite. The undirected hypergraph representation is obtained by putting all the metabolites in each reaction into a single hyperedge. The directed hypergraph representation is obtained by further separating the metabolites on opposite sides of the reaction into the tail and head sets, respectively. Reaction direction is not considered in this study. Finally, we derived standard graph representations based on transformation operations on hypergraphs that amount to commonly adopted representations in existing studies. In particular, we considered the *substance model*, the *substrate-substrate model* and the *substrate-product model*, which correspond to the *primal*, *cis-primal* and *trans-primal* of a hypergraph, respectively. Further, we considered reaction graphs, where nodes correspond to reactions, and two nodes are connected if their reactions share any reactants; this corresponds to the PLGT of a hypergraph. The hypergraph data and the original reaction lists are available from the author's website: `www.cs.rice.edu/~wz4/metabolic_hypergraph.tgz`.

2.3 Metabolic reaction size and null model reconstruction

2.3.1 A binomial distribution of reaction sizes and its effects

When transforming a hypergraph into a standard graph, under any of the aforementioned transformations, the information on the hyperedge cardinality is lost. The question, then, is whether ignoring the hyperedge cardinality distribution affects the properties elucidated from abstracted standard graphs. Further, if the answer is posi-

tive, how should this information be integrated into null models of generating random metabolic graphs in analytical studies.

To address the first question, we begin by inspecting the degree distributions of primal graphs generated randomly in a way to account for hyperedge constraints. It is analytically very hard to establish the degree distribution of the primal of randomly generated hypergraphs, since the overlap between hyperedges creates dependencies among the degrees of the nodes. Therefore, we study this issue in simulations. Given a metabolic hypergraph $H = (V, \mathcal{E})$, where $|\mathcal{E}| = m$ and the maximum cardinality of any hyperedge $E \in \mathcal{E}$ is k, the primal of H has ℓ edges, where $\ell \leq m \cdot k(k-1)/2$. One method for generating random (standard) graphs in this context, while accounting for a fixed hyperedge cardinality k is to use m as the constraint; i.e., generate a hypergraph with m hyperedges, each of cardinality k, and compute its primal. In other words, a hyperedge of cardinality k is generated by randomly sampling (without repeats) a subset of k nodes and connecting them by a hyperedge, and the process is repeated m times (another method is to generate "enough" hyperedges, each of cardinality k, in the hypergraph to yield (approximately) ℓ edges in its primal;).

In the case of the *E. coli* metabolic network, the hypergraph has $n = 1193$ nodes and $m = 1168$ hyperedges, and its primal has $\ell = 5718$ edges. For each combination of n, m, ℓ and hyperedge cardinality $k \in \{2, 3, 4, 5\}$, we generated 300 random (standard) graphs based on the above method, and plotted the median degree distributions of these graphs, along with that of the primal of the metabolic hypergraph of *E. coli*. The results are shown in Fig. 2.2, where the four panels, from left to right, correspond to fixed hyperedge cardinalities of $k = 2, 3, 4, 5$, respectively.

Notice that hypergraphs with different hyperedge cardinalities give rise to standard graphs with different degree distributions. In general, the degree distribution of the primal of a random undirected hypergraph with hyperedge cardinality larger

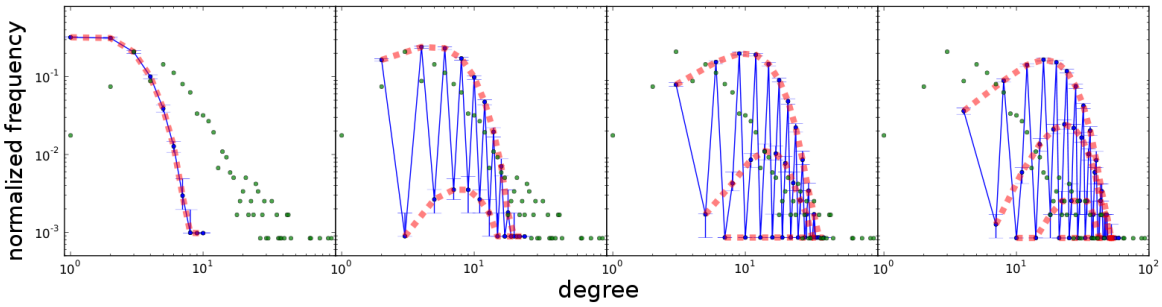

Figure 2.2: **The degree distributions of the primal graphs of random hypergraphs.** Each of the hypergraphs has 1193 nodes and 1168 hyperedges. Columns from left to right correspond to fixed hyperedge cardinalities of 2, 3, 4, and 5, respectively. The results in each panel are based on the 300 randomly generated hypergraphs (replica). For each well represented degree value (contained in at least 10 replica), the median is plotted. Error bars indicate quartiles. Green dots correspond to the degree distribution of the primal graph of the (undirected) metabolic hypergraph of *E. coli*. All plots are on log-log scales.

than 2 has a zig-zag shape when the degree value is low and becomes more complex as the degree value increases. This is due to the fact that the metabolic hypergraphs we consider are very sparse.

In a hypergraph with n nodes, the maximum number of distinct hyperedges of cardinality k, for $2 \leq k \leq n$, is $\binom{n}{k}$. And, if we exclude the trivial hyperedges (those that have a single node or the entire set of nodes), the maximum number of distinct hyperedges is

$$\sum_{k=2}^{n-1} \binom{n}{k} = 2^n - n - 1.$$

In the case of the *E. coli* metabolic network, we have 1168 hyperedges on a set of 1193 nodes. Even if we consider only standard edges (hyperedges of cardinality 2), this hypergraph is very sparse, since the maximum number of distinct hyperedges of cardinality 2 is $1193 * 1192/2 = 711028$ which is $\gg 1168$.

Now, consider a node v that is included in only two hyperedges each of which is of cardinality k. If the hypergraph is sparse, the probability that the two hyperedges would share nodes besides v is very low. Therefore, the primal of this hypergraph is

more likely to have node v with degree $2k-2$ than with degree in between k to $2k-3$. In other words, since each hyperedge contributes $k-1$ to the degree of each of its nodes in the primal, more nodes with degrees at integer folds of $k-1$ are observed if the underlying hypergraph is sparse (when contributions from different hyperedges have less chance to overlap). Hence, it might be visually desirable to classify the degree values into $k-1$ equivalence classes by $d_1 \equiv d_2 \pmod{k-1}$ ("mod" denotes the modulo operation) and connect data inside each equivalence class (dashed bold lines in Fig. 2.2).

Clearly, the hypergraphs of different hyperedge cardinalities contribute to different but overlapping ranges of degree values. In particular, the leftmost panel of Fig. 2.2 corresponds to the binomial degree distribution of random Erdös-Rényi graphs [177] with 1168 edges and probability $p = 1168/711028 \approx 0.001$ of linking two randomly chosen nodes by an edge. The degree distribution of the primal of metabolic hypergraphs is a mixture of degree distribution obtained based on different hyperedge cardinalities.

Indeed, in the case of metabolic hypergraphs, neither do all the hyperedge cardinalities take one same value nor do they follow a simple uniform distribution. Their effect on the properties of the abstracted standard graphs has not been studied. In Fig. 2.3 we plot the hyperedge cardinality distribution of the *E. coli* metabolic hypergraph. The mean value of the distribution is 4.19 and the range is roughly from 2 to 10. A comparison to Poisson and binomial distributions show that the shape is narrower than a Poisson distribution with the same mean and is much closer to a binomial distribution with sample size of 5.

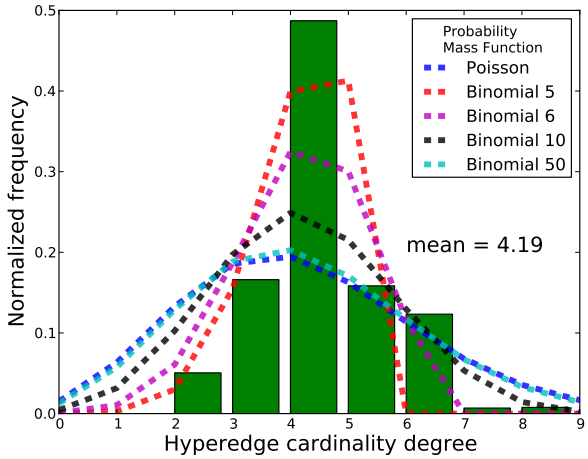

Figure 2.3: **The hyperedge cardinality distribution of the metabolic hypergraph of *E. coli*.** Poisson distribution and Binomial distribution with different sample sizes are shown in dashed lines. Parameters of these distributions (μ for Poisson and p for binomial) are chosen such that their means equal the actual value (4.19).

2.3.2 Incorporating the reaction size distribution into a null model

Based on the above results, we believe it is important for a null model for generating random graphs in the context of metabolic networks to use both the number and cardinality distribution of hyperedges. We study a null model where a random graph is generated from the metabolic hypergraph by first rewiring the hypergraph (thus, keeping the number and cardinality distribution of hyperedges unchanged) and then abstracting the random hypergraph (through a *trans*-primal transformation) into a standard graph. We compare the degree distribution of the real metabolic graph against the new null model and another null model that rewires the metabolic standard graph (also through a *trans*-primal transformation from the metabolic hypergraph) directly (see Fig. 2.4 for an illustration of the generation of the null models on a toy hypergraph).

Notice that this wiring process does not guarantee that the generated random

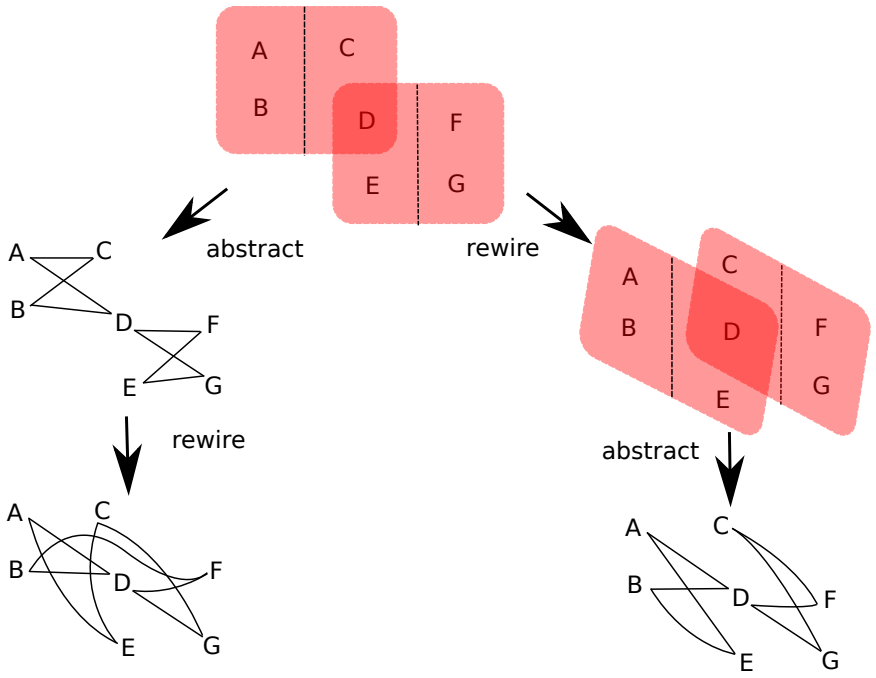

Figure 2.4: **Comparison of the two null models on a toy hypergraph.** The hypergraph-graph abstraction follows the *trans*-primal procedure. Left: Traditional way of the null model generation by first abstracting the hypergraph and then rewiring. Right: New way of null model generation, maintaining the reaction size distribution, by first rewiring the hypergraph and then abstracting it. Note that the rewiring process does not necessarily preserve the overlapping reactants (e.g., reactant D in the toy hypergraph).

networks are mass balanced; this is a very important constraint, but integrating it into a random network generation procedure is beyond the scope of this chapter.

To rewire the metabolic standard graph of *E. coli*, we perform 20,000 operations each of which randomly removes an edge and links a new pair of previously unconnected nodes. Similarly, to rewire the metabolic hypergraph of *E. coli*, we perform 20,000 operations each of which randomly removes a hyperedge, resamples a new set of nodes of the same size (same size for the tail set and the head set if a directed hypergraph is concerned), and connects the new set with a hyperedge. In this way, we keep the number and cardinality distribution of hyperedges unchanged along the rewiring process. Further, we make sure that the same set of nodes is not selected

more than once, to keep all hyperedges distinct. Finally, to obtain statistically significant results, we generate 200 random networks, each of which is rewired in both ways as above 20,000 times.

The degree distributions of the *trans*-primal of *E. coli*'s metabolic network and the random networks generated by the two rewiring procedures are shown in Fig. 2.5. Each data point and its error-bar indicate the median, 5-th and 95-th percentiles, respectively. Since not all the degrees are well represented in all 200 replicas, we plot results only for degree values present in at least 10 replicas.

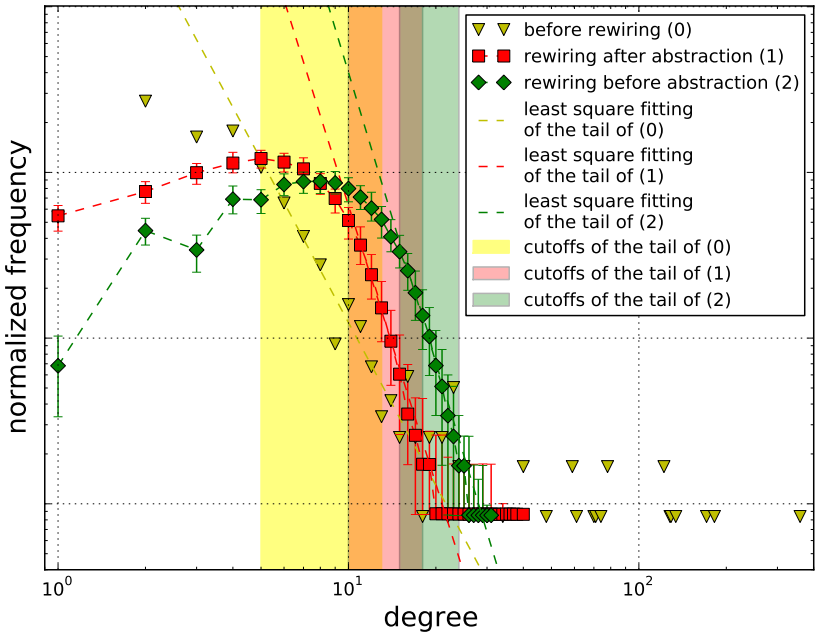

Figure 2.5: **Comparison of the degree distributions of the metabolic standard graph of *E. coli* against two different null models.** The degree distributions are derived based on three versions of the metabolic hypergraph of *E. coli*: The *trans*-primal of the hypergraph (0), the rewired *trans*-primal of the hypergraph (1), and the *trans*-primal of the rewired hypergraph (2). Least-squares fitting of the tail of (0) and the medians of (1) and (2) to $p(k) = \beta k^{-\alpha}$ yields values of $\alpha = 3.26$ for (0), $\alpha = 5.59$ for (1), and $\alpha = 5.82$ for (2). For degree $10^0 = 1$, the point for (0) coincides with that for (1).

We also fit the tail of the degree distribution of the standard graph of *E. coli* and the median of the rewired graphs to $p(k) = \beta k^{-\alpha}$ using the least squares fitting. By

inspecting the data, the fitting region for standard graphs is manually set to $[5, 13]$ (shaded region in Fig. 2.5). For rewired graphs, the end of the fitting region is defined as the smallest degree at which the 95-th percentile is higher than the frequency at count 1 (in other words, 95% of the replicas have more than one nodes with this degree). The start of the fitting region is determined by finding the first pair of neighboring degrees with slope in medians below a certain threshold (4.0) as one moves from the end of the fitting region to degree 1. We set our fitting region as such since (1) existing studies have focused on fitting degree distributions excluding their heads, for detecting "scale-freeness" [178], and (2) real-world degree distributions are always constrained by the fact that the frequency has to be no smaller than the one corresponding to count 1 (since 0 is invalid on a log-log plot).

Two observations are in order based on Fig. 2.5:

1. The tail shifts to the higher degree region in the graphs abstracted after rewiring the metabolic hypergraph compared with the graphs rewired after being abstracted from the real metabolic hypergraph. Comparison with similar situation in undirected hypergraphs (Fig. 2.2) indicates contribution from higher-order hyperedge cardinality.

2. The *trans*-primal of the rewired hypergraph preserves the zig-zag pattern in the low-degree region of the distribution (the head). The rewired *trans*-primal graphs, on the other hand, lose such shape in its "head". This indicates that the zig-zag pattern in the low degree region of the original degree distribution is due to abstracting the hypergraph with a certain hyperedge cardinality distribution into a standard graph.

These two observations are in agreement with the statement of Wagner and Fell [21] that "k-regular random graphs would be particularly poor statistical models of metabolic networks." However, our observations challenge the use of such a random

model for a statistical definition of 'key metabolites'. In particular, the *trans*-primal graphs of repeatedly rewired hypergraphs have a degree distribution whose tail is power-law (just like metabolic networks) and whose head is a zig-zag shape (again, just like metabolic networks). This raises the possibility that while adaptive forces may have shaped the cellular metabolism, neutral evolution forces (mutation, recombination, and random genetic drift) may have defined a large part of the network connectivity. This is in agreement with the observations of Lynch [179] and Wagner [180].

2.4 Origin of the scaling of clustering coefficients

It has been proposed that metabolic graphs are *hierarchical* (e.g., [53]), which can be characterized by the scaling of the average clustering coefficient $C(k)$ of nodes with certain degree k, against k. For example, Ravasz *et al.* found that $C(k) \propto k^{-1}$ for a variety of metabolic networks, including that of *E. coli* [53]. Further, they hypothesized that such a hierarchical structure corresponds to functional organization of the metabolic system. The question we investigate is whether the scaling of clustering of the average clustering coefficient is statistically supported when using a null model that incorporates the reaction size (hyperedge cardinality) distribution.

In Fig. 2.6, we show average clustering coefficient as a function of node degrees, $C(k)$, for four types of graphs:

(I) The primal of the *E. coli* hypergraph (1193 nodes and 5719 edges).

(II) Erdös-Rényi random graphs with 1193 nodes and 5719 randomly chosen edges.

(III) Random graphs generated by 100,000 rewiring operations applied to the graph in (I), where in each rewiring operation, a pair of non-adjacent edges are selected, and the neighbors of an endpoint of one edge are swapped with the

neighbors of an endpoint of the other edge. This procedure generates random graphs with the same degree distribution as that of the graph in (I).

(IV) The primal of hypergraphs generated by 100,000 rewiring operations applied to the *E. coli* metabolic hypergraph (the same method used in the previous section).

Very similar patterns were observed when taking *cis*-primals of directed hypergraphs. Slight difference in *trans*-primals of directed hypergraphs is due to the break of the clique structure in randomization.

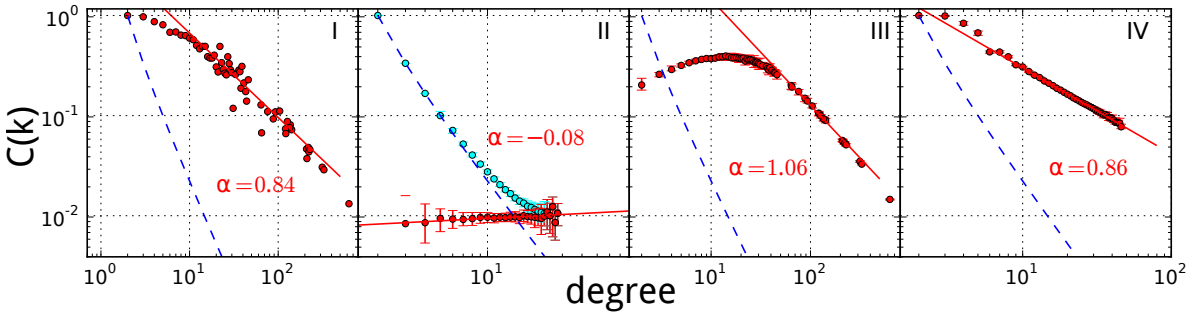

Figure 2.6: **Scaling of average clustering coefficients $C(k)$.** (I) The primal of *E. coli's* metabolic hypergraph. (II) Erdös-Rényi random graphs. The cyan dots are $C(k)$ calculated excluding nodes with $C = 0$. (III) Random graphs with same degree distribution as (I); (IV) primal of rewired versions of *E. coli*'s metabolic hypergraph (see text for more details). For II, III, and IV, the results are based on 100 replica, where the red dots denote the medians of the 100 replica. The red lines are least-squares fitting to $C(k)$ against k using power law. Power coefficients of the fitting are labeled. In each panel, the dashed blue curve corresponds to the points $[k, \gamma_k]$, where γ_k is the smallest C value that a node with degree k can take.

Very similar patterns were observed when taking *cis*-primals of directed hypergraphs. Slight difference in *trans*-primals of directed hypergraphs is due to the break of the clique structure in randomization.

For an Erdös-Rényi random graph with 1193 nodes and 5710 edges, a small value of $C(k)$ is expected as the connectivity is very sparse; this is shown in Fig. 2.6(II). However, if we exclude nodes whose clustering coefficient is 0, $C(k)$ scales almost

exactly the same with the smallest non-zero C values that a node with a particular degree k can take (blue dashed line in Fig. 2.6). This smallest non-zero clustering coefficient equals the reciprocal of the total number of connections among the k neighbors of the node we consider, which is $2/(k^2 - k)$, and thus scales with $\alpha = 2$ when k is large (that is, $2/(k^2 - k) \approx bk^{-2}$ for large k). In other words, for a sparse Erdös-Rényi graph, the scaling of C with $\alpha = 2$ is very likely.

If we rewire the primal of *E. coli*'s hypergraph in such a way that we preserve the degree distribution, then we obtain graphs whose tail of clustering coefficient distribution scales with an $\alpha = 1.06$, as shown in Fig. 2.6(III). This, to a certain degree, weakens the statistical significance of the scaling observed in Fig. 2.6(I).

However, when we employ the null model like that of the previous section (see Fig. 2.4), where the hyperedge cardinality distribution is preserved, we observe that not only do the clustering coefficients scale, but that the scaling has an almost identical value of α; see Fig. 2.6(IV). This finding challenges the statement that hierarchical connectivity of metabolic networks corresponds to functional organization. Or, even if such a correspondence still exists, our finding here does not support the hypothesis that such structure is selected for, since random graphs generated based on the new null model exhibit similar scaling properties.

We also studied the clustering coefficient on reaction graphs obtained through PLGT (see Fig. 2.1). Contrary to the previous observation that the average clustering coefficient $C^T(k)$ scales as $C^T(k) \propto k^{0.08}$ [175], $C^T(k)$ does not show clear scaling in this study (see Fig. 2.7).

Further, in this case we find that the clustering coefficients are greatly affected by the presence of metabolites that participate in a large number of reactions, or the so-called "currency metabolites", such as water. With water removed from the original hypergraph, the entire rightmost vertical strip in the PLGT's clustering co-

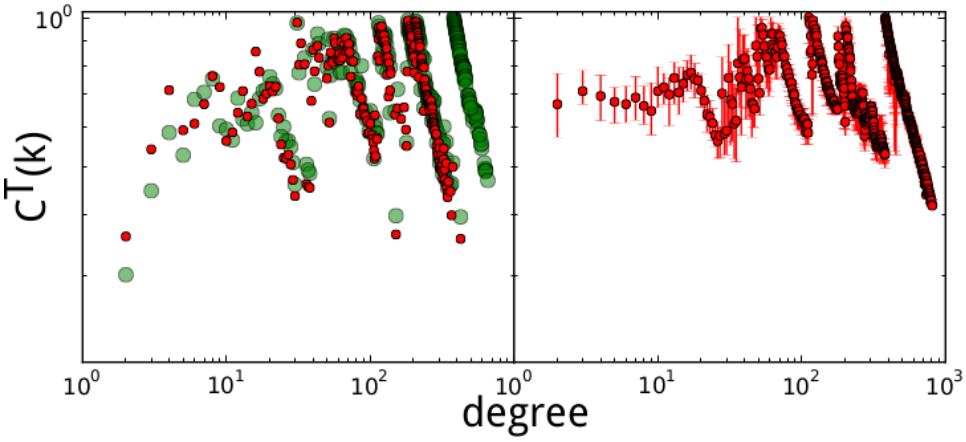

Figure 2.7: **The scaling of averaged clustering coefficients in the reaction graph obtained via PLGT.** Left panel: The green dots are the average clustering coefficients of the PLGT of *E. coli*'s hypergraph. The red dots are the same, but with the water molecule excluded. Right panel: The red dots are the average clustering coefficients of the rewired hypergraph, which is generated by first taking the dual of *E. coli*'s metabolic hypergraph and then rewiring it (1×10^6 times, to guarantee convergence).

efficients disappears (red dots in Fig. 2.7) (Effects of the removal of other "currency metabolites" are also studied; data not shown). This is because a node with degree k becomes a hyperedge with cardinality k in the dual hypergraph, giving rise to $k(k-1)/2$ connections in its primal which is the final PLGT product. This has two complications. First, through PLGT, the graph becomes denser. The average degree, or twice the number of edges per node, increased from 9.6 to 228.0. Second, the difference in the contribution to the connection from nodes of different degrees increases significantly, from k to $k(k-1)/2$. The node with the largest degree (water) is at least partially responsible for most of the connections in the PLGT result.

The results of $C(k)$ against k on the PLGT graphs are different from the ones on randomized graphs, whether the graph abstracted is rewired directly or the underlying hypergraph is rewired and abstraction is made thereafter. However, if the dual hypergraph (of which the PLGT is the primal) is rewired while keeping the number of reactions in which each metabolite participates, the results of $C(k)$ against k on

the standard graph abstracted thereafter is similar to the one observed on the PLGT of the *E. coli* hypergraph (right panel of Fig. 2.7). Once again, this result stresses the implications of the used null model, and how this affects the significance of values computed on biological networks.

The question, then, is: why is this scaling of clustering coefficients? Or, why is this hierarchical structure of graphs abstracted from hypergraphs? We believe that this is simply an artifact of the way standard graphs are abstracted from metabolic hypergraphs. For example, the primal of an undirected hypergraph connects all the reactants in the same reaction, thereby forming cliques in the abstracted standard graph. These cliques contribute the same number of 2-paths and triangles in computing the clustering coefficient of a reactant. Since the number and size of such cliques remain unchanged as a hypergraph is rewired, their contribution remains the same as well. The similarity between the scaling of $C(k)$ in metabolic standard graphs and ones abstracted from randomized hypergraphs indicates that cliques thus formed probably dominate the value of clustering coefficients and thus their scaling in the context of the real-world metabolic networks. In other words, the scaling of $C(k)$ is kept largely by the hyperedge cardinality distribution which is intrinsic to the structure of biochemical reactions but not to how the metabolic hypergraph is organized using these reactions.

In order to figure out whether the scaling of clustering coefficients is due to the inherent "hierarchy" of the metabolic graph, or is just a consequence of the graph abstraction process and the hyperedge cardinality distribution, we computed the hypergraph clustering coefficient using a new measure we devised to apply directly to hypergraphs (see Methods). Results are shown in Fig. 2.8 for *E. coli*'s hypergraph (left panel) and its dual (right panel). The result of clustering coefficient computed using the measure of [174] are similar.

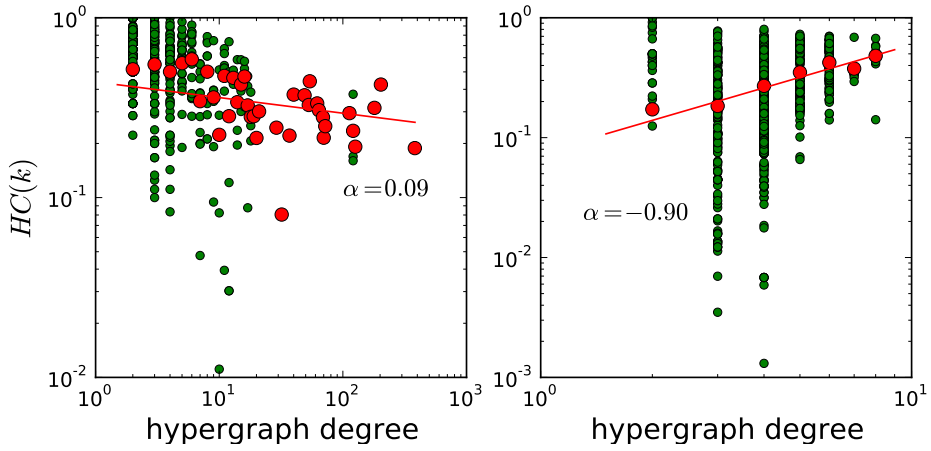

Figure 2.8: **The scaling of hypergraph clustering coefficient.** The green dots are the local clustering coefficients. The red dots are averaged value of the local clustering ceofficients for each degree. Left panel: *E. coli*'s hypergraph. Right panel: The dual of *E. coli*'s hypergraph.

The hypergraph average clustering coefficients show very weak scaling. The individual clustering coefficients are more scattered around. The value of α (0.09) is much smaller than what is observed on the standard graph (0.84, Fig. 2.6(I)) and the value of 1.1 as reported in [175]. As for the dual hypergraph (right panel of Fig. 2.8), we find that the clustering coefficients of the dual hypergraph, from which the line transformed reaction graph is abstracted, shows better scaling but with an α of a larger magnitude. Still, the actual values of the clustering coefficients are very scattered and show no scaling.

To summarize, we believe topological characteristics of metabolic networks, such as scale-free degree distributions and scaling of clustering coefficients, are not necessarily a ground for invoking natural selection or making connections to functional organizations. Instead, these properties may lose statistical significance when a null model taking into account of the reaction sizes is used, and may even disappear when computations are done on the appropriate representation of metabolic networks.

2.4.0.1 Clustering coefficients on hypergraphs

A commonly used statistic for elucidating properties of metabolic networks, such as modularity [53] and small-worldness [181], is the *clustering coefficient*. Among the various existing definitions of the clustering coefficient, the *local clustering coefficient* by Watts and Strogatz [181] and the *global clustering coefficient* by Barrat and Weigt [182] are the most widely used.

According to [181], the local clustering coefficient, C_{local}, for any given node v (with $d(v) \geq 1$) in an undirected (standard) graph is defined as the fraction of the number of edges linking pairs of v's neighbors over the number of all such possible edges (which equals $\binom{d(v)}{2}$). For a node with $d(v) = 0$, we have $C_{\text{local}}(v) = 0$. Intuitively, C_{local} measures, for a node v, the probability that a randomly chosen pair of its neighbors would be seen connected.

According to [182], for an undirected graph with at least one 2-path (three distinct nodes connected via two edges), the global clustering coefficient C_{global} is defined as the fraction of the number of 2-paths with linked end points (i.e., triangles) over the number of all possible 2-paths. Intuitively, C_{global} measures the probability of having an edge (u, w), given that edges (u, v) and (v, w) exist, with u, v, w being three distinct nodes.

For a proper extension of C_{local} and C_{global} to the domain of hypergraphs (denoted by HC_{local} and HC_{global}, respectively), the following intuitive properties may be desirable, in addition to reflecting the extent of clustering in a hypergraph:

P1 The values of HC_{local} and HC_{global} fall in the range $[0, 1]$.

P2 HC_{local} and HC_{global} should reduce to C_{local} and C_{global}, respectively, when every hyperedge connects exactly two nodes (i.e., the hypergraph is a standard graph).

P3 $HC_{\text{local}}(v)$ should reflect the extent of connectivity among neighbors of v due

to hyperedges other than ones connecting v with those neighbors.

The rationale behind property P1 is to retain the probabilistic interpretation of the clustering coefficient statistic, as well as to enable comparing two different hypergraphs under the statistic. The rationale behind property P2 is to allow treating hypergraphs and standard graphs (which are a special case of hypergraphs) in a uniform manner. Property P3 reflects the fact that neighbors of a node can also be neighbors simply since all three belong to the same hyperedge—a case that should be treated carefully to reflect a proper notion of clustering.

Based on these properties, we define $HC_{\text{local}}(v)$ and $HC_{\text{global}}(H)$ as follows for a hypergraph $H = (V, \mathcal{E})$ and $v \in V$:

$$HC_{\text{local}}(v) = \begin{cases} \dfrac{1}{\binom{|\mathcal{M}(v)|}{2}} \displaystyle\sum_{\substack{E_i, E_j \\ \in \mathcal{M}(v)}} EO(E_i, E_j) & \text{if } d(v) > 1 \\ 0 & \text{if } d(v) = 1 \end{cases} \tag{2.4}$$

$$HC_{\text{global}}(H) = \begin{cases} \dfrac{1}{|\mathcal{I}|} \displaystyle\sum_{\{E_i, E_j\} \in \mathcal{I}} EO(E_i, E_j) & \text{if } \mathcal{I} \neq \emptyset \\ 0 & \text{if } \mathcal{I} = \emptyset \end{cases} \tag{2.5}$$

where, $\mathcal{I} = \{\{E_i, E_j\} \subset \mathcal{E} \mid E_i \cap E_j \neq \emptyset \wedge E_i \neq E_j\}$, and the *extra overlap* of two intersecting hyperedges E_i and E_j is defined as:

$$EO(E_i, E_j) = \frac{|N(D_{ij}) \cap D_{ji}| + |N(D_{ji}) \cap D_{ij}|}{|D_{ij}| + |D_{ji}|}. \tag{2.6}$$

where $D_{ij} = E_i - E_j$. For two hyperedges E' and E'' such that $E' = E''$, we define $EO(E', E'') = 0$. Fig. 2.9 provides examples of the values of EO and HC_{local} under a

variety of scenarios. For HC_{global}, the numerator is the sum of extra overlap between any pairs of hyperedges that contain v, and the denominator is the number of all possible pairs of such hyperedges.

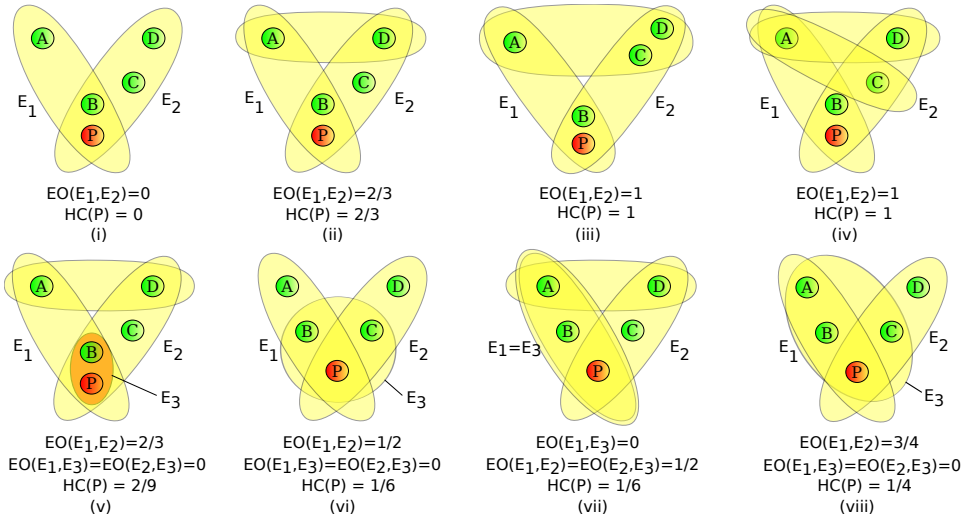

Figure 2.9: **Illustration of the *Extra Overlap* and the *Local Clustering Coefficient* for hypergraphs.** $EO(E_i, E_j)$ denotes the extra overlap between hyperedge E_i and E_j, and HC(P) denotes the local hypergraph clustering coefficient for the node P.

From the definition of EO, we observe the following:

1. $EO(E', E'') \in [0, 1]$ for every pair of hyperedges E' and E''.

2. For two non-identical, intersecting hyperedges, E_i and E_j, each of cardinality 2, $EO(E_i, E_j) = 1$ when their non-shared elements are linked by a third hyperedge, and $EO(E_i, E_j) = 0$ otherwise.

3. For any two sets $E, E' \subseteq V$, where $E' \subseteq E$, $EO(E, E') = 0$.

It follows from these observations that HC_{local} and HC_{global} satisfy the three aforementioned properties P1—P3.

Note that we are not the first to define clustering coefficient measures for hypergraphs. Estrada and Rodríguez-Velázquez [174] defined their (global) clustering

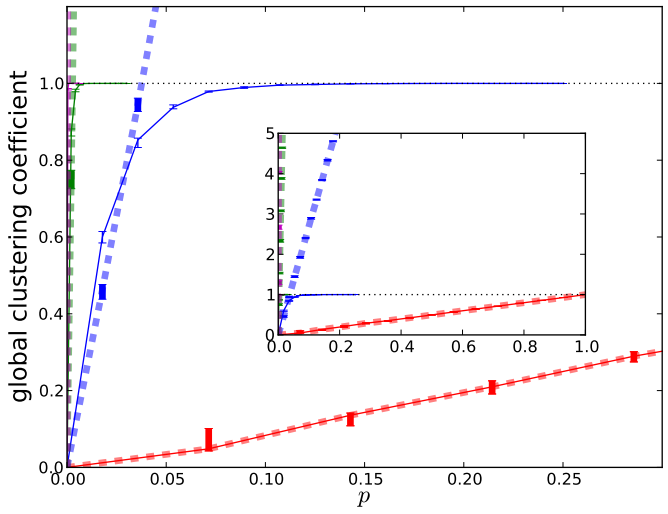

Figure 2.10: **Comparison of the two global hypergraph clustering coefficient measures on random hypergraphs.** The x-axis shows the probability p with which a hyperedge with a fixed cardinality is added, and the y-axis shows the value of the global clustering coefficient. Each hypergraph has 30 nodes. Solid and dashed lines correspond to our measure (Eq. 2.5) and the ERV measure (Eq. 2.7). Red, blue, green and magenta colors correspond to hyperedge cardinalities 2, 3, 4, and 5, respectively. A completely connected hypergraph has $p = 1$. Note that for $|E| = 2$ (red) the two coefficients agree and both degenerate into the standard graph clustering coefficient [182]. Each data point shows the median of 15 replica and the error bar shows the upper and lower quartiles.

coefficient for hypergraphs, denoted ERV hereafter, as

$$\frac{6 \times \text{the number of hyper-triangles}}{\text{the number of 2-paths}} \tag{2.7}$$

where a hyper-triangle is a set of three nodes and three hyperedges that connect them, and a 2-path is a sequence $\{u, E_1, v, E_2, w\}$, where u, v, w are three distinct nodes, E_1, E_2 are two distinct hyperedges, $\{u, v\} \subseteq E_1$ and $\{v, w\} \subseteq E_2$. The numerator is essentially the number of the closed-walks of length 3 without reusing hyperedges or revisiting nodes except at the end points [174].

To analyze how the two measures of global clustering coefficients compare, we conducted a simple test, where we generated random hypergraphs with increasing

connectivity and applied the measures to them. More precisely, we generated a random graph by starting with 30 disconnected nodes, and then, for each subset of m nodes, we connected them by a hyperedge with probability p. Finally, we applied the two measures to the generated graph. In our experiment, we used $m = 2, 3, 4, 5$ and varied p between 0 and 1, and for each combination of values of m and p, we repeated the experiment 15 times, plotting the median of the 15 runs in Fig. 2.10.

Three observations are in order. First, the two measures yield identical results in the case of standard graphs (where $m = 2$), since they both reduce to the standard global clustering coefficient statistic on standard graphs when all hyperedges have cardinality 2. Second, the two measures begin to deviate as our measure approaches 1. In particular, the ERV measure is not bounded from above, and goes beyond 1 quickly for hyperedge cardinality higher than 2. This makes hard the interpretation of values computed by the ERV measure, since they cannot be treated in a probabilistic manner. Further, the ERV measure would not allow for comparing two hypergraphs in terms of their clustering coefficients since the values are not bounded. Last but not least, in both definitions of the hypergraph clustering coefficient, the hypergraphs with higher hyperedge cardinalities approach 1 much faster in their global clustering coefficient. The reason for this is that the total number hyperedges of a given hyperedge cardinality (which equals $\binom{|V|}{|E|}$) grows exponentially with the hyperedge cardinality value $|E|$. Therefore, the density $p = |\mathcal{E}|/\binom{|V|}{|E|}$ is diminished by the same factor if the number of hyperedges $|\mathcal{E}|$ is kept fixed. This further illustrates the fact that hyperedge cardinality plays a significant role in the clustering coefficient computed on hypergraph and beyond. Similar patterns were observed for the local clustering coefficients measures.

Conclusion

In this chapter, we investigated the impact of choosing a null model that incorporates the hypergraph property of the metabolic system such as the reaction size distribution to the networks' connectivity analyses. By reanalyzing the degree distribution and clustering coefficient we found that the reported scaling of the clustering coefficients in the metabolic graphs and its specific power coefficient may be an artifact of the hypergraph abstraction, and is not supported when biochemical reactions are atomically treated as hyperedges. Also we found that by taking into the reaction size distribution, a null model can explain some of the details in the shape of the degree distribution that have not been explained otherwise, further highlighting the necessity of using appropriate null models in exploring adaptive evolution, along with the analysis of their support in biological systems.

Chapter 3

Network pruning and pathway inference

3.1 Introduction

Graph representation of a metabolic network connectivity map provides a simple representation of certain relationships among the network's entities. Analyses of such graphs have provided various insights into the properties of metabolic networks, yet not without controversy. For example, the finding of a short average path length in metabolic networks (e.g., [22, 21]) has been challenged in that it was based on the "raw" metabolic graphs, without first *pruning* them [183].

To illustrate the concept of "pruning," consider the following reaction analyzed in [184]:

$$\begin{aligned} &\text{N-Acetylornithine} + \text{L-Glutamate} \quad \Leftrightarrow \\ &\qquad\qquad\qquad\qquad \text{L-Ornithine} + \text{N-Acetyl-L-glutamate} \end{aligned} \tag{3.1}$$

In their network assembly, the authors only linked N-Acetylornithine to L-Ornithine and L-Glutamate to N-Acetyl-L-glutamate and omitted the link between L-Glutamate and L-Ornithine and the link between N-Acetylornithine and N-Acetyl-L-glutamate. Depending on the edge semantics of the network and the subsequent analyses, this pruning step may or may not make a difference. From the perspective of *causality*

of biochemical transformation and *pathway inference* [184], this pruning makes sense: no chemical content is conserved between L-Ornithine and L-Glutamate, and the acetyl group that is conserved between N-Acetylornithine and N-Acetyl-L-glutamate is not sufficiently representative for linking the two chemical compounds. However, if one is modeling the *propagation of perturbation* in the concentration of metabolites (e.g., [21]), then not removing these last two edges makes sense, since they do capture how a perturbation to certain metabolites may spread throughout the metabolic system. Under this semantics, metabolic graphs are built by connecting all the metabolites that participate in a reaction [164], and subsequently analyzed, without post-processing the connectivity [53], to elucidate properties on information transfer, network robustness and resilience, etc. In this paper, we focus on the first of the edge semantics, namely causality of biochemical transformation.

To build metabolic graphs for pathway inference, all metabolites participating in a reaction are connected to form the *raw graph*, and then, via *connectivity pruning*, edges that may result in the inference of biochemically implausible pathways [185, 184] are pruned (hypergraph-based pathway inference techniques, such as the network expansion [32], require different treatment and are beyond the scope of this paper).

Several methods exist for pruning metabolic graphs including hub deletion [36], removal of currency metabolites [34, 69], manual curation [27] and RPair typing [186]. However, the ambiguity inherent in these *ad hoc* methods, and the lack of a systematic one, may confound analyses of metabolic graphs [27, 187, 23]. Here we propose a simple criterion, the *strength of chemical linkage* (SCL), for systematic pruning of metabolic graphs. By analyzing the metabolic graph of *E. coli*, we demonstrate the power of this criterion in yielding biochemically meaningful pathways. Further, we characterize the commonly used pruning heuristics in terms of the strength of chemical linkage, and discuss the ambiguity in these methods and the superiority of

using the SCL criterion. Finally, we demonstrate the utility of the criterion in pruning the search tree used in pathway inference methods to gain in accuracy and efficiency compared with other graph-based search heuristics (e.g., [42]).

3.2 Methods

3.2.1 Reaction data and reference pathways

1383 reaction equations were obtained from KEGG Ligand database [176]. For each reaction with any gene in *E.coli* annotated to produce an enzyme that catalyzes the reaction, we assembled a graph connecting every pair of metabolites that sit on opposite sides of the reaction (the raw graph). Both reaction-enzyme mapping information and enzyme-gene mapping information were downloaded from KEGG. Following common practice [188], we removed any reaction that appeared in the reference pathway yet did not have a definite gene annotation; e.g., reaction R07765 has only the generic EC number 1.3.1.- even though there are genes in *E.coli* that are annotated for that EC number. For reactant pairs that exist in the KEGG RPair database, we information on the molecule alignment. "Markush structures" and groups with label "R" were taken as one atom.For reactant pairs that do not exist in the database, we manually set the alignment number to 0. This treatment is dependent on the coverage of RPair database to all possible reactant pairs with non-zero alignments. We manually verified that the coverage is satisfactory. Out of 1383 reactions that exist in the *E.coli* network, 1104 reactions needed to be treated with additional specification to connections with alignment number 0. Out of 2642 connections with alignment number 0 added in these reactions, 98 connections have actual nonzero chemical linkage. The percentage of unsatisfactory connections was less than 4%. Moreover, from a closer inspection, many of these linkages are hard

to process because of the use of generic compounds and unbalanced reactions in the KEGG Ligand database [189, 190].

Reference pathways were obtained from the KEGG KGML pathway files. Reactions that exist in the reference pathways but are not validated by the presence of clearly defined enzymatic information, and thus do not appear in the total set of reactions from which the raw metabolic graph is assembled, were removed. Annotated pathways for the pathway inference validation were obtained from the aMAZE database [191]. We excluded those pathways that are duplicates in terms of using the same sequence of compounds and those that have fewer than two steps. Two other manually curated reference pathways were also included. The name of the pathways are listed in Table 3.1.

3.2.2 The Strength of Chemical Linkage (SCL) criterion

We define the *strength of chemical linkage*, or SCL, for two reactants as the proportion of chemical content conserved between them in a reaction, normalized by the maximum chemical content of either of the two reactants. If there is more than one mechanism that involve the same two reactants, SCL takes the maximum result computed over all such mechanisms. For example, in some rare cases, two reactants can be converted to one another via more than one mechanism even in one reaction; e.g., C00022-C00900 in R00006 (the C and R labels are standard indices used in KEGG [176]). Chemical content can be quantified in many ways; in this paper, we use the *absolute atom counting*. Under this quantification, given the set $\mathcal{C}(A)$ of non-hydrogen atoms in molecule A (due to the fact that hydrogen is not generally considered the backbone of biochemical compounds), the chemical content of the compound is simply $|\mathcal{C}(A)|$. Atoms are mapped in the reaction according to the true chemistry. While in this study we use the KEGG RPair database for molecule

alignments, SCL depends only on the physics of the real chemical reaction and is independent of the data source.

Formally, for a compound pair (A,D) that sit on two sides of a reaction (e.g., $A + B \Leftrightarrow C + D$), we define

$$\begin{aligned} SCL_{\text{self}}(A|D) &= |\mathcal{C}(A) \cap \mathcal{C}(D)|/|\mathcal{C}(A)| \\ SCL &= |\mathcal{C}(A) \cap \mathcal{C}(D)|/\max(|\mathcal{C}(A)|, |\mathcal{C}(D)|) \\ &= \min(SCL_{\text{self}}(A|D), SCL_{\text{self}}(D|A)). \end{aligned} \tag{3.2}$$

$SCL_{\text{self}}(A|D)$ measures the contribution of chemical content from D to A, or equivalently, how much chemical content of A comes from D. A high $SCL_{\text{self}}(A|D)$ value indicates a greater importance in chemical composition of D when A is produced/consumed in the reaction. When compound D is clear from the context (e.g., when there is only one reactant on the other side), we write $SCL_{\text{self}}(A)$. Further, in this context, we use the definition $SCL_{\text{other}}(A) = SCL_{\text{self}}(D|A)$. High SCL is an indication of stronger chemical causality in terms of chemical content between the two reactants in a specific reaction.

3.2.3 A pathway inference method

In order to demonstrate the quality of SCL-based pruning, we introduce a simple, SCL-based algorithm for identifying a set of pathways from a source to a target metabolite in a given metabolic graph; see Algorithm 1. The algorithm first identifies a set of candidate paths within a maximum length in a breadth-first manner (Lines 5–10), extending paths (Lines 9–10) only using edges with an SCL value higher than a certain threshold T (in our study, we use $t = 0.4$; see Section 3.3.1 for a discussion of this choice).

After exploring all the valid paths of a certain length, all paths are reordered

according to the minimum SCL_{self} value of all steps along the path (Line 11). Only the top (if there are more than) $N = 1000$ paths are saved for further exploration in the next round (Line 12). Finally the paths are ranked by the minimum SCL_{self} value of any edge they contain, where paths with lower values are ranked higher (Line 13).

Algorithm 1: **InferPathway.**

Input: Source metabolite: s; Target metabolite: t; Threshold of SCL: T; Maximum path length: L;
Output: A list of ranked pathways $\mathcal{S}_{\text{Result}}$.
2 $\mathcal{S}_{\text{ToExplore}} \leftarrow$ a path that contains only t;
4 $\mathcal{S}_{\text{Result}} \leftarrow \emptyset$;
6 **while** $\mathcal{S}_{\text{ToExplore}} \neq \emptyset$ *&&* *path length* $< L$ **do**
8 $\mathcal{S}_{\text{Temp}} \leftarrow \emptyset$;
10 **foreach** *path* p *in* $\mathcal{S}_{\text{ToExplore}}$ **do**
12 **foreach** *neighbor* n *of the last node* l *in* p **do**
14 **if** $n = s$ **then**
16 $\mathcal{S}_{\text{Result}} \leftarrow p$ extended by n;
17 **end**
19 **else if** $n \notin p$ *&&* $SCL_{self}(l|n) > T$ **then**
21 Extend p using n and add the new path into $\mathcal{S}_{\text{Temp}}$;
22 **end**
23 **end**
24 **end**
26 Sort $\mathcal{S}_{\text{Temp}}$ by the minimum SCL_{self} on each path;
28 $\mathcal{S}_{\text{ToExplore}} \leftarrow$ top 1000 paths in $\mathcal{S}_{\text{Temp}}$;
29 **end**
31 Sort $\mathcal{S}_{\text{Result}}$ by the minimum SCL_{self} on each path;
33 **return** $\mathcal{S}_{\text{Result}}$;

3.3 Results

To assess the quality of SCL, we conducted three tasks on *E. coli*'s metabolic graph. First, we computed the distribution of SCL values for edges in the raw graph, categorized based on presence/absence in the reference pathways to assess the power of SCL in selecting connectivities. Second, we explored pathway inference guided by SCL values to assess its power in finding compound-to-compound linear/cyclic biochemical pathways. Third, we studied existing metabolic graph pruning methods in the light of the SCL criterion.

3.3.1 The distribution of SCL values

Fig. 3.1 shows the distribution of SCL values of raw graph edges present/absent in the reference pathways. Most of the edges present in the reference pathways have high

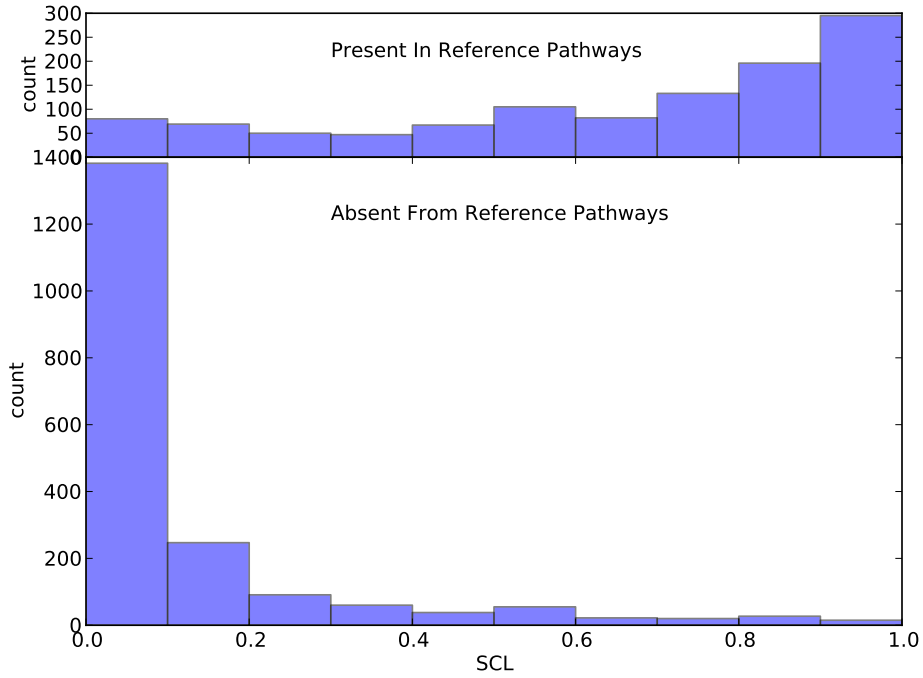

Figure 3.1: The distribution of SCL values based on the 1124 edges in the raw metabolic network of *E.coli* that are present in (top panel) and 1957 ones that are absent from (bottom panel) the reference pathways.

SCL values, whereas a great majority of those missing from the reference pathways have SCL values lower than 0.2. These results show that SCL is strongly correlated with absence/presence in reference pathways; stated differently, SCL can be used as a selection criterion for deciding which connections to include in a reference pathway.

Next, we studied whether pruning the raw graph based on thresholding the SCL values (keeping only edges with SCL values no less than a threshold T) produces the conventional pathway connectivities. Pathway maps are organized into five categories according to the hierarchy provided by the KEGG pathway database: carbohydrate metabolism, nucleotide metabolism, lipid metabolism, amino acid metabolism, and

metabolism of cofactors and vitamins. In addition, we considered two higher-level pathway unions provided by KEGG, namely eco01100 (Metabolic Pathways) and eco01110 (Biosynthesis of secondary metabolites). ROC curves for some of the pathways in the Carbohydrate Metabolism and the pathway unions are shown in Fig. 3.2.

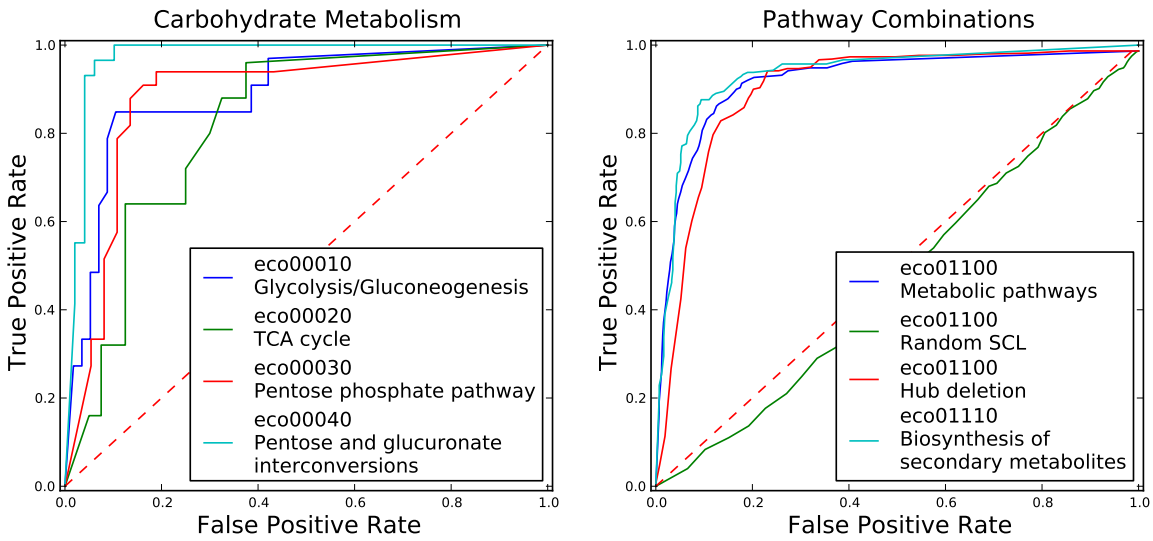

Figure 3.2: ROC curves based on thresholding the SCL values. Each curve is based on the raw graph restricted to a particular pathway map and on 50 threshold values evenly distributed in the range $[0, 1]$. Similar results have been obtained on other pathway maps (data not shown). The right panel is based on a combination of pathways where the ROC curve of hub deletion and one based on random SCL assignment are also shown in red and green. The ROC curve on hub deletion is obtained by tuning the degree threshold of the hub definition, thus changing the presence/absence of the network connections.

For a specific pathway, we count the number of edges from the raw metabolic network that are present in the reference pathway. Positives (P) (Negatives (N)) are defined as the connections that do (do not) exist in the reference pathway. For each value of the threshold T, only edges with SCL values $\geq T$ are kept, and the rest are removed. The true positives (TP) are the positives that also exist in the thresholded network. The false positives (FP) are the positives that do not exist in the thresholded network. The true positive rate (TPR) equals $|TP|/|P|$, and the false positive rate (FPR) equals $|FP|/|N|$. Notice that when $T = 0$, TP = P and

FP = N, giving $TPR = FPR = 1$. On the other hand, when $T > 1$, TP = FP = $\emptyset$, giving $TPR = FPR = 0$. In each panel, T increases in the direction from the upper-right corner to the lower-left corner. The concave shapes of the ROC curves indicate that thresholding SCL has strong power of selection for connections that appear in pathways, as opposed to those that are missing. This further validates our implicit reasoning that conventional pathway connectivity is a reflection of the chemical linkage strength.

While no SCL threshold seems to exist for perfect retrieval of established biochemical pathways, our detailed study of the Glycolysis/Gluconeogenesis pathway (with threshold $T = 0.4$) revealed two main reasons (beside the issue with KEGG's RPair database coverage) behind the false cases.

The first reason is the presence of reactant pairs with special roles. Not all the reactant pairs actively participate in the mass circulation, but are required for, e.g., energy dependencies. For example, ATP-ADP drives a reaction towards a certain direction [184]. Many such reactant pairs have strong chemical linkage with each other, yet weak linkage with other reactants. They are commonly perceived as carriers of small chemical moieties, such as proton (NAD, NADH), phosphate (ATP, ADP; Protein-histidine, Protein N-phospho-L-histidine), and acetyl group (CoA, Acetyl-CoA).

These reactant pairs with special roles usually cause false positives, i.e., reactant pairs absent from the reference pathways but with a high chemical linkage. Nonetheless, these false positives are completely tolerable and, in our view, are even better to be preserved in the network. For pathways where these reactant pairs are used for non-mass circulation reasons (e.g., eco00010), they are usually disconnected from the bulk network component due to a low SCL value with other reactants (data not shown). Therefore, their presence would not confuse the pathway inference by creat-

ing biochemically unintuitive shortcuts. Moreover, it makes sense to preserve these reactant pairs in the network and the reference pathways since they represent the way how energy is consumed. For example, when ADP is used to make ATP, such cycling would be unclear if the consumption of ATP is missing from the network. More importantly, being a "carrier" is in itself ambiguously defined, since all reactant pairs "carry something." In order to be a "carrier", same reactant pair should appear in multiple reactions to "load" and "unload" chemical groups. We have observed that although there exist certain reactant pairs that participate in a large number of reactions, there is in general no clear-cut boundary for being a "carrier". (data not shown)

The second reason is the inappropriate quantification of chemical content by *absolute atom counting*. For example, in the following reaction [Pyruvate + Thiamin diphosphate $\Leftrightarrow$ 2-(alpha-Hydroxyethyl)thiamine diphosphate + CO_2], reactant pair (CO_2, Pyruvate) is missing from the reference pathway and its presence in the thresholded network can potentially give rise to shortcuts between Pyruvate and other irrelevant compounds via CO_2. The total number of non-hydrogen atoms of Pyruvate is 6 and that of CO_2 is 3. Although they share 3 atoms in the reaction, 2 of them are oxygen and only 1 is carbon, which is traditionally considered to be the backbone of Pyruvate. Therefore when all the non-hydrogen atoms are included $SCL = 3/6 = 0.5$. However, the value becomes $SCL = 1/3 \approx 0.3$ when only carbon atoms are considered, since Pyruvate has 3 carbons, CO_2 has 1, and they share 1 carbon in the reaction. By counting only carbons, the false positive connection can be avoided. This not only reflects the fact that the atoms are treated as biochemically different, but also suggests that alternatives to *absolute atom counting* (e.g., by counting only carbons) might improve the performance of the criterion on certain (but not all) reactions.

Further, a large collection of atoms may form some chemical group that functions as a single unit. In the context of one particular pathway, their detailed composition, creation and degradation is not relevant. But, again, if we count only the number of non-hydrogen atoms, the criterion might be biased. For example, in reaction [ATP + Acetate + CoA $\Leftrightarrow$ AMP + Diphosphate + Acetyl-CoA], the reactant pair (CoA,Acetyl-CoA) is falsely present (false positive) while the pathway reactant pair (Acetate, Acetyl-CoA) is missing (false negative). 48 non-hydrogen atoms are conserved in the former reactant pair, while only 3 are conserved in the latter. Under *absolute atom counting*, the former reactant pair is stronger in chemical linkage while the latter is weaker. However, biochemically, all 48 atoms in the former comes from CoA which functions as a single unit in the context of glycolysis pathway, while the 3 atoms conserved in the latter contribute to 3 out of 4 non-hydrogen atoms in acetate. If we count all atoms in CoA as 1, we obtain $SCL = 1/4 = 0.25$ for the (Acetate, Acetyl-CoA) pair and $SCL = 3/4 = 0.75$ for the (CoA, Acetyl-CoA) pair, which would avoid both false cases. However, to do this throughout the metabolic graph, finer delineation of functional groups for metabolites is needed, which we target as future work.

By scanning the threshold from 0 to 1 with increment 0.01, we find that the range of threshold value that minimizes the false cases (both false positive and false negative) is from 0.38 to 0.39. Nevertheless, as shown in Fig. 3.2, when individual pathway is concerned, there does not exist a threshold value of SCL that suits all (data not shown). We found that the range of optimal threshold in different pathways varies not only in magnitudes, but also in lengths. Some pathways reach optimal pruning under a wide range of threshold values. Besides, the optimal threshold of some pathways can be explained by their biochemical function. For example, pathways involved in fatty acid metabolism have a lower threshold. This reflects the fact that

links in these pathways are responsible for the extension of a long fatty acid chain by one small residue which is weak in the sense of relative mass conservation.

3.3.2 Using SCL in pathway inference

In order to investigate the effectiveness of SCL in pathway inference, we applied Algorithm InferPathway (see Methods) to source/target pairs of 8 reference *E.coli* pathways obtained mostly from the aMAZE database [191]. Results are shown in Table 3.1.

Table 3.1: **Inference of aMAZE pathways** [191]. 'Length' is the number of reactions from the source to the target compound in the reference pathway. 'Rank' is the place of the reference pathway as identified by Algorithm InferPathway.

Pathway name	Length	Rank
Arginine Catabolism	3	1
Arginine Utilization	4	1
Chorismate Biosynthesis	4	1
Glucuronate Catabolism	3	1
Lysine Biosynthesis	7	1
Threonine Biosynthesis	3	1
Oxidative Pentose Phosphate Pathway	4	1
Glycolysis	6	2
Methionine Biosynthesis	5	95

In Fig. 3.3, we show two of the reference pathways which are correctly returned by our method as the top result, namely, the Lysine Biosynthesis and Oxidative Pentose Phosphate pathway, as well as the two pathways that differ from our top results, namely, the Glycolysis and Methionine Biosynthesis pathways.

For Glycolysis (iii of Fig. 3.3), the only difference between the reference pathway and our top result (ranked second; see Table 3.1) is the use of reaction R01827 instead of a combination of reactions R04779 and R01070. The shortcut is a documented step in the KEGG pathway map, but only in the Pentose Phosphate Pathway

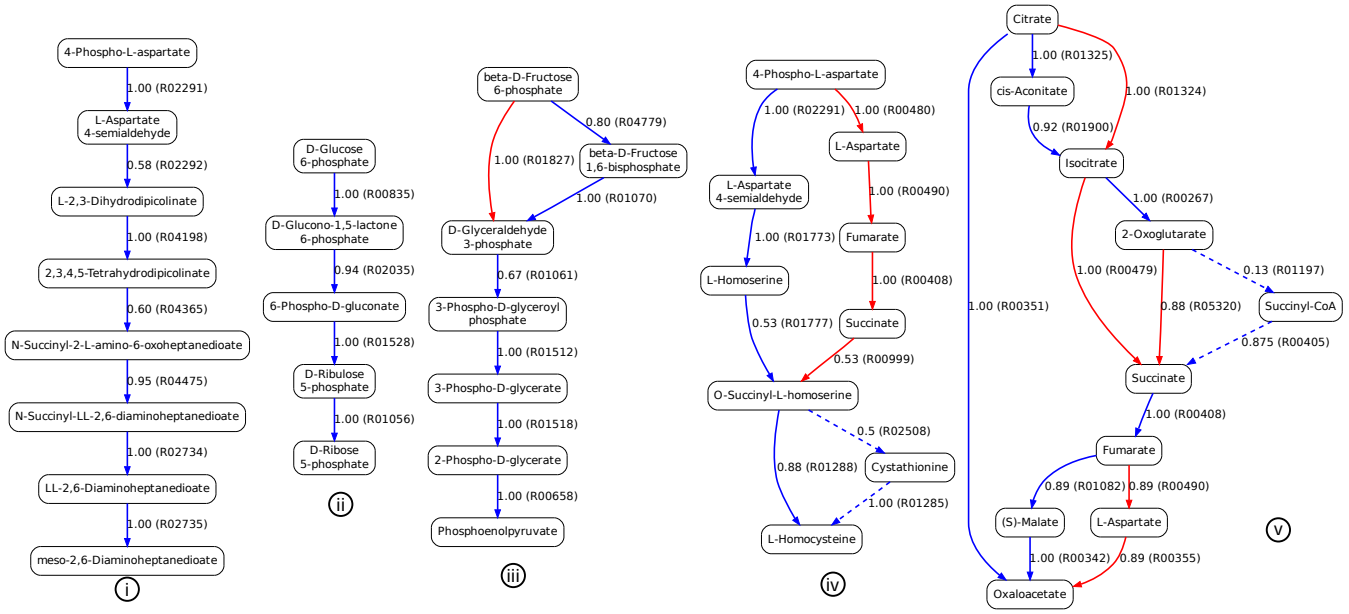

Figure 3.3: Pathways inferred by applying algorithm **InferPathway** onto the *E.coli* metabolic network. Nodes are labeled by the metabolite name used in KEGG. Edges are labeled by the SCL_{self} of the link, along with the reaction ID used in KEGG of one of the reactions that make available the transition shown in parentheses. The solid connections (both blue and red) correspond to the top 9 results returned from our method. **(i)**: Lysine Biosynthesis. **(ii)**: Pentose Phosphate Pathway. **(iii)**: Glycolysis. Blue connections are the annotated pathway and the second result returned by our method. The top result uses the shortcut shown in red. **(iv)**: Methionine Biosynthesis. The blue connections are the top result returned by our method. The annotated pathway differs from the first result by the dashed connections. The second result differs from the first result by the connections shown in red. **(v)**: TCA (tricarboxylic acid) cycle. The annotated TCA cycle is shown in blue (both solid and dashed, the dashed links are missing from the result).

map, indicating that the exclusion of the single reaction step is of manual, rather than biochemical, origins. For the Methionine Biosynthesis pathway (iv of Fig. 3.3), our method fails to return the annotated pathway as the top result. The only step that is consistently missing from our inference is the shortcut from O-Succinyl-L-homoserine to L-Homocystein without passing through Cystathionine, as is the case in the annotated pathway. Two clarifications are in order here. First, the shortcut step is structurally valid but infeasible in terms of free energy—information that is not incorporated into the reaction equation. The reaction's direction is to the left, as

Hydrogen sulfide takes the gas form under room temperature and leaves the system quickly once formed:

[O-Succinyl-L-homoserine + Hydrogen sulfide $\Leftrightarrow$ L-Homocysteine + Succinate]. Second, the missing step in the annotated path, composed of the two reactions [O-Succinyl-L-homoserine+L-Cysteine $\Leftrightarrow$ Cystathionine+Succinate] and [Cystathionine+ $H_2O \Leftrightarrow$ L-Homocysteine + NH3 + Pyruvate] has a low SCL_{self} value of 0.5. This is because to add just an SH (mercapto group), a Cysteine is recruited and Pyruvate is released subsequently. However, when computing SCL_{self} of the link from O-Succinyl-L-homoserine to L-Cystathionine, the entire Cysteine contributes to the number of non-conserved atoms, although only SH (1 non-hydrogen atom) is preserved after the subsequent step. The major part of Cysteine (6 non-hydrogen atoms) is released as Pyruvate. This issue can be solved by tracking the identity of each atom and recording for each intermediate metabolite the set of its atoms that are conserved all the way to the target.

We also studied the capability of the algorithm to find cyclic pathways by applying it to contiguous metabolites on the TCA cycle pathway (v in Fig. 3.3). The pathway connections corresponding to the top 9 results returned are shown. Indeed, all intermediate metabolites, except for the Succinyl-CoA, in the TCA cycle are recovered as well as all other connections that are documented in the pathway maps of KEGG. The low SCL value (0.13) on one of the two missing steps involving Succinyl-CoA is due to a reason same as discussed above, namely that CoA contains many atoms yet functions as one unit.

A satisfying consequence of this pruning strategy is its capability of not only getting pathways but also rejecting cases where in between the given source and target there is no linear unbranched pathway that is biochemically meaningful. If the synthesis of a metabolite requires contributions from many different sources, this

advantage would be reflected in our method as a low minimum SCL_{self} of all paths returned. To further illustrate the efficiency of pathway inference guided by SCL_{self}, we compare the search efficiency by only considering nonzero SCL (or equivalently, the presence/absence of annotations for reactant pairs in the RPair database), by pruning of extension using the SCL_{self}, by the pruning of exploration using minimum SCL_{self} on the path and by a combination of both pruning. We found first that no matter how large the results, minimum SCL_{self} always sorts out the reference pathway to a high rank (2 in case of glycolysis) while sorting using path length does not (170 ties with the highest rank being 65). With the same accuracy of the reference pathway, the result set (Fig. 3.4) and total number of node visits (data not shown) is greatly reduced when our pruning strategies are applied. Same observations have been obtained from other pathways. These results combined demonstrate the utility of SCL in not only sorting out the best pathway from the result set but also being effective in pruning the search tree of path finding.

3.3.3 Existing pruning methods in the light of SCL

Here, we compare existing pruning methods in terms of the SCL criterion, and discuss the necessity and superiority of SCL.

3.3.3.1 Hub and currency metabolites deletion

Hub deletion [36]) and currency metabolites [69] are compared with SCL-based methods. Fig. 3.5 shows that the average SCL value of a graph increases as hubs (nodes of high degree) are removed from the graph, which is in agreement with the rationale behind hub deletion [36]. Despite its similar effectiveness in pruning the network connection (red curve in Fig. 3.2), the degree of a metabolite depends only on the global layout of the network which has little meaning in the local chemistry of each

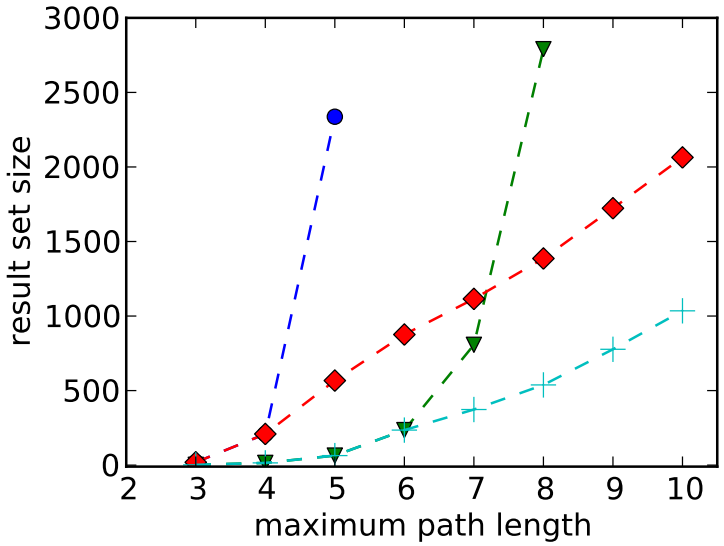

Figure 3.4: Comparison of pruning strategies. The size of results computed on Glycolysis pathway as the maximum path length increases. **Blue**: No SCL pruning. Only presence and absence of RPair is used. **Red**: Pruning of path exploration using minimum SCL on the pathway (see Methods). **Green**: Pruning of path extension using SCL_{self}. **Cyan**: Combination of both path exploration pruning and path extension pruning. In all cases, the reference pathway ranks the second in the result set (see Table 3.1).

reaction. Indeed, the lack of smoothness of the curves indicates a poor correlation between the node degree and the SCL (similarly shown by [186]). Hub deletion is known to suffer from several issues [27, 183], one of which is the coarse-grainedness in the sense that connections are pruned by deleting metabolites as well as all their connections. Accordingly, we have observed that not all the connections of these hub metabolites are of low SCL values.

Along the same line, some metabolites, defined in an *ad-hoc* fashion, and referred to as *pool metabolites* [192] or *currency metabolites* [184], which largely coincide with the hub metabolites [42] are often defined for network pruning. A widely used example is ATP [190, 27]. Although ATP in many cases serves as a carrier of phosphate groups and an energetic driver of reactions, it is also actively involved in the mass circulation of nucleotide metabolism. The versatility of ATP can be demonstrated by a *signature*

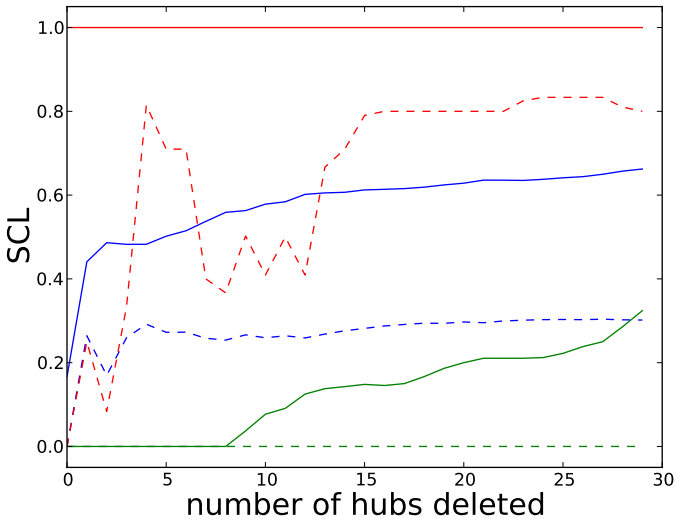

Figure 3.5: Change in SCL values as more hubs are deleted by decreasing order of their degrees. Solid line: presence in the pruned graph. Dashed line: absence from the pruned graph. Red line: upper quartile (75% in SCL). Blue line: median (50% in SCL). Green line: lower quartile (25% in SCL).

based on SCL_{self} and SCL_{other} (Fig. 3.6). The functionality of serving as a carrier of small chemical moieties is reflected in the signature by two groups of dots, one in the top-left corner and the other in the bottom-right corner. The others correspond to other functions of ATP. Some of these involve a high SCL_{other} value—an indication of contribution to the mass circulation. For the same reason, we see in Fig. 3.5 that as more and more hubs are deleted from the graph, more connections of high SCL values are eliminated (the increase in the dashed curves) as well.

This suggests that a graph with high SCL values is hard to obtain without losing important connectivity information of the graph. From Fig. 3.6, we also observe that some pool metabolites serve multiple functions (as indicated by multiple dots in the signature; e.g., ATP and Pyruvate) while others are functionally specific (as indicated by a single dot in the signature; e.g., NAD^+ and H_2O). Compounds that are usually released as part of other bulk metabolites have dot(s) only in the bottom-left corner of their signatures (e.g., H_2O, CO_2 and NH_3).

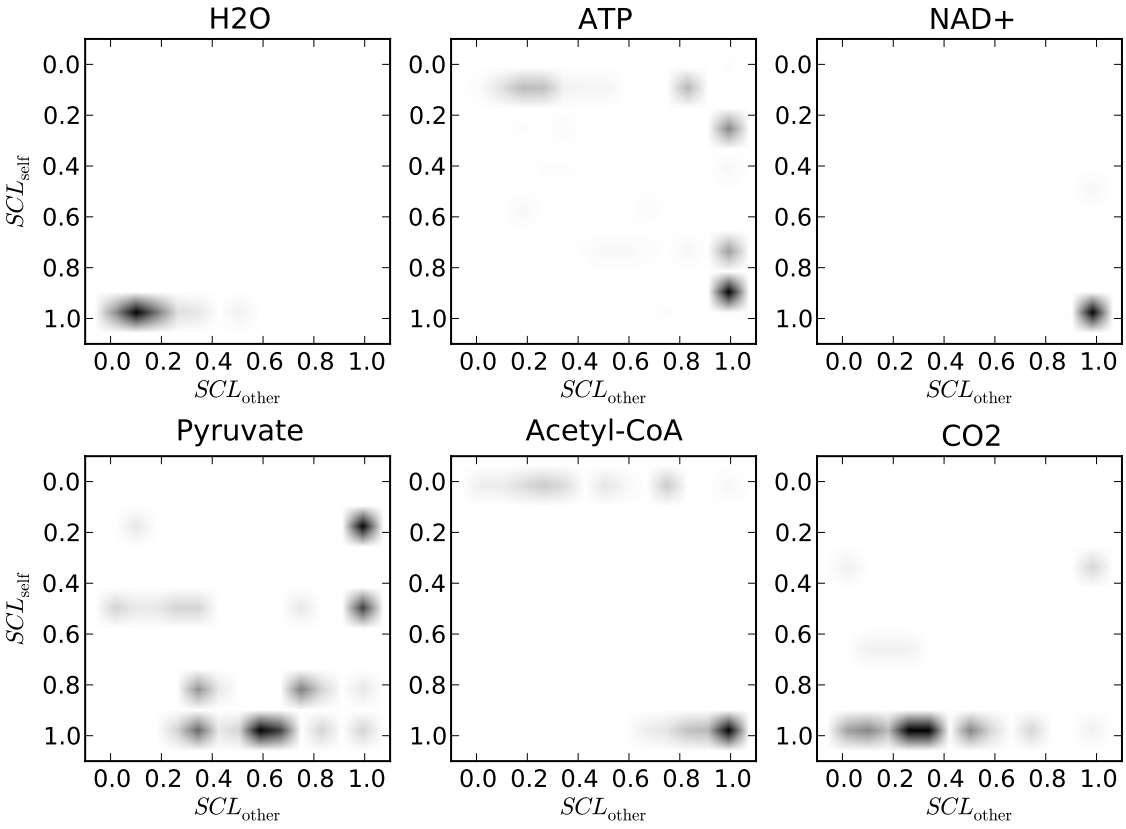

Figure 3.6: The SCL signature for six metabolites: H_2O, ATP, NAD^+, Pyruvate, Acetyl-CoA, and CO_2. The darker the dot for a metabolite C, the more reactions exist in E.coli's metabolic network with connections of $(SCL_{\text{other}}(C), SCL_{\text{self}}(C))$ combination.

3.3.3.2 Manual curation

In addition to being labor-intensive and error-prone, we show that pruning by manual curation [184, 27, 187] may also be ambiguous. Consider reaction (3.1) above. The reactant pair N-Acetylornithine and N-Acetyl-L-glutamate is missing because the acetyl group is not sufficient to represent the link between the two compounds. Now, consider reaction

$$\text{Acetyl-CoA} + \text{Formate} \Leftrightarrow \text{CoA} + \text{Pyruvate} \tag{3.3}$$

In this case, Acetyl-CoA loses the Acetyl group to Formate to form Pyruvate. The question is: should we eliminate the connection between Acetyl-CoA and Pyruvate? In this case, the acetyl group is important, as 2/3 of the carbon backbone of Pyruvate comes from it. Hence, this connection should be kept in the network. Although reactions (3.1) and (3.3) show exactly the same mass transformation pattern ({Acetyl-Group I, Group II} on one side of the reaction and {Group I, Acetyl-Group II} on the other side), we make different decisions on whether to prune away the connection between Acetyl-Group I and Acetyl-Group II. In fact, the intrinsic nature of chemical transformations implies that the amount of chemical moieties that are conserved in any reactant pairs is arbitrary (Fig. 3.1). A combination of the SCL pruning criterion with the objective quantification of chemical content can help ameliorate this problem.

3.3.3.3 RPair types

Pruning methods also include filtering specific class(es) of reactant pairs inside a reaction [186]. The KEGG RPair database provides such classification by assigning reactant pairs to five different categories: "main", "cofac", "trans", "ligase", and "leave" [193]. The categorization is reaction-dependent: one reactant pair may be of different types in different reactions. The typing is based on the classification of the enzymes (e.g, oxidoreductase, transferase, etc.) that catalyze the reaction and the role of the reactant pair in the reaction. However, the five types are manually curated, thus resulting in the same problems as discussed above.

In Fig. 3.7, we plotted the distribution of SCL values of reactant pairs in the five categories. Reactant pairs of "main" and "cofac" tend to have higher SCL values, while reactant pairs of "leave" and "trans" tend to have lower SCL values. Pathway connections are composed of reactant pairs belonging to different categories, although the dominant categories are "main" and "trans" (Fig. 3.7). This is in agreement with

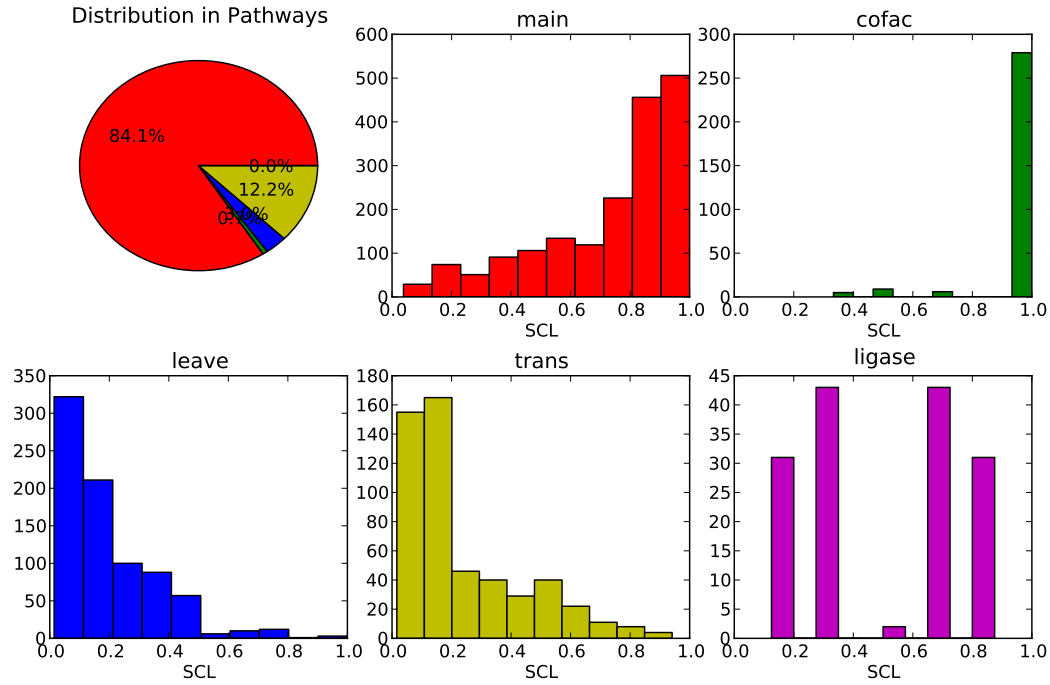

Figure 3.7: Distribution of SCL values of the five RPair types. The pie chart shows the distribution of RPair types in the connections that are annotated in KEGG reference pathway. Colors in the pie chart correspond to the colors in the histograms. The histograms shows for a given RPair type its distribution of SCL values based on all connections that are annotated with RPairs in the metabolic network of *E.coli*.

the previous practice of using only these two types for metabolic network assembly [186]. Reactant pairs of type "ligase" have a symmetric distribution, since this type consists of reactions where a large piece of chemical compound is decomposed into two components giving rise to two reactant pairs complementary in SCL values. For example, reaction [ATP + Deamino-NAD$^+$ + NH3 $\Leftrightarrow$ AMP + Diphosphate + NAD$^+$] gives rise to both ATP-AMP (high SCL) and ATP-Diphosphate (low SCL) as reactant pairs of type "ligase".

3.4 Conclusion

In this paper, we introduced the strength of chemical linkage, or SCL, as a criterion for pruning metabolic graphs. The use of the conserved chemical content according

to the actual reaction mechanism in this work is in contrast to the previous approach where only structural similarity between the two reactants is considered [194]. Although we also use the RPair database, unlike [186], whose pruning depends on the presence/absence of RPairs or the certain classes of reactant pairs annotated from KEGG based on the type of enzyme that catalyze the reaction [193], we used only the information for molecule alignments. Such information can also be obtained from other sources, such as [195].

We showed that the SCL criterion is biochemically intuitive and has power of selection for the conventional pathway connectivity when thresholded. False positive and false negative cases are caused mainly by improper quantification of chemical content as well as flaws in the data. The utility of using SCL on pruning the searching tree in pathway inference was evaluated. Biochemically meaningful pathways can be found by implementing a simple search program using the SCL criterion. Further, we compared several commonly used connectivity pruning heuristics and *ad hoc* methods, such as hub deletion and manual curation. We found that SCL values reflect the rationale behind these heuristics, yet the SCL is more objective, systematic and robust to annotation error. Many ambiguities of these heuristics are rooted in lacking an objective criterion and quantification of chemical content.

Note that although, we focus here only on graphs whose nodes are compounds, SCL can also be adapted in the assembly of networks whose nodes are reactions or reaction-derived entities such as enzyme class or genes. This is done by considering nodes for linking reactions only ones that appear in sufficiently strong reactant pairs. One potential improvement to the quantification of chemical content is to partition every chemical compound into functionally independent groups. The amount of chemical content is measured in terms of the number of such functionally independent groups, instead of the absolute number of non-hydrogen atoms. Further, the partition

of a compound into functionally independent groups is flexible yet objective, relying on at most a delineation of a set of specific pathway-related reactions. Two atoms in a compound are considered in the same group if they are linked by covalent bond(s) that does not break in all the chemical transformations under that delineation.

Chapter 4

Atom tracing and chemical graph symmetry

4.1 Introduction

The study of metabolism at the atomic level has made significant progress in the last few decades [196, 197, 190, 198]. With detailed information of atom transition for each metabolic reaction, the fate of atoms can be traced in a pathway composed of concatenated reactions [199, 200]. Besides the study of the circulation of the mass [201, 199], tracing atoms in a metabolic pathway or network has found two major applications, namely pathway inference and the simulation of isotopomer distribution for the purpose of estimating reaction fluxes. Known and novel metabolic pathways can be inferred from a metabolic network under the principle of atom economy [202], which seeks, among all the pathways that connect a source compound to a target compound, the one that maximizes the number of atoms traced [203, 186, 204]. Simulating isotopomer distribution is a task of predicting for a compound the relative abundance of its isotope-labeled forms, called isotopomers, from 1) the isotopomer distribution of a given source compound; 2) the reaction fluxes of the metabolic network and 3) the atom mapping relationship as yielded by each reaction mechanism [199]. Experimentally, information on the steady-state isotopomer distribution,

such as the mass distribution from mass spectrometry [205] or the distribution of atoms in certain microenvironments as can be detected by nuclear magnetic resonance [206, 207], can be used to derive the steady-state flux distribution. Isotopomer distribution is represented *in silico* as an Isotopomer Distribution Vector (IDV) which contains the relative abundance of each isotopomer whose location in the IDV can be coded into a binary vector with each bit projected to an atom in the compound and its value indicating whether the atom is isotopically labeled [208]. The transition from the IDV of one compound to the IDV of another compound can be summarized in a matrix called Isotopomer Mapping Matrix (IMM) whose rows are IDVs of the source compound and columns are IDVs of the target compound [208]. Computing an IMM reduces to tracing atoms from the source compound to the target compound.

Computational tracing of atoms in a reaction requires knowledge of the full atom mapping of each reaction on the pathway. Compounds in a reaction atom mapping are usually represented with atoms labeled to distinguish different atom instances. However, there are cases where atoms or compound instances are chemically indistinguishable. For example, in 4.1(A) when the σ-bond linking the central atom and the star rotates, atoms 1, 2 and 3 are replaced by atoms 2, 3 and 1, respectively, without changing the nature of the compound. Linking, for tracing purposes, multiple reactions via compounds that contain indistinguishable atoms might bring about alternative tracings (see Methods section). This equivalence in atoms or compound instances is generally referred to as symmetry.

3-D molecular symmetry has been well studied for more than half a century using the theory of point groups [209]. The main purpose of studying molecular symmetry is to elucidate chemical properties of a molecule such as selection rules in vibration spectroscopy [210]. However, there are several challenges facing a large-scale analysis that applies this theory. First, the identification of the 3-D symmetry operations is

hard to automate. The complete and accurate curation of symmetry groups and their resolution into subgroups requires *ad hoc* knowledge from the molecular point group theory. In fact, most symmetry studies have been conducted on small compounds [211]. Secondly, most reaction atom mapping data available, including manual curations such as the RPAIR data [193] and automated methods such as ones based on the Maximum Common Subgraph (MCS) heuristics [212, 213] and the ones based on the minimum graph edits [214, 215], fail to consider 3-D molecular symmetries. It is impossible to generate correct alternative atom mappings if the reaction atom mappings are not produced aware of the 3-D configuration.

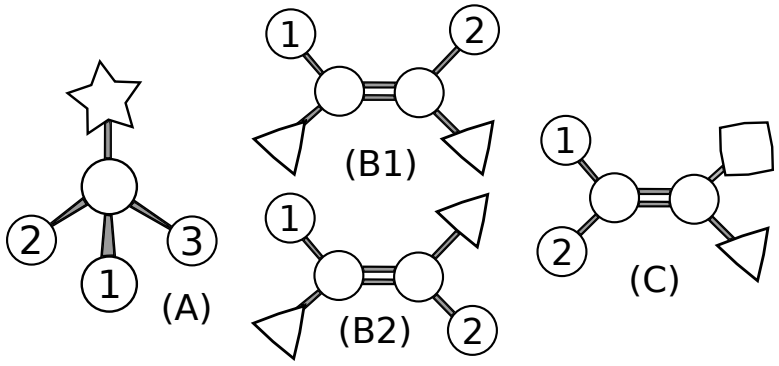

Figure 4.1: **Comparison between 3-D symmetry operations and automorphisms of the graph representation.** (A): Under graph representation, every permutation of the three labeled atoms is a valid symmetry mapping. $\{1 \to 2; 2 \to 3; 3 \to 1\}$ is a rotational symmetry operation and $\{1 \to 1; 2 \to 3; 3 \to 2\}$ is a mirror symmetry opertation. The mirror symmetry operation is not physically feasible. (B): Swapping 1 and 2 is a mirror symmetry in (B1) and rotational symmetry in (B2). However, under graph representation, (B1) and (B2) are indistinguishable. (C): atom 1 and 2 cannot be swapped. However, under graph representation, swapping 1 and 2 is a false positive symmetry.

In this work, we take a graph-theoretic approach that represents both compounds and reactions as *chemical graphs*, which are attributed relational graphs [216], and formulates the problem of the symmetry as an automorphism problem on the chemical graphs to facilitate automated computing. Although this approach fails to capture all the symmetry operations of compounds in 3-D (see 4.1 case C), it captures most

symmetry mappings as well as the group structure (see 4.1 case A and B). Its calculation can be automated and the resulting symmetries integrate well with reaction atom mapping data that do not consider 3-D features such as prochirality [217] and *cis-trans* isomerization. We developed automated methods to identify symmetries of both compounds and reactions. Symmetry-breaking atom mappings are defined and demonstrated to impact atom tracing. We find that 257 out of 1251 reactions in *E.coli* and 176 atom mappings from the KEGG RPAIR database [193] are symmetry-breaking with respect to a selection of symmetry subgroups. We devise a decomposition scheme for efficient storage of symmetry in the database. Alternative atom transitions in KEGG are explicitly listed when symmetric compounds are present. Finally, we evaluate, using metabolic pathways in the model organism *E.coli*, the impact of alternative atom tracings on pathway inference, in terms of atom economy, and isotope labeling analysis, in terms of differential IMM calculation. We regenerate atom mapping data between reactants by first augmenting the RPAIR data into full reaction atom mappings and then composing the reaction atom mappings with symmmetries on both sides of a reaction to get all the alternative atom mappings in cases where symmetries are broken by the reaction. Compared to a previous approach that considers symmetry and curates atom mapping in one organism and one specific model [196], our work can be readily extended to many organisms.

4.2 Formulation of the symmetry of chemical graphs

In chemical informatics, a chemical compound can be represented as an attributed relational graph G [216], whose nodes $V(G)$ correspond to atoms and edges $E(G)$ correspond to chemical bonds. Each node $v \in V(G)$ refers to an atom and each edge $(u, v) \in E(G)$ refers to a chemical bond linking the atoms to which nodes u and v refer. Nodes in $V(G)$ inherit attributes such as elements and isotopic status from the atoms

they refer to. Likewise, bonds in $E(G)$ inherit attributes such as bond types (single, double, aromatic, etc.). Multiple compounds, such as all compounds in a reaction, can be represented as the individual connected components in one graph. Together with graphs for a single compound, they are collectively refered to as *chemical graphs* in this study. For example, all the compounds that participate in a reaction can be represented in a chemical graph. So can all compounds produced from a reaction. In this study we call the two graphs *reactant graph* and *product graph*, respectively. Atoms in a chemical graph are labeled to distinguish their identities in the atom mapping.

An *atom mapping* from a chemical graph G to a chemical graph H is a bijection $g : V(G) \to V(H)$ such that if $g(u) = v$, then u and v refer to the same phyiscal atom identity. When defined for a reaction atom mapping, $g : V(G_r) \to V(G_p)$, where G_r and G_p are the reactant graph and product graph of the reaction. A node $u \in V(G_r)$ is mapped to $v \in V(G_p)$ if u, v refer to the same atom before and after the reaction process, assuming the chemical reaction does not incur atom-level modifications such as nuclear fission/fusion, which is true for most biochemical reactions. We denote the domain and image of a function by $Dom(\cdot)$ and $Img(\cdot)$.

A symmetry of a chemical graph is defined as an automorphism f of G (that is, a bijection $f : V(G) \to V(G)$), that preserves the chemical properties which may include, but are not limited to: 1) The atom's element: For all $p \in V(G)$, p and $f(p)$ have the same element; 2) The atom's incident bonds: For every pair of atoms p and q, there is a bond between $f(p)$ and $f(q)$ if and only if there is a bond between p and q with the same bond type. The bond types we consider here includes single, double, triple and aromatic. A nontrivial symmetry is a symmetry mapping that is not the identity mapping (where each atom maps to itself). A compound with at least one nontrivial symmetry is said to be symmetric. For simplicity, we do not consider 3-D

chemical features such as prochirality [217] in this study. We target this as future directions and discuss the issue in the Discussion and conclusion section.

Chemical graphs are graphs of bounded valence, which means that the degrees of the nodes are bounded by a small constant (the number of electrons available for forming covalent bonds). The more general isomorphism problem, the problem of finding a bijection with aforementioned properties but between two arbitrary graphs and of which automorphism is a special case, can be tested in polynomial time on graphs of bounded valence [218]. Due to the constraints imposed by the nodes' and edges' attributes, the problem is in practice not hard to solve. We devise the subroutine GraphIso(G_1, G_2) for this problem, which is a straightforward adaptation from the VF2 algorithm [219] by further testing the nodes' element and edges' bond types in the feasibility function. In other words, GraphIso takes two chemical graphs G_1 and G_2 and returns all the isomorphisms preserving the chemical properties between the two. A compound's symmetries are found by invoking GraphIso(G, G) where G is the compound's graph representation. The method takes seconds to return all the isomorphisms for all the compounds in KEGG [176], each of which may contain up to hundreds of atoms.

4.2.1 Symmetry decomposition

All the symmetry mappings of a compound form a symmetry group [1] (in the algebraic sense). The group operation is function composition and the identity is the mapping what projects every atom to itself. Enumerating and storing all the symmetry mappings explicitly can be infeasible for compounds with multiple symmetry subgroups. The number of mappings increases exponentially with the number of subgroups (as can be seen from Lagrange's theorem in group theory which states that the order

[1] A group is a set together with a defined group operation satisfying the group properties including the existence of identity, closure, associativity and etc. [220]

of the subgroup divides the order of the original group). In this study, we sought a nonexhaustive decomposition of symmetry groups. Certain chemical substructures of putative symmetry subgroups are detected from the chemical graph. They include: 1) subgroups of the form XYn where n Y atoms are covalently linked only to X, forming a permutation group; e.g., PO4, PO3, PO2, CO2, SO4, SO3, SO2, NO3, CC2, CC3, CN2, RuN4 and RuN5. Although Y atoms in some of these substructures are not all equivalent in their microenvironment (e.g., the four oxygens in PO4 are not all equivalent and cannot be freely exchanged as they appear in the graph representation), they are treated equivalently in the RPAIR data that we use to expand the reaction atom mapping, with the rationale that these bonds exchange electrons at a rate much faster than what can be captured by a GC/MS analysis (quick equilibration of labeling caused by resonance stabilization [221]). 2) Single-atom ions forming a permutation group, e.g., Na^+, Cl^-. 3) Hydration water. 4) Mirror image in benzene groups and cyclohexane groups (analogous to the σ_h operation in the 3-D symmetry theory). 5) Remaining symmetries. The remaining subgroup may be further reduced or may be irreducible. The symmetry mappings from the remaining symmetry subgroup are explicitly computed. As 4.2 shows, most data points (circles) are above the y=x line which means that the above decomposition greatly reduces the number of explicit mappings one needs to store (x-axis of 4.2) compared to the actual number of symmetries (y-axis of 4.2). To detect benzene and cyclohexane substructure, we employ the ring perception algorithm by Balducci and Pearlman [222].

4.2.2 Detecting symmetry-breaking atom mappings

Consider an atom mapping g from compound x to compound y. Compound x has a symmetry mapping f_x and compound y has a symmetry mapping f_y. In theory, one can generate all the alternative mappings from g through function composition:

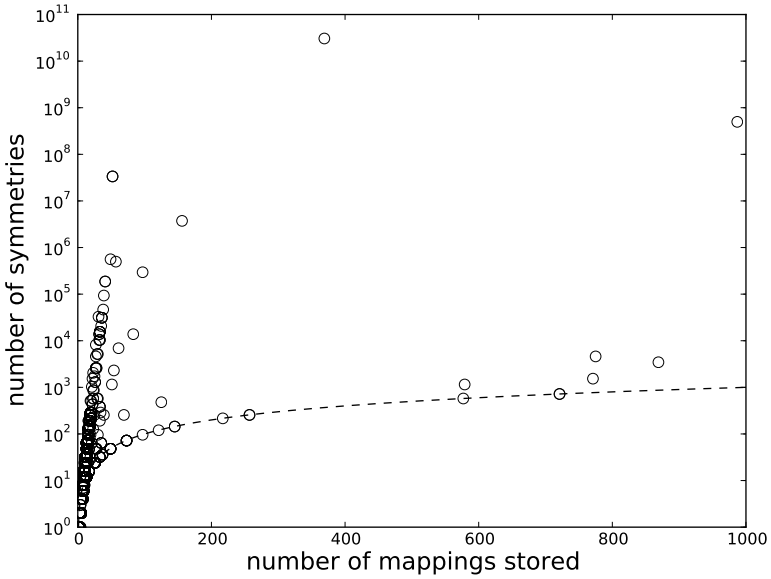

Figure 4.2: **The reduction of the number of explicit mappings stored after decomposition.** All the compounds in KEGG ligand database are plotted. The x-axis is the number of symmetry mappings explicitly stored in the database. The y-axis is the number of actual symmetries (on the log scale). The dashed line is the y=x line.

$f_x \circ g \circ f_y$. However, not all alternative mappings constructed this way can introduce alternative routes in pathway inference. The iterative composition could exponentially increase the number of alternative tracings as more and more mappings are traced (see 4.3). Only symmetries that are broken by the mapping of each reaction need to be considered. Here we give the formal definition of symmetry breaking in chemical graphs.

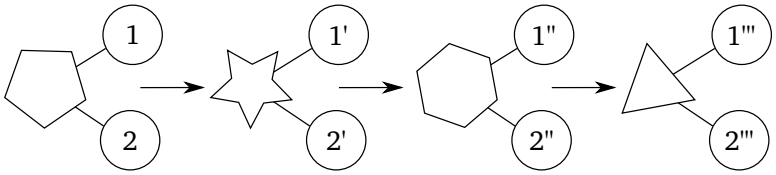

Figure 4.3: **Illustration of non-symmetry-breaking reactions.** The circles represent atoms of the same element. Tracing from the left-most compound to the right-most compound through three atom mappings gives rise to 2^3 alternative tracings (given by $g_1 \circ g_2 \circ g_3$ for $(g_1, g_2, g_3) \in \{\{1:2', 2:1'\}, \{1:1', 2:2'\}\} \times \{\{1':2'', 2':1''\}, \{1':1'', 2':2''\}\} \times \{\{1'':2''', 2'':1'''\}, \{1'':1''', 2'':2'''\}\}$).

Definition 1. *Consider two chemical graphs G, H and an atom mapping g from $V(G)$ to $V(H)$. A symmetry f on G is broken by g if either 1) $Dom(g) \neq f(Dom(g))$; or 2) there is no symmetry of H whose restriction to $Img(g)$ is $g \circ f \circ g^{-1}$.*

If the atom mapping is a complete reaction atom mapping, then $Dom(g) = V(G) = f(V(G)) = f(Dom(g))$. The atom mapping breaks the symmetry only by the second condition of Defintion 1. For example, myo-inositol (see 4.4) has 11 nontrivial symmetries, 5 by rotating the carbon ring and 6 by reflecting with respect to 6 mirror symmetry axes. Only one mirror symmetry (marked with * in 4.4) is preserved through its transformation into myo-inositol-3-phosphate. The algorithm used to find symmetry-breaking atom mapping strictly follows Definition 1.

For each atom mapping g in the RPAIR database [193], we evaluated the symmetry of both of its constituent compounds. Consider a compound x and its symmetry f_x, y is the other compound reached through the atom mapping from x. If $Dom(g) = f_x(Dom(g))$ and if there exists a symmetry f_y of compound y such that for each atom $a \in Dom(g)$, $g(f_x(a)) = f_y(g(a))$ ($g \circ f_x \circ g^{-1}$ can be embedded into f_y), the atom mapping g is said to preserve the symmetry f_x. Otherwise, g breaks f_x.

Breaking the compound/reaction symmetry in a reaction has an implication when atoms are traced across multiple reactions. In 4.5, carbon labeled $*$ in ornithine can reach carbon $*$ in agmatine via carbon $*$ in putrescine according to the annotated atom mapping in RPAIR. Since putrescine is symmetric (as shown in the mirror symmetry axis) and the symmetry is broken when it gets transformed into agmatine, carbon labeled $*$ in ornithine can as well be mapped into carbon $\diamond$ in putrescine and carbon $\diamond$ in agmantine. Therefore carbon $*$ in ornithine has alternative fates if one considers symmetry while tracing. It is easy to see that the symmetry of a reaction graph is broken only if either the permutation symmetry of its components or any of their inherent symmetry is broken.

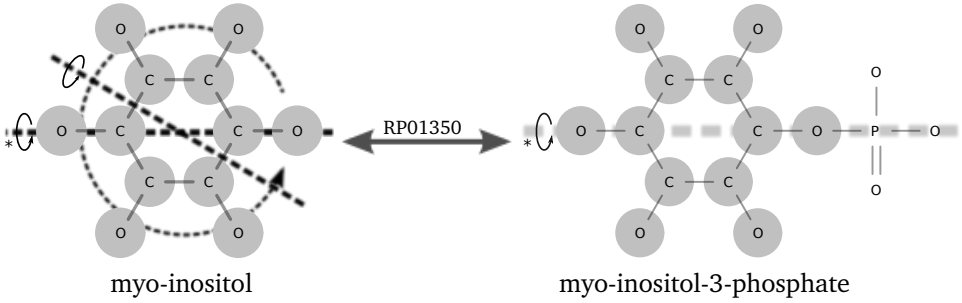

Figure 4.4: **Symmetry breaking between compounds.** Each compound is shown with its atoms' elements and labeling. Symmetry axes and operations corresponding to these symmetries are marked in dashed lines. Only two mirror symmetry axes of myo-inositol are shown. Atoms mapped in the two compounds are shaded in circles. RP01350 is the ID corresponding to the atom mapping given in RPAIR database. Atoms mapped are shaded. Note that this is not a complete reaction atom mapping (and therefore not mass-balanced) and irrelevant reactants are omitted from the figure for simplicity.

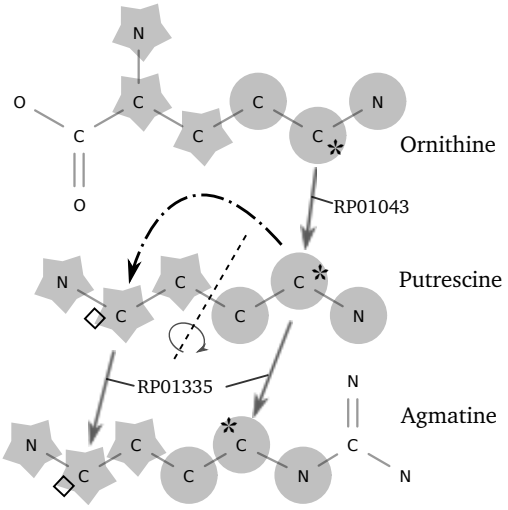

Figure 4.5: **Symmetry breaking introduces alternative atom fates.** The symmetry in putrescine brings its atoms shaded in circles to atoms shaded in stars. Atoms in ornithine and agmatine are shaded in the same way as they are in putrescine. Atoms that are not mapped to ornithine and armatine are not shaded. Dashed line corresponds to the symmetry mapping from C4 to C3 in putrescine. Note that these are not complete reaction atom mappings (and therefore not mass-balanced) and irrelevant reactants are omitted from the figure for simplicity.

4.2.3 Chemical graph standardization

Because of inconsistency in the representations of compounds in the KEGG RPAIR database (multiple labeling for the same compound in different RPAIRs), compound graphs are first standardized to the representation used in KEGG LIGAND database [176]. We identified 394 inconsistencies in compound representations from the KEGG RPAIR database. These representations are standardized by running GraphIso(G_R, G_L) (where G_R is the representation used in RPAIR and G_L is the (standard) representation used in LIGAND) and finding a graph isomorphism (an arbitrary one if there are many) preserving chemical facts. Four compounds contain irreconcilable representations which are not amenable to standardization because of discrepancy in the compound structure. Three such compounds (C02419, C03624 and C06185) are manually standardized.

4.3 Symmetry in metabolic networks

4.3.1 Compound symmetry

We study the symmetry of the entire reaction by first considering the inherent symmetry of each reactant. 14066 compounds in the KEGG COMPOUND database [176] with a defined representation were analyzed for symmetry. We identified 9131 compounds in the KEGG LIGAND database that possess nontrivial symmetry mappings assembled from all subgroups and 5912 compounds with nontrivial symmetry mappings assembled from the chosen subgroups (see Methods). They include the most studied examples of rotationally-symmetric compounds such as fumarate and succinate [221]. In 4.6, we plot the distribution of the number of symmetry mappings in the entire KEGG COMPOUND database and in assembled networks of the microbial organism *E.coli*. 4.1 tabulates some exceptionally highly symmetric compounds (with

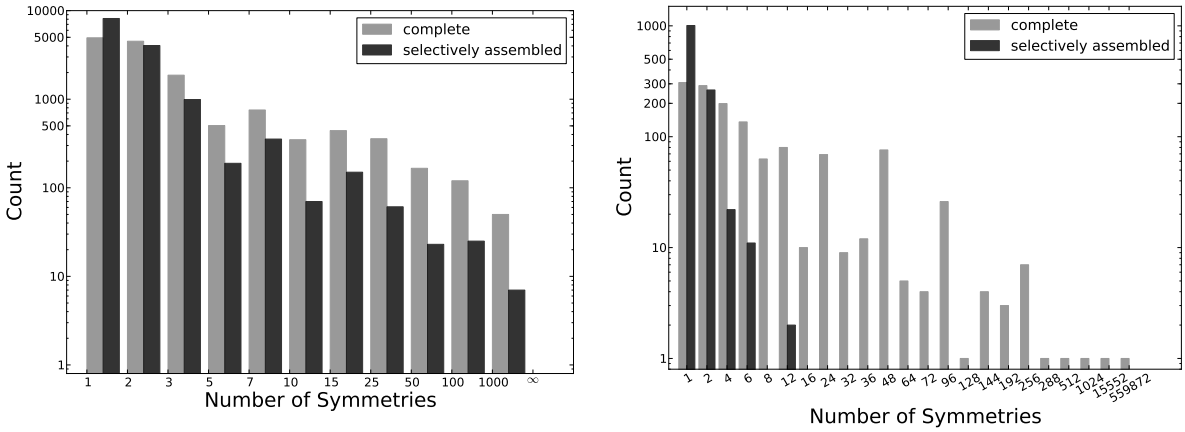

Figure 4.6: **Distribution of symmetric compounds.** (Left) All compounds in KEGG LIGAND; (Right) Compounds in assembled E.coli metabolism; Bars heights are in log scale. Number of symmetries includes the identity mappings, meaning that compounds with one symmetry are essentially asymmetric. Selectively assembled symmetries exclude non-carbon subgroups (using only CC2, CC3, Benzene Mirror and the remaining subgroup, see Methods section.)

more than 1000 symmetry mappings found in KEGG). We see that most compounds have a small number of symmetry mappings. Most highly symmetric compounds are drugs or non-biomolecules (see 4.1) and do not appear in the *E.coli* metabolism (see 4.6).

Carbon and oxygen are the two mostly exchanged elements in a symmetry mapping (shown in 4.7). Next are nitrogen and chloride. Seletive assembly reduces symmetries involving oxygen, nitrogen and sodium most significantly.

4.3.2 Reaction symmetry

Besides inherent compound symmetry, symmetry of atoms in a reactant/product graph also results from multiple occurrences of the same reactant or product. This is referred to as non-1-0 stoichiometry, which RPAIR data fails to handle properly [204]. Using pruned symmetries subgroups of compounds, we explicitly enumerated all the reaction symmetries originated from inherent compound symmetry as well as

ID	common name	classification	#symmetries	#selectively assembled symmetries
C00374	heparin	drug	31104	1
C00925	Heparan sulfate	drug	31104	1
C01204	myo-Inositol hexakisphosphate	phytic acid	559872	12
C06042	lipoteichoic acid	polymer	33554432	1
C06043	D-Alanyl-lipoteichoic acid	polymer	33554432	1
C07373	Probucol	drug	20736	20736
C07974	Suramin	drug	93312	2
C11174	1-Diphosinositol pentakisphosphate	metabolite	186624	2
C11526	5-Diphosphoinositol pentakisphosphate	metabolite	186624	2
C12933	Rolitetracycline nitrate	drug (antibiotic)	13824	64
C13523	secretin	polymer	32768	512
C13553	Collistin sodium methanesulfonate	drug (antibiotic)	3732480	4
C13704	TX-TM-calixarene	polymer	10368	8
C13723	Dextran Sulfate	polymer	46656	1
C13768	Collistin Sulfate	drug (antibiotic)	30576476160	3840
C13932	Ruthenium Red	dye	497664000	2
C14287	4,4'-Methylenebis(2,6-di-tert-butylphenol)	curing agent	10368	10368
C15435	Fenbutatin oxide	pesticide	294912	294912
C15990	Spheroidene	metabolite	15552	2
C15991	Myo-inositol pentakisphosphate	metabolite	15552	2
C16001	Reactive Black 5	dye	497664	4
C17142	[Heparan sulfate]-N-sulfoglucosamine	drug	31104	1

Table 4.1: **Highly symmetric compounds in the KEGG LIGAND database.** Selectively assembled symmetries exclude non-carbon subgroups (using only CC2, CC3, Benzene Mirror and the remaining subgroup).

from permutating compounds with non-1-0 stoichiometry. 20 reactions from KEGG database are intractable due to high stoichiometry which leads to the difficulty in enumerating all the permutations. They are R00918, R05185, R05464, R06448, R06453, R06458, R06459, R06480, R06481, R06482, R06483, R06635, R06636, R06637, R06641,

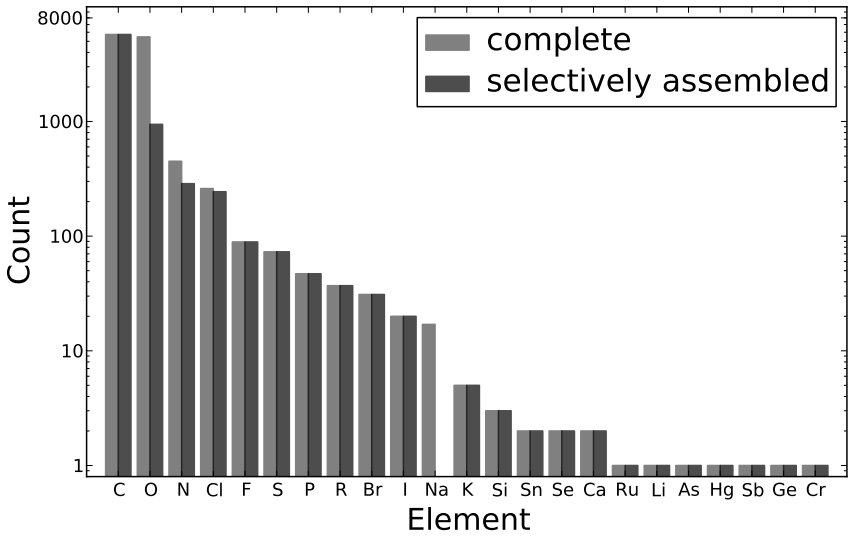

Figure 4.7: **Element composition of symmetry mappings.** The height of each bar is the number of symmetry mappings that alter at least one atom of the element. Selectively assembled symmetries have reduced composition in oxygen, nitrogen and sodium. Selectively assembled symmetries exclude non-carbon subgroups (using only CC2, CC3, Benzene Mirror and the remaining subgroup, see Methods section).

R06643, R06644, R06645, R07251 and R08649. And these reactions are lumped representations of multiple elementary reaction steps and do not appear in the assembled *E.coli* metabolic network. Out of 8163 KEGG reactions that are amenable to explicit symmetry enumeration, 7194 reaction sides contain non-trivial symmetries (every reaction has two sides). When restricted onto the *E.coli* metabolic network, we found that out of 5647 cases where a compound appears in a reaction, only 149 ($\sim 2.6\%$) appear more than once. In 121 of these 149 cases, compounds participate exactly twice. In only 27 cases, compounds participate more than twice. Out of these non-1-0 stoichiometry compounds, even fewer contain inherent symmetry (see the left panel of 4.8), indicating rare cases where the total number of symmetries is attributed to both non-1-0-stoichiometry and inherent compound structures. Out of 1398 metabolic reactions in *E.coli*, 539 of the reactions have nontrivial symmetries on either side. Among them, 77 reactions have nontrivial symmetry from non-1-0 sto-

ichiometry. 494 reactions inherit their symmetries from inherent symmetries of the component reactants or products. 32 reactions have symmetries coming from both sources. The distribution of the number of symmetry mappings in all reactant and product graphs of all reactions in *E.coli* has a power-law shape, as shown in the right panel of 4.8.

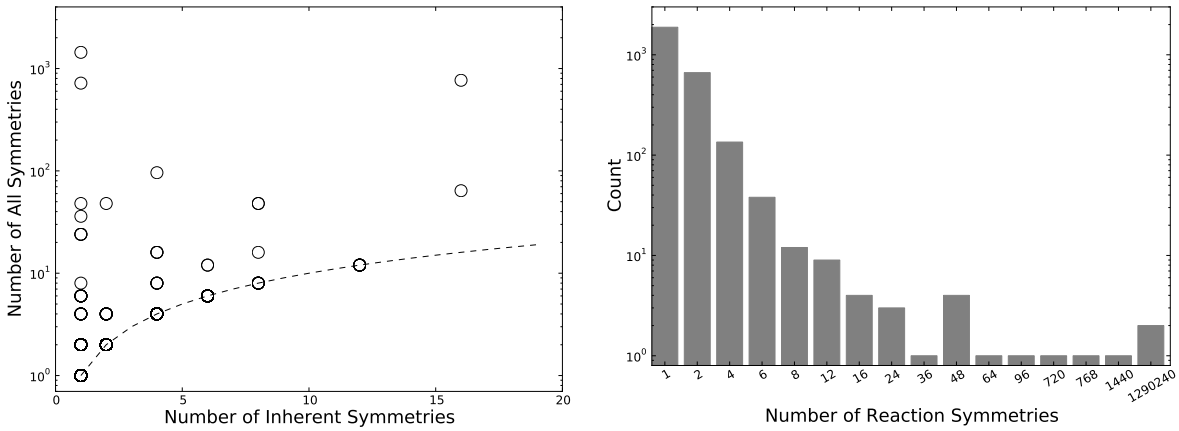

Figure 4.8: **Distribution of reaction symmetries.** (Left) An increase in the number of symmetries from non-1-0-stoichiometry. The dashed line corresponds to $x = y$, that is, symmetries come solely from inherent structures of the compounds. (Right) Distribution of the number of symmetric forms of both reactant graphs and product graphs of all reactions in the metabolic network of *E.coli*.

4.3.2.1 Decomposition of reaction symmetries

The symmetry of each reactant graph and product graph of a reaction is obtained by composing the inherent symmetries of their constituent compounds and the permutation of those compounds (see 4.9 for an illustration). For inherent symmetries, only symmetries composed from subgroups CC2, CC3, benezene and cyclohexane mirror and remaining symmetries are considered since the other symmetry subgroups do not involve carbon, which is of utmost interest in tasks such as the prediction of the carbon fate and carbon-13 simulation [199]. Such selective assembly of the compound symmetry greatly reduces the amount of space needed for the explicit storage of the

symmetries of reactant graphs and product graphs of all the metabolic reactions in *E.coli*, facilitating the detection of symmetry breaking reactions (shown below). The only exception is reaction R06447 which contains 1290240 symmetry mappings. The number of symmetries of a reactant graph or product graph can be calculated as

$$\prod_i s_i! \times c_i^{s_i}$$

where s_i is the stochiometric coefficient of the i^{th} compound in the chemical graph and c_i is the number of inherent symmetries of the i^{th} compound.

permutation:
P_0 {1:1;2:2;3:3;4:4;5:5}
P_1 {1:3;2:4;3:1;4:2;5:5}

Inherent:
I_0 {1:1;2:2;3:3;4:4;5:5}
I_1 {1:1;2:2;3:4;4:3;5:5}
I_2 {1:2;2:1;3:3;4:4;5:5}
I_3 {1:2;2:1;3:4;4:3;5:5}

composition:
$P_0 \circ I_0$ {1:1;2:2;3:3;4:4;5:5}
$P_0 \circ I_1$ {1:1;2:2;3:4;4:3;5:5}
$P_0 \circ I_2$ {1:2;2:1;3:3;4:4;5:5}
$P_0 \circ I_3$ {1:2;2:1;3:4;4:3;5:5}
$P_1 \circ I_0$ {1:3;2:4;3:1;4:2;5:5}
$P_1 \circ I_1$ {1:4;2:3;3:1;4:2;5:5}
$P_1 \circ I_2$ {1:3;2:4;3:2;4:1;5:5}
$P_1 \circ I_3$ {1:4;2:3;3:2;4:1;5:5}

Figure 4.9: **Illustration of calculating reaction symmetries on a graph with three compounds.** Two compounds are of the same type and composed of two bonded atoms of the same element. Each circle represents an atom. The *permutation* mapping (labeled as P_i, i being the index) and the *inherent* symmetry mapping (labeled as I_i, i being the index) of each component compounds are listed in curly braces. The enumeration of the final reaction symmetry mappings (the right column) are realized by iteratively composing each permutation mapping with each inherent mapping. Function composition of two functions P_i and I_j is defined as a new function $P_i \circ I_j$: $x \rightarrow I_j(P_i(x))$ for each $x \in Dom(P_i)$ given $Img(P_i) = Dom(I_j)$.

4.3.3 Symmetry-breaking reactions

In studying symmetry breaking reactions, we studied two classes of atom mappings, atom mappings coming from entire reactions and atom mappings restricted to the

atom transition between two compounds (as is the case of the RPAIR data).

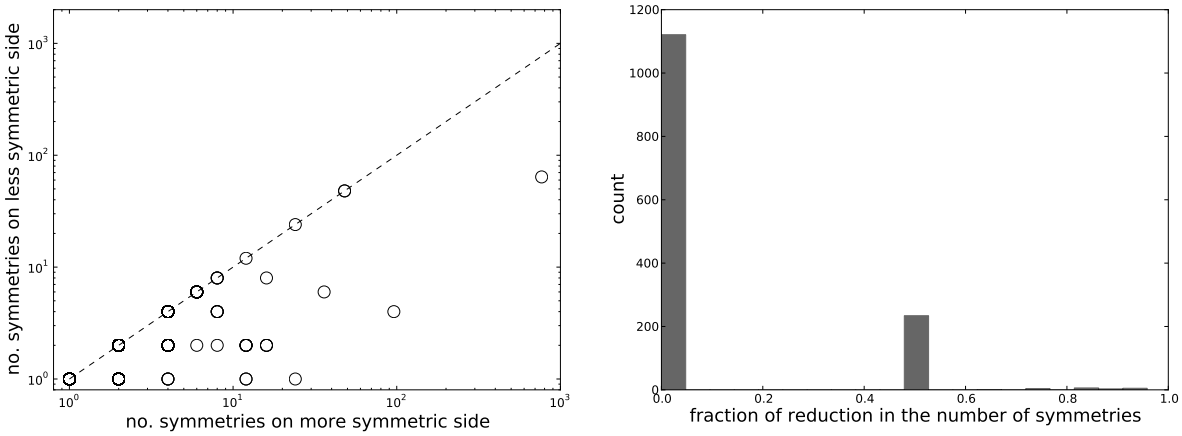

Figure 4.10: **Distribution of symmetry-breaking reactions in the** *E.coli* **metabolic network.** (Left) Only reactions with all reactants structually defined is plotted. One dot on the plot can contain more than 1 reactions. (Right) Distribution of the reduction in the number of symmetry mappings of reactions in the metabolic network of *E.coli*. The relative reduction is calculated by normalizing the difference between the number of symmetric forms of reactant graph and of product graph by the larger of the two. Symmetry mappings exclude non-carbon subgroups (using only CC2, CC3, Benzene Mirror and the remaining subgroup, see Methods section. Note that having equal numbers of symmetries on the two sides of a reaction does not exclude the possibility of symmetry-breaking.

We first look at the number of symmetries on different sides of the reaction. In 4.10, we observe that a substantial number of reactions have discrepant numbers of symmetries on the two sides. This signifies the breaking and emergence of symmetries. Based on the source of the symmetry the atom mappings break, we classify all the symmetry breaking reactions into two types. Type 1 involves the breaking of symmetries from non-0-1 stoichiometry. Type 2 involves the breaking of reaction symmetries that are inherent to the participating compounds. The inherent symmetries are assembled from compound symmetries excluding non-carbon subgroups (using only CC2, CC3, Benzene Mirror and the remaining subgroup). There are 1398 metabolic reactions in *E.coli*. Out of 1257 reactions with defined atom mappings, we found 56 reactions having type 1 symmetry breaking of either the reactant or the

product graph (12 of which have type 1 symmetry breaking on both sides). For type 2 symmetry breaking, 217 reactions are identified and 15 of them are broken from both sides of the reaction. 257 reactions have symmetry breaking of either types. 16 of them have symmetry breaking of both types.

4.3.3.1 Computing whole reaction atom mappings by minimizing the graph edit distance

In order to trace atoms across metabolic reactions, we need the knowledge of atom mapping for each metabolic reaction. The whole reaction atom mappings are found in two steps. For each reaction, we aimed at finding only one valid reaction atom mapping. Subsequent symmetry analysis will give all the possible reaction atom mappings. This is done by first obtaining a partial mapping for a reaction via composing all the associated RPAIRs (subroutine GreedySeed, elaborated below) and then expanding the partial mapping to a complete reaction atom mapping by minimizing the graph edit distance based on a cost assignment [215]. In this way, we both take advantage of the manual curation of the reactant pair mapping from the RPAIR database [193] and the power of the automated reaction atom mapping analysis by the criterion of minimum graph edit distance. The method only fails (in minutes) to yield a atom mapping for reaction R06447 because of its high graph edit distance. The resulting reaction atom mappings are manually inspected for correctness.

In GreedySeed, we first find the set Θ of all the possible atom mappings that are derived from any of the RPAIRs linked to the reaction in either direction. Each atom mapping $\theta \in \Theta$ maps atoms in a substrate to atoms in a product. We make use of a subroutine SetMapping that takes as input a partial mapping m, augments it by the mappings in Θ and returns the new partial mapping that is expanded. We verify that the domain of θ has no overlap with the domain of m before the augmentation.

Algorithm 2: GreedySeed

input : Θ the set of all mapping derived from RPAIRs linked to a given reaction.
output: a partial mapping of the reaction

```
1  Q = an empty priority queue of partial mappings prioritized by the number of
   atoms mapped;
2  foreach θ ∈ Θ do
3  |  m = an empty partial mapping;
4  |  m = SetMapping(m, θ);
5  |  push m into Q;
6  end
7  expandable = True;
8  while expandable do
9  |  m = first element in Q;
10 |  expandable=False;
11 |  foreach θ ∈ Θ do
12 |  |  if θ does not contradict m then
13 |  |  |  m' = SetMapping(m, θ);
14 |  |  |  push m' into Q;
15 |  |  end
16 |  end
17 end
18 return first element in Q;
```

4.4 The regeneration of RPair atom mapping

Several studies have depended on the KEGG's RPAIR database for inferring metabolic pathways [203, 186, 204]. But due to the lack of explicit incorporation of symmetry-induced alternative mappings, RPAIR data is insufficient for the task of pathway inference [196].

In this study, atom mappings in RPAIR database are first checked to see whether they break the computed symmetries of the constituent compounds. Out of 2369 reactant pairs that are involved in reactions from the *E.coli* metabolome, we identified 176 reactant pairs that break the symmetry of at least one of the two constituent compounds. The identification is based on the definition of symmetry breaking of compound graphs (see Methods). 257 reactant pairs are added to the data of KEGG RPAIR database by composing the compound(s) symmetry that each rectant pair breaks with the atom mapping from the reactant pair itself.

Note that in some cases where stoichiometry is higher than 1, RPAIR database already provides multiple atom mappings for the reaction if these atom mappings are different (e.g., R00006 has RP00440 and RP12733 for pyruvate). But due to the incompleteness of RPAIR database in covering the reaction atom mappings, we regenerated the atom mapping between every pair of reactants in all the KEGG reactions except the 20 reactions with high stoichiometries and infeasible to explicit enumeration of all their symmetry mappings (see Reaction symmetries section). We built 15196 pairwise compounds atom mapping data from 6809 reactions. The pairwise reactant atom mapping data regenerated not only covers all the atoms in each reactions, but also has taken into consideration of the reaction symmetry.

4.5 Impact of alternative tracing on atom economy and isotopomer distribution vector

Taking each of the 1261 metabolites from E.coli network as the source, we repeated 200 times the tabu search (see Methods) and collected, upon encountering of a symmetry-breaking reaction, the information on whether it gives rise to any of the three consequences: differential in IMM, differntial in conservation and differential in size (see Methods). We plotted the number of sources from which any of the 200 searches within a prespecified maximum length (the search scope) would result in alternative tracings differential in IMM, conservation and conservation size respectively (see 4.11). We observe that as the search scope increases, so does the chance of encountering alternative tracings with any of the three consequences. The increase in the probability of seeing any alternative tracing plateaus after the maximum path length exceeds 5. For more than 350 metabolites out of 1261 ($\sim 27.8\%$), at least one search out of 200 random paths of length higher than 5 yielded alternative tracings differential in IMM. More than 200 ($\sim 15.9\%$) of them are differential in conservation, and more than 100 ($\sim 7.9\%$) exhibit difference in the size of the atom set conserved. In fact, any difference in atom tracing would be reflected in the Atom Mapping Matrix (AMM), hence changing Isotopomer Mapping Matrix (IMM). Indeed, in every case where an alternative tracing arises, there is a difference in IMM.

4.5.1 Tabu search for evaluating the impact of alternative tracings

We devised a simple tabu search algorithm (Algorithm 3 below) to investigate the impact of alternative tracings by conducting a random walk on the metabolic network from a given source metabolite. Three consequences of alternative tracing are

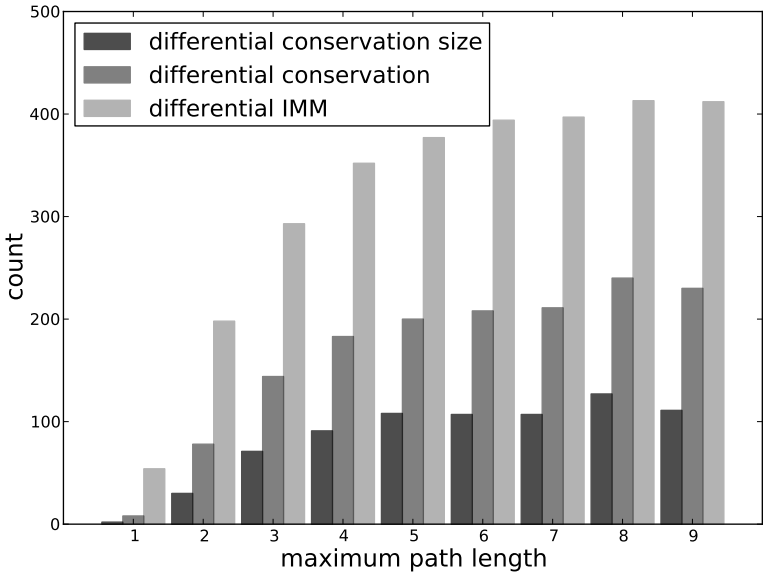

Figure 4.11: **Distribution of the three consequences of alternative tracings in tabued random walks of the metabolite network of *E.coli*.**

investigated along the walk.

1. Differential in Isotopomer Mapping Matrix (IMM)

   This is the case when source IDVs are traced to different target IDVs. In other words, under different atom mapping matrices, the isotopomer mapping matrices are different. For example, consider atoms 1, 2, 3 in the source metabolite and atoms 7, 8, 9 in the target metabolite; mapping $\{1 \Rightarrow 7, 2 \Rightarrow 8, 3 \Rightarrow 9\}$ and mapping $\{1 \Rightarrow 8, 2 \Rightarrow 7, 3 \Rightarrow 9\}$ are two alternative tracings that will give rise to differential IDVs. For instance, a source IDV with only atom 1 labeled will be mapped to a target IDV with either atom 7 labeled or atom 8 labeled depending on the tracing.

2. Differential in conservation

   This is the case when source atoms are traced to different sets of target atoms. For example, mapping $\{1 \Rightarrow 7, 2 \Rightarrow 8, 3 \Rightarrow 10\}$ and mapping $\{1 \Rightarrow 8, 2 \Rightarrow 7, 3 \Rightarrow 9\}$ are two alternative tracings that will give rise to differential conser-

vation, namely set $\{7, 8, 10\}$ and set $\{7, 8, 9\}$. However, the previous example raised for differential IDV is not differential in conservation (both results in set {7, 8, 9}). It is easy to see that every pair of alternative tracings differential in conservation is differential in IMM.

3. Differential in conservation size

   This is the case when source atoms are traced to different numbers of target atoms. For example, mapping $\{1 \Rightarrow 7, 2 \Rightarrow 8\}$ and mapping $\{1 \Rightarrow 8, 2 \Rightarrow 7, 3 \Rightarrow 9\}$ are differential in conservation size with the former conserving 2 atoms with the source and the later conserving 3 atoms with the source. Differential in conservation size has an implication in calculating the atom economy and thus affects pathway inferences based on such criterion. Likewise, every pair of alternative tracings differential in conservation size are differential in conservation.

In this algorithm, a priority queue of so-called *pathway states* is kept. Each pathway state belongs to a metabolite and specifies the set of its atoms conserved along the pathway to the source. Pathway states are sorted by the size of conservation. Initially, the only pathway state in the queue is the chosen source metabolite and the set of all its atoms. In each round, a pathway state is popped from the queue and tracked through all the reactions in which the metabolite is involved. New pathway state spawned from the tracking is pushed into the queue only when its conservation cannot be included in any of the existing pathway states of the metabolite to whom the tracking leads.

We first describe an auxiliary subroutine ReactionTrack, which takes a reaction r, its atom mapping m_r, a set s of atoms and a tracking direction d as inputs and returns a tracking result. The tracking result γ is a list containing ordered pairs (m, ψ), where m is a metabolite that a subset of s is tracked into and ψ is a mapping that projects the atom(s) tracked to atom(s) in the source metabolite. For example,

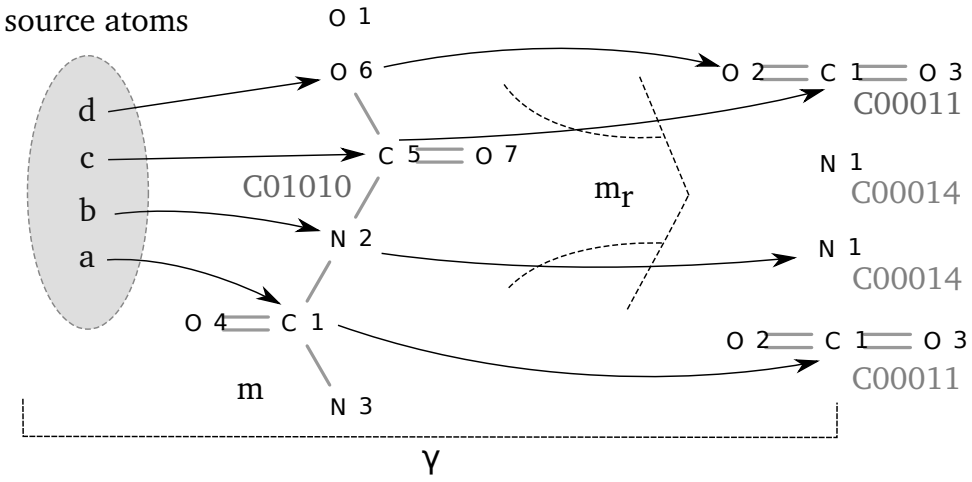

Figure 4.12: **Schematic illustration of a tracking result.** Hydrogens and their associated bonds are not shown.

consider reaction R00005 (C01010+C00001 $\Leftrightarrow$ 2C00011+2C00014, see 4.12), a pathway state of metabolite C01010 with source projection $\{1 \Rightarrow a, 2 \Rightarrow b, 5 \Rightarrow c, 6 \Rightarrow d\}$ (suppose a, b, c, d are source atoms) is tracked through the reaction and the tracking direction d is from left to right. According to one of the reaction mechanisms, atom 1 maps to atom 1 in C00011, atom 2 maps to atom 1 in C00014, atoms 5 and 6 map to atoms 1 and 2 in the second C00011. The resulting list is $\gamma = [(\text{C00011}, \{1 \Rightarrow a\}), (\text{C00014}, \{1 \Rightarrow b\}), (\text{C00011}, \{1 \Rightarrow c, 2 \Rightarrow d\})]$. Note that in γ one metabolite can appear multiple times with different projections to the source if its stoichiometry in the reaction is higher than 1 (as is the case for C00011 in the example). Three other auxiliary subroutines CheckDiffIMM (Algorithm 4), CheckDiffConserv (Algorithm 5) and CheckDiffConservSize (Algorithm 6) test if a set Γ of tracking results is differential in IMM, conservation, or conservation size, respectively.

4.6 Discussion and conclusion

Although previous studies on pathway inference largely ignored the impact of symmetry in compound and reaction [223], several automated methods exist that are able to

Algorithm 3: RandomTabuSearch

input : maximum path length l, source metabolite m_s
output: boolean variables diff_i, diff_c, diff_s, indicating whether the tracking from m_s has given rise to alternative tracings differential in IMM, conservation or the conservation size respectively.

diff_I = False;
diff_c = False;
diff_s = False;
$m_c = m_s$;
c = the set of all atoms of m_s;
while path length $< l$ **do**
    r = a randomly chosen reaction that m_c is involved;
    m_r^0 = a reaction atom mapping of r;
    d = a randomly chosen tracking direction if the metabolite appears on both sides of r, or from the side where the metabolite appears if otherwise.;
    tracking_results: $\Gamma = \emptyset$;
    if r is symmetry-breaking from d **then**
        spawn the set of reaction atom mappings $M_r = \{m_r\}^i$ by composing m_r^0 with the reaction symmetries that are broken from d.;
        foreach $m_r^i \in M_r$ **do**
            $\Gamma = \Gamma \cup \{\text{ReactionTrack } (r, m_r^i, c, d)\}$;
        end
        if not diff_I **then**
            $\text{diff}_I = \text{CheckDiffIMM}(m_s, M_r)$;
        end
        if not diff_c **then**
            $\text{diff}_c = \text{CheckDiffConserv}(\Gamma)$;
        end
        $\text{diff}_s = \text{CheckDiffConservSize}(\Gamma)$;
        if diff_s **then**
            Break;
        end
    else
        $\Gamma = \{\text{ReactionTrack } (r, m_r^0, c, d)\}$;
    end
    m_c, c = randomly choose a metabolite and its conservation from Γ;
end
return diff_I, diff_c, diff_s;

Algorithm 4: CheckDiffIMM

input : source metabolite m_s, reaction atom mappings M_r
output: a boolean variable indicating whether the tracking from m_s has given rise to alternative tracings differential in IMM.

X = a random sample of IDVs of m_s;
foreach $x \in X$ **do**
  foreach $m_r^i,\ m_r^j \in M_r \times M_r \wedge m_r^i \neq m_r^j$ **do**
    if $m_r^i(x) \neq m_r^j(x)$ **then**
      return True;
    end
  end
end
return False;

Algorithm 5: CheckDiffConserv

input : a list of tracking results Γ
output: a boolean variable indicating whether the tracking from m_s yields alternative tracings differential in conservation.

$S = \emptyset$;
diff_c = False;
foreach $\gamma \in \Gamma$ **do**
  $s = \emptyset$;
  foreach $(m, \psi) \in \gamma$ **do**
    $s = s \cup \{(m, Dom(\psi))\}$
  end
  if $s \notin S \wedge S \neq \emptyset$ **then**
    diff_c = True;
    Break;
  else
    $S = S \cup \{s\}$;
  end
end
return diff_c;

Algorithm 6: CheckDiffConservSize

input : a list of tracking results Γ
output: boolean variable diff_s indicating whether the tracking from m_s yields alternative tracings differential in conservation size.

```
S = ∅;
diff_s = False;
foreach γ ∈ Γ do
    s = ∅;
    foreach (m, ψ) ∈ γ do
        s = s ∪ {(m, |Dom(ψ)|)}
    end
    if s ∉ S ∧ S ≠ ∅ then
        diff_s = True;
        Break;
    else
        S = S ∪ {s};
    end
end
return diff_s;
```

generate alternative mapping directly from scratch [215, 214, 196, 221]. BioNetGen™ provides a method that makes proper corrections for symmetries in counting molecule observables [224]. Antoniewicz *et al.* [221] provided a method for computing Elementary Metabolite Units (EMU) considering equivalent atoms without giving a solution to how these equivalent atoms can be identified. Algorithms designed by Heinonen *et al.* [215] as well as Crabtree and Mehta [214] can compute alternative reaction atom mappings in the form of equally optimal solutions. However, there is no apparent extension to predict, using their algorithm, the number of equally optimal solutions, which can be prohibitively large when the number of symmetry mappings are high. Our work can supplement the automatic generation of reaction atom mapping in estimating the number of equally optimal solutions.

Ravikirthi *et al.* [196] curated the atom mappings for 2077 reactions present in a genome-wise construction of E. coli network, combining an automated method based

on heuristics such as Maximum Common Subgraph [212] with manual curation of not only symmetry but also chirality and prochirality [217]. Particularly, these authors manually treated "equivalent oxygen atoms" and "rotational symmetric molecules". The authors found 653 reactions that contained compounds with at least one kind of symmetry. For each atom mapping they curated, alternative mappings ("mapping degeneracies") are investigated. However, Maximum Common Subgraph heuristic is reported to be less accurate in returning the true reaction atom mapping [215, 214] compared to ones that minimize the number of bond break and formation. Moreover, the curation in Ravikirthi *et al.* [196] is limited to one model organism, *E.coli*, and one of its reconstructed model, iAF1260 [64]. On the contrary, the curation from RPAIR is organism-independent but does not handle non-1-0 stoichiometry properly [204] and is not complete in term of covering all the atoms in the reaction. We took advantage of this but employed a graph-theoretic, minimum graph edit based method for expanding the atom mappings from RPAIR to complete the atom mapping for each entire reaction.

In all the previously mentioned studies, symmetry was only implicitly accounted for. No symmetry mapping was computed. In this study, we explicitly computed compound and reaction symmetries by using an adapted method for finding graph automorphisms.

From the random walking experiment, we observe that symmetry-initiated alternative tracings is nonnegligible (with around 27.8% chance of emergence in the *E.coli* metabolism), and could result in miscalculation of atom economy (around 7.9% in the *E.coli* metabolism) and Isotopomer Mapping Matrices (everytime an alternative tracing is seen). This observation highlights the significance of having symmetry-awared atom mapping data when one calculates atom economy in pathway inference and computes the distribution of isotopomers.

There are some obstacles in the accurate *in silico* computation of whole reaction atom mappings. The first involves reactions that are formed from multiple elementary reactions aggregated together. On those reactions, the automated computation of atom mapping becomes infeasible when the graph edit distance is too large. Moreover, the high stoichiometry gives high symmetry from permutating the reactants of the same kind.

The second involves the curation of 3-D symmetry operations. In this work, we have ignored prochirality for simplicity. When we consider 3-D configurations, some prochiral substituents can be differentiated, leaving some symmetries detected using graph-theoretic methods invalid. In other words, symmetries that arises soly from exchanging two prochiral substituents are invalid under a 3-D point of view. We employed the following *ad-hoc* method for a preliminary detection of prochiral carbon centers. For each carbon of a compound, we studied all of its covalently linked substituents. We compared every pair of the substituents by whether they are isomorphic. When testing graph isomorphism, in each chemical graph of the substituent, we disconnect the center carbon atom from all of its neighbors except from the substituent under study. The center carbon's element attribute is uniquely relabeled to differentiate from other carbons. The number of distinct (in the graph isomorphism sense) substituents are then coded in a sorted list. For example, a carbon has the code (2,1) if it has 3 substituents, 2 of them are isomorphic and non of the two is isomorphic to the third substituent. We detect among all the KEGG compounds those that have carbons with code (2,1) or (2,1,1). These carbons are putative prochiral centers and the exchange of the two isomorphic substituents are invalid symmetries in the 3-D point of view. We find 2904 out of 14066 compounds from KEGG with such prochiral carbon center(s). These prochiral carbon centers are special cases of the more general 3-D restriction which is beyond the scope of this

paper. We identify the full resolution of 3-D symmetry as a future direction.

Thirdly, there are further uncertainties from not knowing detailed reaction mechanisms in this automated approach. These mechanisms could alter atom mappings and symmetry operations in ways such as triggering uncommon atom transitions which could be of very high graph edit distance, or having special contraints from timing and synchronization that prohibits certain symmetry operations. How much these uncertainties affect our reported results remains a question whose answering requires more detailed information on enzymes and reaction thermodynamics.

To summarize, in this study, we formulated the problem of compound and reaction symmetry as a graph automorphism problem. We explicitly computed symmetry mappings of reactions either from non-1-0 stoichiometry or inherent symmetries of the reactants/products. We motivated the concept of symmetry-breaking reactions and studied its extent. Random walk on the metabolic network revealed significant impact of alternative tracings to pathway inference and isotopomer distribution simulation. Technically, we augmented the KEGG RPAIR data by first expanding atom mappings from RPAIR and then composing symmetries that are broken to complete the whole reaction atom mapping.

Chapter 5

Modular structure

5.1 Introduction

Analyses of biological networks have revealed modular structures [225, 226, 227, 228, 229]. Parter et al. [29] found that bacterial species living in variable habitats have metabolic networks with significantly higher modularities than bacterial species living in less variable habitats. According to one explanation, since modularity promotes evolvability, enabling bacteria to quickly adapt to varying environments, having a more modular metabolic network is an evolutionarily favored trait for species living in open habitats such as soil and sea. In other words, high modularity is selected for by evolution for species living in these varying habitats (edge 1 in Fig. 5.1). The robustness of metabolic networks, a concept related to modularity [230], as measured by the maintenance of a phenotype (e.g., growth) under perturbation (e.g., mutation or gene loss), has been shown, both *in vivo* and in simulation, to have risen from fluctuating environments [231, 232]. An alternative explanation can be formulated from the other direction: because species with a higher modularity in their metabolic networks are more capable of adapting to changes in environment, they colonize a wider range of habitats, giving rise to the observation that bacteria living in varying

habitats have more modular metabolic networks (edge 2 in Fig. 5.1). In another recent study of an Archaea data set [233], such relationship between modularity and habitat variability was not found, which calls for more investigation of alternative explanations.

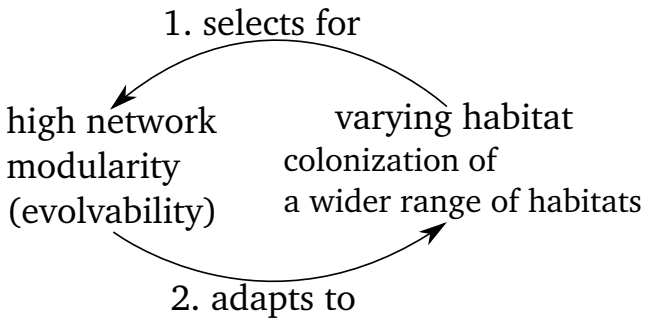

Figure 5.1: **A feedback loop between modularity and habitat variability.** Two different explanations of the association of the modularity score with the habitat variability.

Modularity as a graph-theoretic concept, when studied on biological networks, can be quantified in different ways [234, 235, 236, 237, 29, 30]. In the works of Parter et al. [29] and Kreimer et al. [30], modularity is based on the definition of Newman and Girvan [238]. This definition quantifies the extent to which the graph connectivity of a network exhibits a modular structure, that is, communities with a majority of the connections falling within, rather than across, communities. Roughly speaking, the modularity score Q [238] (see Methods), which is a quantity associated with a partition of the network, indicates how much more likely it is for an edge to be placed inside a community from that partition than would be expected from a random selection of neighbors for a node of a certain degree. The partition of nodes that gives rise to the maximum Q value is regarded as the community structure of the graph, and the score itself is taken to be the graph's modularity.

Although the modularity score depends on the community structure, similar modularity scores may arise from different community structures. It is natural to ask (and

is currently unknown) whether a specific modularity (high or low) of metabolic networks is the result of acquiring a similar community structure or of achieving different community structures. More specifically, assuming that network modularity plays an adaptive role [239], as is the case for the first explanation (Fig. 5.1), is it the modularity score that confers higher fitness regardless of the community structure giving rise to it, or is it the community structure that is the unit of selection and modularity is conserved only as a consequence? If modularity is achieved via similar community structures, it might be the community structure that is the unit of selection under different environments. That said, any observed association of modularities with the environmental features [29, 30] or growth conditions [233] would naturally give rise to a question as to whether such a correlation arises due to similar community structures (which, by definition, would have similar modularity scores) or different community structures with similar modularity scores.

In this work, we analyzed metabolic networks of species spanning three kingdoms of life by computing their community structures and modularity scores (see Methods for details on metabolic network reconstruction). We compared the difference in community structures against the difference in modularities and the genetic distance, to investigate the correlation, or lack thereof, among the three. The results suggest that the difference in community structures does not parallel the difference in modularity scores we compute, except when community structures are extremely similar. That is, we find that larger community structure differences do not necessarily mean larger differences in modularity scores and vice versa, which is an indication of convergent evolution of modularities via different underlying community structures. To further understand the evolutionary driving force behind such convergent evolution, we revisited the analysis of Parter et al. [29], which first associated modularity with habitat variability, but under different aspects of the microbial life styles, including

temperature preference and oxygen requirement. We also confirmed the finding of Kreimer et al. [30] that the size of the metabolome (the number of enzymes) is a major determinant of the modularity score, even after the score is normalized and believed to be size-independent on general (non-metabolic) networks.

From a computational perspective, a contribution of this paper is an improved heuristic based on spectral decomposition for modularity optimization [240] using a self-organizational *merge and resplit* refinement. The goal of this improvement is to deterministically identify more optimal modularity scores and community structures efficiently. We show, on well-studied benchmark data sets, that compared to the original algorithm of Newman [240] and some other existing algorithms [238, 241, 242, 243], our algorithm achieves higher Q scores at the cost of only a moderate increase in time.

5.1.1 Community detection and modularity

The modularity score of a network is defined as follows [238]: consider a network with its set of nodes V and set of edges E, the Q score is defined as a function of a partition $\mathcal{P}$ of V,

$$Q(\mathcal{P}) = \sum_i (e_{ii} - a_i^2) \tag{5.1}$$

where e_{ii} is the fraction of edges in community i (over all edges in the network) and a_i is the fraction of edges that are incident on a node in community i. The highest Q score attained over all possible partitions, $\arg\max_{\mathcal{P}} Q(\mathcal{P})$, is defined as the network's modularity. Two communities are neighbors if there is an edge connecting any pair of their members, i.e., C_i is a neighbor of C_j if there is some $p \in C_i$ and $q \in C_j$ such that $(p, q) \in E$. Several algorithms have been devised to estimate the modularity together with its corresponding community structure; see [244] for a review. In this work, we improve the algorithm of Newman [240] to optimize the modularity score.

The improvement is achieved by global merge and resplit and is given in Algorithm 7.

Algorithm 7: Merge-Resplit

Input : Graph $g = (V, E)$.
Output: A partition $\mathcal{P}$ to maximize Q.

1 $\mathcal{P}$ = RecursiveBipart(V, E);
2 **do**
3 **for** C_i, C_j = *neighbors in* $\mathcal{P}$ **do**
4 $C_{merge} = C_i \cup C_j$;
5 $\mathcal{P}' =$ RecursiveBipart(C_{merge}, E);
6 **foreach** $v \in C_{merge}$ **do**
7 $S(v) = \begin{cases} 1 & \text{if} \quad v \in C_i \\ -1 & \text{if} \quad v \in C_j \end{cases}$;
8 **end**
9 $\mathcal{P}'' =$ KirnighanLin(C_{merge}, E, S);
10 $\mathcal{P} = \text{argmax}_{\mathcal{P} \in \{\mathcal{P}', \mathcal{P}''\}} Q(\mathcal{P})$;
11 **end**
12 **while** $\mathcal{P}$ *is varying* ;
13 **return** $\mathcal{P}$

Procedure RecursiveBipart on line 1 and 5 follows Newman [240] which recursively bipartitions its input graph using spectral decomposition by [245] and [246], with the KirnighanLin (on line 9) procedure interleaved on each level of bipartitioning. Following Newman [240], given any bipartion (C_i, C_j), if we define Q as a quadratic product of graph Laplacian L and the membership vector S (as defined in line 6).

$$Q = \frac{1}{2} S^T L S \tag{5.2}$$

Optimal Q is achieved by finding S with the leading eigenvalue of L. Eigen problems are solved using shifted power method. Each step in KirnighanLin procedure both on line 9 and inside RecursiveBipart (following Newman [240]) optimizes the boundary of two communities by greedily swapping a pair of nodes whose exchange results in the largest increase in Q. The intermediate state with the highest Q is

returned.

After the initial decomposition from RecursiveBipart, each pair of communities thus obtained are merged and fed again into RecursiveBipart, whose spectral property guarantees that the computed partition, which might contain one, two or more subsets, yields no lower Q. The new partition obtained is compared with a partition obtained by directly applying the KirnighanLin procedure to the boundary between the two original communities. The partition that gives rise to the larger Q is kept. This is to ensure the new partition will lead to better optimization than the current one. Such merge-resplit process continues until the partition no longer varies after completely traversing the boundaries between all pairs of the neighboring communities, thereby reaching a self-organized state (a state in which boundaries between any two neighboring communities can not be further improved). The modified algorithm outperforms the existing deterministic algorithms and some computationally heavy stochastic methods, in maximizing Q, as is shown in Supplementary Material Table 2 (see Supplementary Material Table 3 for the computation time at each benchmark data set). A C implementation of the improved algorithm is available at `http://www.bioinfo.cs.rice.edu/`.

5.1.2 Normalized modularity

Following Parter et al. [29], normalized modularity is defined as

$$\frac{Q - Q_{\text{rand}}}{1 - 1/M - Q_{\text{rand}}}. \tag{5.3}$$

where M is the number of communities in the real network and Q_{rand} is the mean Q value of randomized networks. To determine the number of rewiring operations in computing Q_{rand}, we use the leveling of global clustering coefficient [182] of the net-

work as the signal for convergence. For each edge semantics, the number of rewiring operations required to make level the global clustering coefficient of the largest network is used for all species when we rewire its metabolic network of the particular edge semantics (see Supplementary Material Fig. 21). Each rewiring operation involves swapping the ends of two randomly chosen edges. This process keeps the networks' degree distribution. Alternative null models can involve the constraint of the number of short cycles. We do not consider the constraint due to difficulty in identifying all the cycles and ambiguity in determining the length of the cycles constrained.

5.1.3 Mutual information

Given two partitions $\mathcal{A}$ and $\mathcal{B}$ (in this work, $\mathcal{A}$ and $\mathcal{B}$ are the community structures of networks from two different species), the mutual information $MI(\mathcal{A}, \mathcal{B})$ [247] is defined as,

$$\frac{2 \times (H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{AB}))}{H(\mathcal{A}) + H(\mathcal{B})}, \tag{5.4}$$

where the marginal entropy is defined as,

$$H(\mathcal{A}) = \sum_{i \in \mathcal{A}} \frac{N_i}{N} log(\frac{N_i}{N}), \tag{5.5}$$

N_i is the number of nodes that belong to set $i \in \mathcal{A}$ and N is the total number of nodes common to both networks. The joint entropy is defined as,

$$H(\mathcal{AB}) = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} \frac{N_{ij}}{N} log(\frac{N_{ij}}{N}). \tag{5.6}$$

and N_{ij} is the number of nodes that belong to both set $i \in \mathcal{A}$ and set $j \in \mathcal{B}$.

5.1.4 Data

We obtained manually annotated metabolic networks of 1021 species from the KEGG database [176] (see Supplementary Material Fig. 22 for a summary of enzymatic annotations). The networks were assembled following Kreimer et al. [30]. Reaction direction information was extracted from the pathway KGML file provided by KEGG. Altogether there are 3548 KEGG reactions with direction identified, leaving 4635 reactions denoted as reversible. From these data, we assembled four types of networks using four different semantics, namely, compound networks where nodes are metabolites, enzyme networks where nodes are enzymes, compound networks with currency deletion where nodes are metabolites and connections are pruned as in [34] and [42], and enzyme networks with currency link deletion where nodes are enzymes and connections are pruned as in [30]. Analyses shown in this work are of enzyme networks with currency link deletion unless stated otherwise. The species' habitat variability, temperature preferences and oxygen requirements are obtained from NCBI Genome Project Organisms Info Tab (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi).

To conduct an evolutionary analysis of the data, we make use of the phylogeny, both branching pattern and branch lengths with branch lengths measuring sequence divergence in the unit of the number of subtitutions per site, inferred by [248]. Out of the 1021 species, only 56 appear in this phylogeny. Therefore, when we compare the community structures and modularity scores against genetic distances, only the 56 species shared by the phylogeny are used. The genetic distance between any pair of species is defined as the sum of the lengths of the branches on the path between the two species on the species phylogeny.

5.2 Modular structure and modularity in metabolic networks

Previous studies have shown the association of modularity of metabolic networks with variability of the living environment of species [29] and the bacterial life style [30]. However, it remains unclear whether or not this association is a consequence of any further association with the underlying community structure. In other words, the relation between the living environment and modularity might be a consequence of the habitats' association with the community structure. To answer this question, we investigate whether for a similar modularity score there exist multiple distinct community structures in metabolic networks of different species.

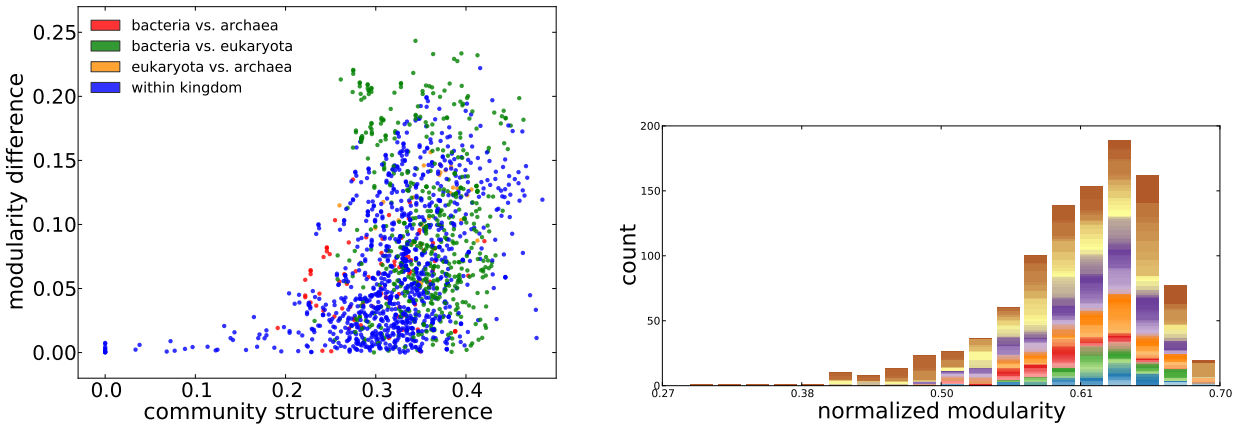

Figure 5.2: **Community structure vs. modularity.** Left) Community structure difference vs. modularity difference: Difference in community structure is computed by $1 - MI$ where MI is the mutual information between the two community structures. Right) Distribution of modularity scores colored by the cluster to which the community structures of the metabolic network belongs (See Methods for the method used to cluster species based on the distance in the community structures). Modularity scores are normalized with respect to scores based on randomized networks (See Methods). The normalized modularity is believed to have network size-dependent factors removed, allowing networks of different sizes and connectivity to be comparable in modularity [249]. Each color corresponds to a community structure cluster. The height of the bar (or bar segment) is proportional to the number of species in each cluster falling into the particular bin of modularity scores.

The results in the left panel of Fig. 5.2 show that a smaller difference in modularity is not an indication of more similar community structures. When the community structures are similar (roughly < 0.2), their modularity scores must be similar. Such dependency is expected from the definition of modularity. Beyond 0.2 in the difference of community structures, modularities vary significantly, from very similar to very different, despite different community structures. In other words, the same modularity score may be achieved via different community structures. Such convergence at the modularity level takes place mostly between bacteria and eukaryota, though also happening between species within the same kingdom, as indicated by the green and blue dots on the bottom right corner of the left panel of Fig. 5.2. To further explore this relationship between modularity scores and community structures on metabolic networks, we plotted the distribution of modularity scores for each community structure cluster (Fig. 5.2) obtained through hierarchical clustering (see Methods). In the right panel of Fig. 5.2, we see that most community structure clusters span many bins of modularities and for each bin of modularity scores, community structures from different clusters can be discerned. This indicates that similar modularity scores found on metabolic networks can stem from different community structures.

5.3 Convergent evolution of modular structures

By comparing community structures of the networks across multiple species, we find that community structures are only specific at the kingdom level but not lower. Clustering of species based on the mutual information of community structures separates species from different kingdoms with some exceptions, as is shown in Fig. 5.3. The discrimination of kingdoms from the community structure of metabolic networks is brought about by the similarity of *enzyme profiles*, or the spectra of all enzymatic activities as are characterized by the sets of Enzyme Commission (EC) numbers,

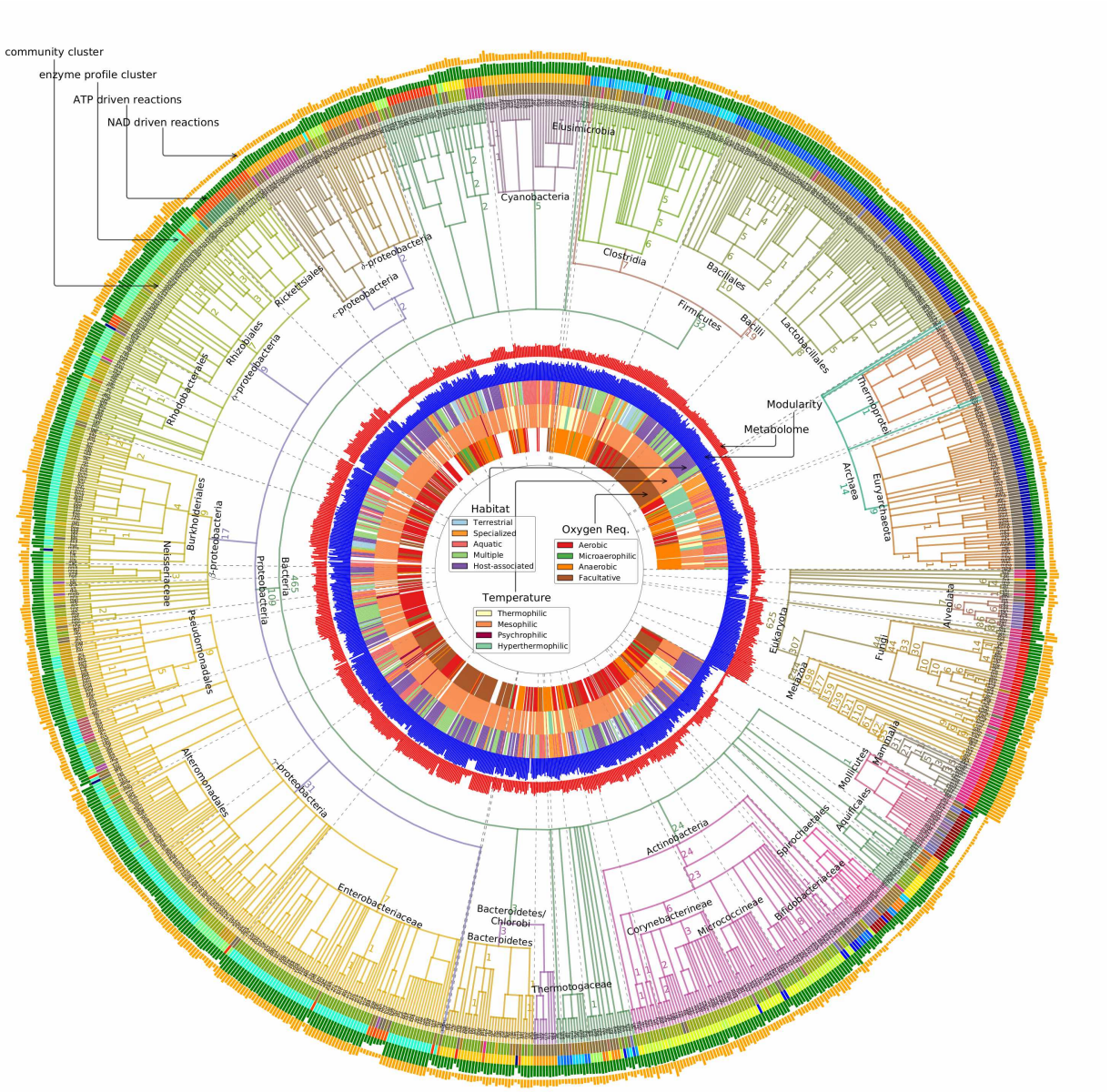

Figure 5.3: **Clustering of community structures.** The outermost track is colored according to the clustering of community structure. The phylogeny and names of the clades are obtained from the NCBI taxonomy. The blue track corresponds to the normalized modularity score (very similar pattern has been observed in unnormalized modularity scores, omitted due to page limit). Enzyme clusters are obtained by flattening the UPGMA linkage such that the cophenetic distance among leaves in each cluster is less than 0.5. The red track indicates the number of annotated enzymes in each species.

among species from the same kingdom. As is shown in Fig. 5.3 where we label on each branch the number of enzymes appearing exclusively in the descendants of the branch (an indication of metabolic innovation specific to the lineage), both bacteria and eukaryotes have their characteristic metabolic capabilities (465 and 625 respectively) while archea tend to share their metabolic capabilities with species from other kingdoms (14 unique enzymes).

Due to the independence of enzyme-reaction relationship from the choice of the species, enzyme profiles directly determine the connectivity, and hence the community structure of the metabolic networks. Any difference in the community structure is a result of some difference in the enzyme profile. To see whether different enzyme profiles would generate similar community structures, we cluster the species by their enzyme profiles using Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [250]. We find that the clusters based on enzyme profiles agree to a substantial degree to the clusters based on community structures (third and fourth tracks from the outer rim in Fig. 5.3).

5.3.1 Clustering of Community Structures

We cluster the community structures by using hierarchical clustering (nearest point algorithm) implemented in the open source SciPy [251] package. The distance between any two networks is $1 - MI$ where MI is the mutual information between their community structures. Clusters are flattened by looking for largest sets of individuals such that the pairwise distance among its members are within a chosen threshold based on inspection. The threshold used is 0.7.

5.4 Modularity evolution and microbial living environments

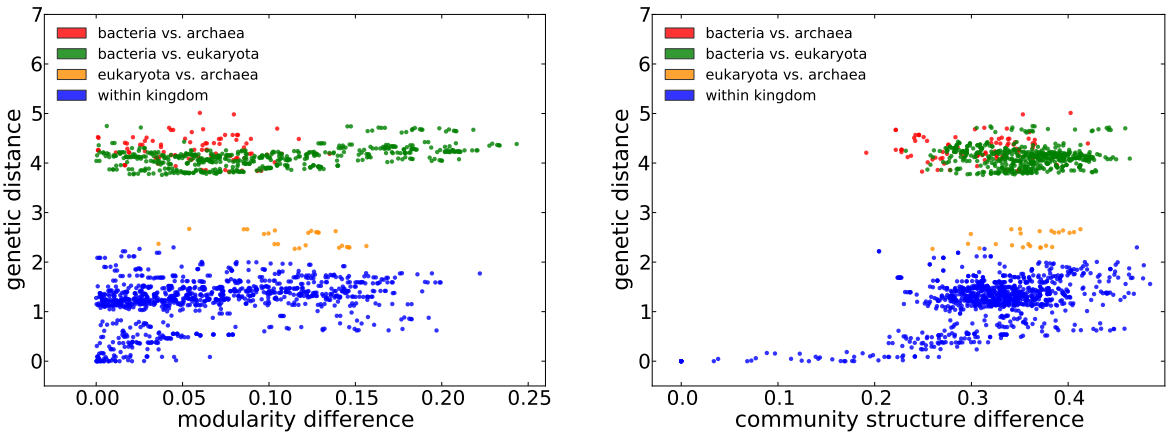

Figure 5.4: **Difference in modularities (left) and community structures (right) vs. genetic distance (in substitutions per site).** The gap in the middle of the plots corresponds roughly to the long branches separating bacteria from the rest (archaea are closer to eukaryota than to bacteria and bacteria are roughly as close to archaea as to eukaryota).

To investigate the evolution of modularity scores and community structures, we plotted for every pair of species the difference in their modularity scores and community structures against their genetic distances (see Methods for the computation of the genetic distances); results are in Fig. 5.4. In the left panel of Fig. 5.4, modularity difference can be close to zero even between species across the kingdoms, which supports the hypothesis of convergent evolution of modularity. On the contrary, community structures are similar only when two species are genetically very close (see the right panel of Fig. 5.4). Since closely-related organisms have similar enzyme profiles (see Supplementary Material Fig. 1) which result in similar metabolic networks' connectivity, and enzyme profile similarities are negatively correlated with community structure differences (Supplementary Material Fig. 2), it makes sense that closely-related organisms also have similar community structures.

Knowing that similar modularity may be achieved independently via different community structures, we revisit the question of what drives the convergent evolution of modularity. We studied several factors ranging from the size of the metabolome (the number of enzymes and the size of the network under the current choice of network semantics) to environmental factors that include temperature preferences and oxygen requirements.

5.4.1 Modularity vs. size of the metabolome

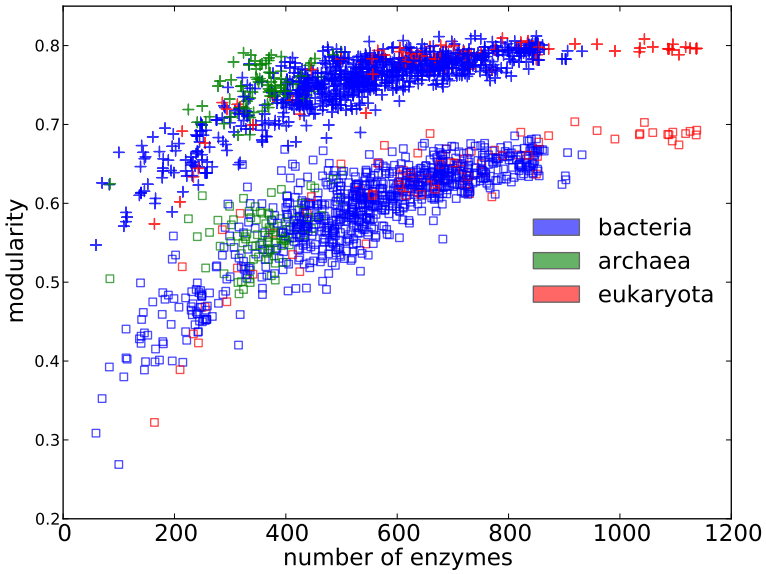

Figure 5.5: **Modularity vs. the number of enzymes.** The numbers of enzymes are significantly correlated with modularity scores ("+" markers). Such observation remains after modularity scores are normalized (square markers).

Network size is reported to be an important determinant of network modularity [30]. We show that: although the normalized modularity is believed to be independent of the network size [249], dependence remains for normalized modularities in the case of enzyme networks (see Methods). In Fig. 5.5, we plot the modularity scores and the number of enzymes. We observe that modularity is significantly correlated with the number of enzymes, whether modularity is normalized or not (Spearman's

ranked $r = 0.85, p = 2.0 \times 10^{-282}$ in the normalized case and $r = 0.80, p = 2.6 \times 10^{-229}$ in the unnormalized case). We also see that species with a reduced metabolome (such as those under the clade of Mollicutes and Rickettsiales) possess smaller modularities in their metabolic networks (see Discussion), which is consistent with our observation here. The dependence of modularity on the number of enzymes is sensitive to rewiring (see Supplementary Material Fig. 3). It is worth mentioning that similar correlation is seen on: 1) synthetic linear graphs (graphs composed of nodes linearly concatenating each other); see Supplementary Material Fig. 4; and 2) the line graph transformations [173] of rewired compound networks with currency metabolites deleted; see Supplementary Material Fig. 5, implying that their resemblance to the organization of metabolic networks may explain the dependence of Newman's modularity on the sizes of the network.

5.4.2 Modularity vs. environmental variability

When revisiting the association of modularity to environmental variability, we find a similar trend as is reported by Parter et al. [29] (left column of Fig. 5.6, with the data set used in [29] plotted in the top row and a larger data set plotted in the bottom row). However, an identical trend is also seen for the number of enzymes (right column of Fig. 5.6). This means that the association of modularity with the environmental variability might be a consequence of the difference in the numbers of enzymes between species living in environments of different variability, given the aforementioned strong correlation between modularity and the number of enzymes. In the study by Parter et al. [29], the category "host associated" in the classification from NCBI was further refined into "obligate" and "facultative" to differentiate bacteria that are able to survive without the host from those that cannot. We find that under this refinement, obligate species have a significantly smaller number of enzymes than

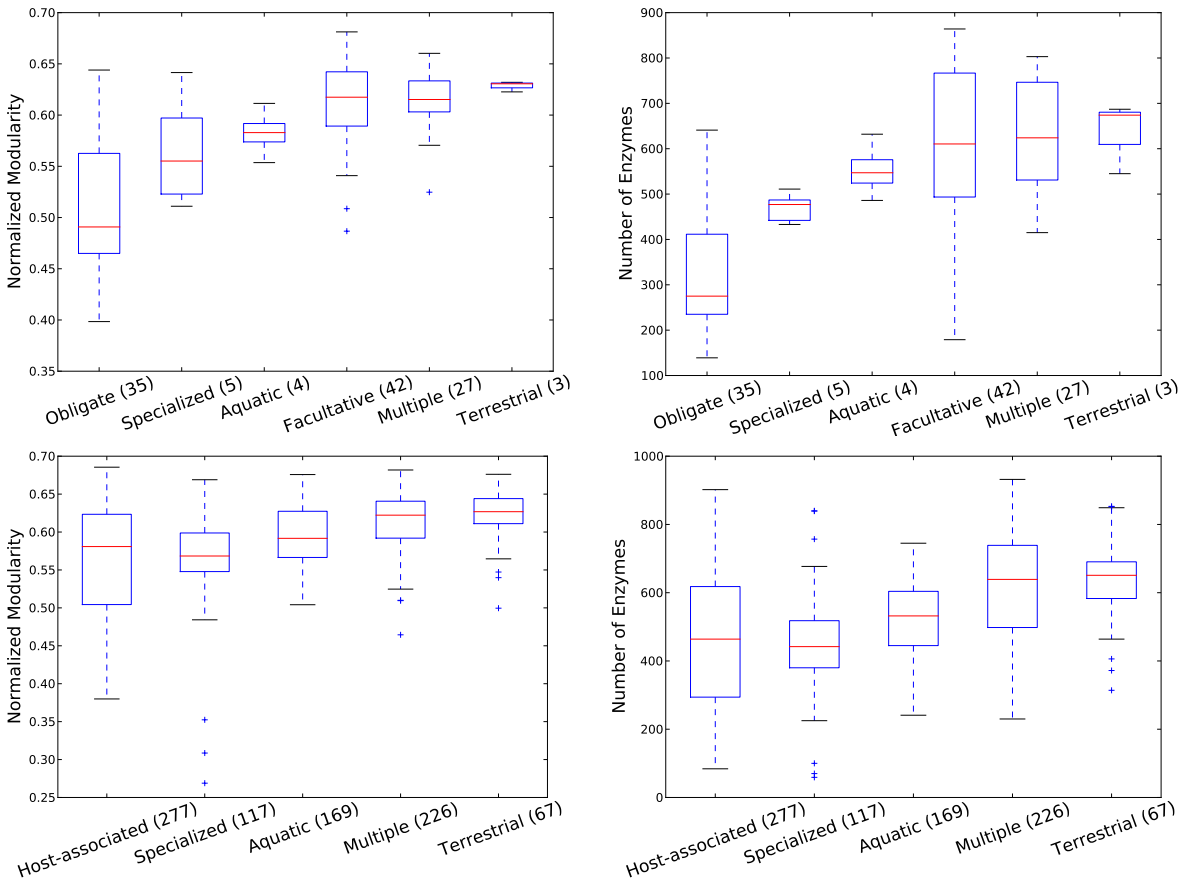

Figure 5.6: **Environment variability and modularity.** (Top row) On a small data set of 116 bacteria, habitat variability vs. normalized modularity (left) (Kruskal-Wallis H-test $p = 5.48 \times 10^{-11}$) and habitat variability vs. the number of enzymes (right) (Kruskal-Wallis H-test $p = 1.03 \times 10^{-10}$). (Bottom row) On a large data set of 806 microbes, habitat variability vs. normalized modularity (left) (Kruskal-Wallis H-test $p = 6.51 \times 10^{-26}$) habitat variability vs. the number of enzymes (Kruskal-Wallis H-test $p = 1.03 \times 10^{-30}$).

facultative ones (one tailed Wilcoxon rank-sum test $p = 4.6 \times 10^{-10}$). Moreover, this refinement is not perfect (for example, the smallest facultative species *B. burgdorferi* is often described as obligate [252, 253] and the second largest obligate species *R. Baltica* in the data set is in fact free-living marine bacteria [254]). Therefore, the difference in the number of enzymes between facultative species and obligate species could in fact be more striking.

It is conceivable that microbes capable of coping with a varying and open habitats

have a larger metabolome and microbes that lead specialized lifestyles have a smaller metabolome. An extreme case is that bacteria leading an obligate lifestyle has a reduced metabolome. One explanation of this phenomenon is that unnecessary genes for living in a specialized niche that only increase the overhead of maintenance were lost during evolutionary history [123, 157, 158, 124]. For example, the γ-proteobacteria *B. aphidicola* lack the genes for the synthesis of tryptophan, riboflavin, fatty acids and phospholipids due to its endosymbiosis with aphids [255, 256]. Here we see that the numbers of enzymes of 8 insect endosymbionts in γ-proteobacteria are significantly smaller than the other species in our dataset (one-tailed Wilcoxon rank-sum test $p = 1.1 \times 10^{-6}$). Even the largest of these endosymbionts (*B. pennsylvanicus*, 366 enzymes) has a smaller metabolome than the smallest non-endosymbiont (*D. nodosus*, 459 enzymes). Modularity scores of endosymbionts are also significantly smaller than non-endosymbionts (one-tailed Wilcoxon rank-sum test, $p = 2.5 \times 10^{-6}$).

To study whether habitat variability truly affects the modularity of the metabolic networks besides the effect of the number of enzymes, we binned the species into groups with the number of enzymes in bins ranging within at most 50 enzymes. Out of 24 bins from 100 to 820 with the number of enzymes incrementing by 30, 16 bins contain at least two categories of species each of which has more than 10 members. Only in 4 of these 16 bins (310$\sim$340, 430$\sim$460, 490$\sim$520, 520$\sim$550) habitat variability significantly (Kruskal-Wallis H-test $p < 0.05$) affects the network modularity. This fact shows that most of the seeming dependence of modularity on the habitat variability may disappear if the number of enzymes is controlled.

5.4.3 Modularity vs. temperature preference

Temperature preferences and oxygen requirements can be more objective measures of environmental variabilities. By comparing the modularities against the temperature

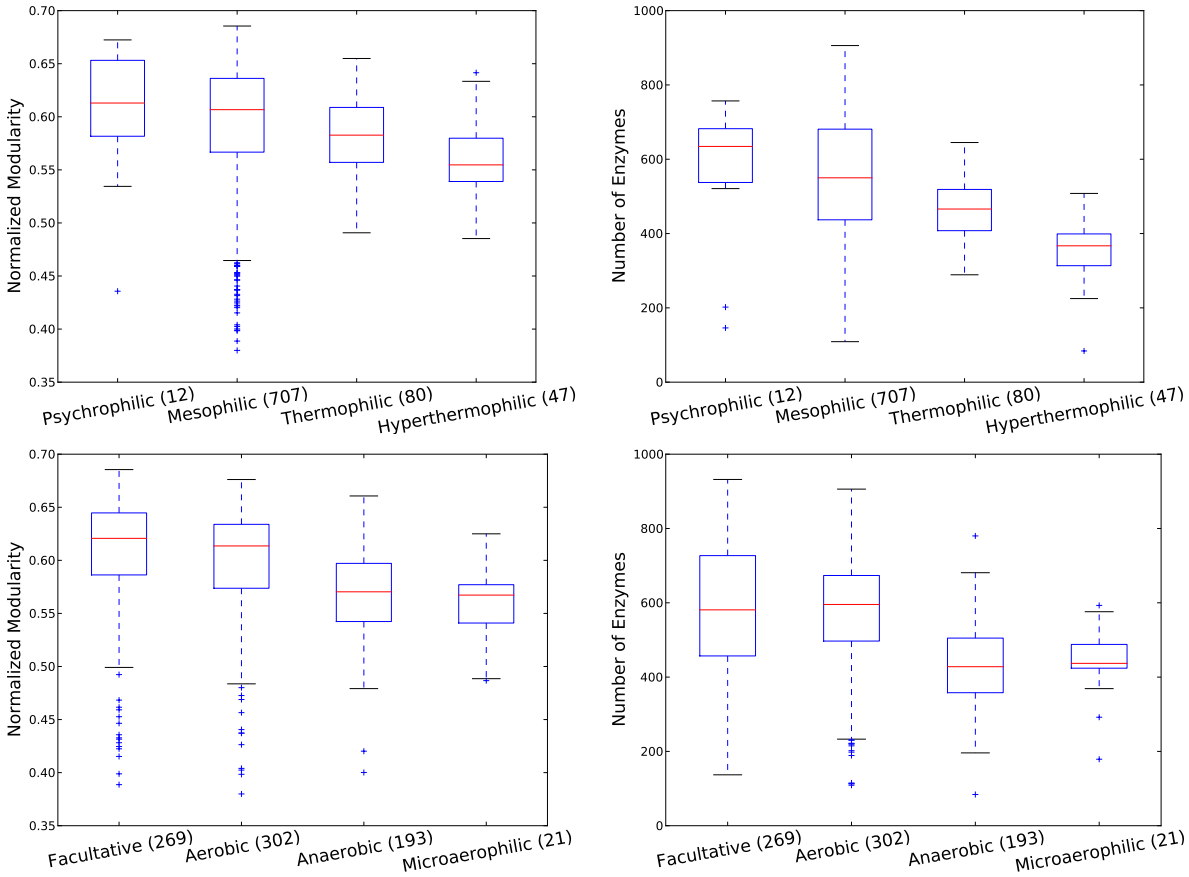

Figure 5.7: **Modularity and environment factors.** (Upper left) normalized modularity vs. oxygen requirement (Kruskal-Wallis H-test $p = 2.96 \times 10^{-25}$); (Upper right) number of enzymes vs. oxygen requirement (Kruskal-Wallis H-test $p = 3.35 \times 10^{-33}$); (Lower left) normalized modularity vs. temperature requirement (Kruskal-Wallis H-test $p = 5.52 \times 10^{-9}$); (Lower right) number of enzymes vs. temperature requirement (Kruskal-Wallis H-test $p = 1.06 \times 10^{-19}$).

(top row of Fig. 5.7), we find that thermophilic and hyperthermophilic bacteria have a lower modularity (see Supplementary Material Table 1 for pairwise comparison). In all the cases where we compare modularity, we also compare the number of enzymes from different categories. We observed a significant difference in every case. And the number of enzymes has a consistent trend as modularity, which again indicates that the association of modularity to the temperature is mediated by the number of enzymes. The variation in the number of enzymes can be understood recognizing the biochemical fact that only a small amount of enzymes can function properly under

elevated temperature.

5.4.4 Modularity vs. oxygen requirements

By comparing the modularities against the oxygen requirements of the species (bottom row of Fig. 5.7), we find that facultative bacteria have the highest modularity. Microaerophilic bacteria have the least modularity. Facultative bacteria are ones that normally utilize oxygen as their electron receptor but can also ferment other endogenous electron receptors such as ethanol and lactate. On the contrary, microaerophiles have the most strict requirement for oxygen. For them, oxygen is not only a requirement for survival, but the concentration of oxygen must also be lower than what is present in the atmosphere. If environmental variability should explain the difference in modularity, the flexibility in oxygen usage, as one way of reflecting environmental variability, supports such explanation: facultative bacteria have higher modularity than strictly aerobic and strictly anaerobic bacteria. And strictly aerobic bacteria have higher modularity than microaerophiles. There is no significant difference in modularity between anaerobic bacteria and microaerophiles (two tailed Wilcoxon rank-sum test $p = 0.40$, same result for the number of enzymes, $p = 0.57$). However, bacteria that are capable of freely metabolizing oxygen (facultative joined with aerobic) have significantly (one tailed Wilcoxon rank-sum test $p = 5.5 \times 10^{-26}$) higher modularities than those who have limited capability of handling oxygen or have to rely on fermentation (microaerophiles joined with anaerobic). The same result is obtained when the number of enzymes are compared ($p = 1.1 \times 10^{-35}$). Comparison between only strictly aerobic microbes against strictly anaerobic microbes also indicates statistical significance (one tailed Wilcoxon rank-sum test $p = 6.9 \times 10^{-16}$ in modularities and $p = 1.2 \times 10^{-30}$ in the numbers of enzymes). Facultative bacteria have significantly higher modularities than strictly aerobic bacteria (one tailed

Wilcoxon rank-sum test $p = 0.0025$). However, a null hypothesis is accepted when it comes to the number of enzymes (one tailed Wilcoxon rank-sum test $p = 0.18$), meaning that the significant difference in modularity between facultative bacteria and strictly aerobic bacteria is not a consequence of the difference in the numbers of the enzymes.

5.5 Biological interpretation of modularity-based communities

Despite the existing studies on the modularity of metabolic networks and reported limitation in modularity-based community detection such as the *resolution limit* [257] (optimizing the modularity score might fail to detect small communities), the *non-locality* [258] (the local delineation of a community depends on the global network connectivity) and the *extreme degeneracy* [259] (there might exist multiple optimal/suboptimal community structures), it remains unclear whether, in this specific case of metabolic networks, modularity-based communties reflect the graph-theoretic intuition of a community structure.

To briefly investigate whether the modularity score (and the corresponding community structures) reflects the intuitive concept of being "modular" (that is, whether a graph with high modularity score can indeed be partitioned into dense subgraphs with sparse connectivity across subgraphs) given the specific topologies of metabolic networks, we compare the communities based on Newman's definition against one of the many other definitions, namely the one by Radicchi et al. [260], where the community structure definition in strong sense requires that for all the nodes in the network, the number of neighbors of the node from the same community (k^{in}) be greater than the number of neighbors of the node from different communities (k^{out}).

The definition in a weaker sense only require the sum of k^{in} be greater than the sum of the k^{out} over all nodes in a community. We computed the k^{in}, k^{out} for all the nodes in the metabolic network of *E.coli*. We find that the partitions obtained via modularity optimization satisfy the weaker definition (see Fig. 5.8). Most communities also satisfy the strong definition (Supplementary Material Fig. 6). In all the 10 nodes in *E.coli* that break the definition in the strong sense, the connections to nodes from the same community outnumber the connections to any one of the other communities to which the node does not belong (even though the sum of outward connections is greater). This explains why these nodes are not classified into any of the other communties. These 10 nodes consist of 2 oxidoreductase, 6 transferase and 2 lyases. No particular preferences of pathway participation from these exceptions was observed.

In order to test the extent of the resolution limit of the modularity based community detection on metabolic networks, we computed densely connected subgraphs using the SIDES program [261]. As shown in the right panel of Fig. 5.8, most of the densely connected subgraphs are contained in the same communities, which is a rough indication of the exemption from the resolution limit.

Despite these findings, the definition of modularity we use might still be problematic when applied to linear/sparse graphs. As we show in Supplementary Material Fig. 4, the longer the linear graph, the higher its modularity, which is problematic given that two line graphs should be intuitively considered equally modular regardless of their lengths.

Another crucial question in studying the modularity of metabolic networks is whether the communities detected carry any functional meaning (in the biological sense). Intuitively, modularity or density-based methods would not identify linear, or more generally sparse, pathways. To answer this question, we investigate the

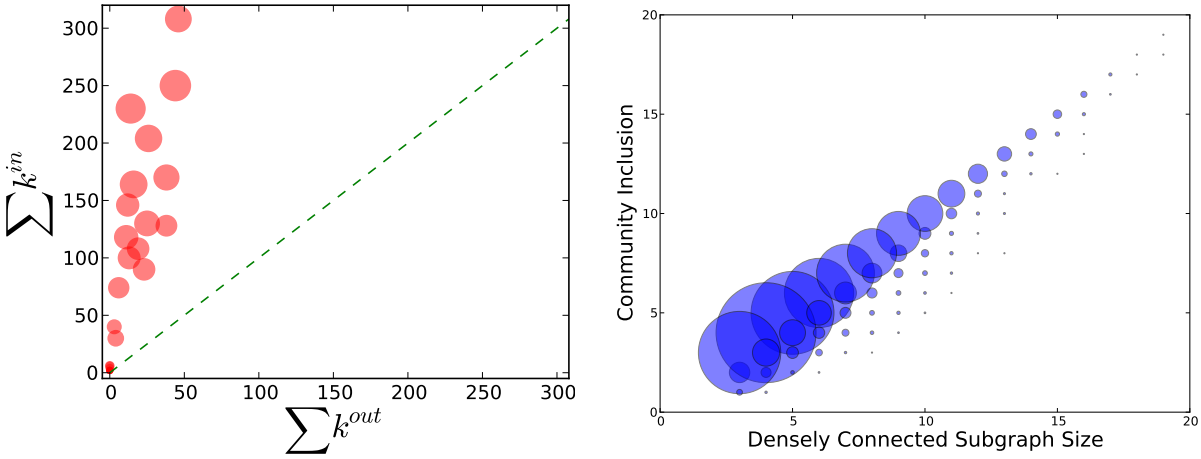

Figure 5.8: **Topological meaning of communities detected.** (Left) Following [260], k^{in} is the intra-community degree and k^{out} is the inter-community degree of each node in the metabolic network of *E.coli*. Summation ($\sum k^{in}$ and $\sum k^{out}$) of intra and inter community degree is over all nodes in each community. The size of the circle is proportional to the size of the community. As is shown, all the communities have $\sum k^{in} > \sum k^{out}$ and most nodes are have $k^{in} > k^{out}$ (Supplementary Material Fig. 6). (Right) Inclusion of densely connected subgraphs in the community. The center of each circle corresponds to an observation of a densely connected subgraph K (in any of the 1021 species investigated). X-axis indicates the size of K and Y-axis indicates the maximum overlap of K with a community among all the communities detected from the metabolic network of the same species. The size of the circle is proportional to the number of instances that give the observation.

functional meaning of communities computed on the metabolic network of *E.coli*. We find that these communities have limited specificity to partitions based on biological functions.

First, we explore how communities overlap with established biochemical pathways. Second, we explore the functional similarity based on the Gene Ontology (GO) [262]. For correlation with biochemical pathways, we computed for each pair of community-pathway the community-wise and pathway-wise specificities, defined as the number of reactions shared by both the community and pathway and normalized by the size of the community and size of the pathway respectively. Based on these definitions, if a community is completely contained within a pathway, its community-wise specificity (with respect to that pathway) is 1, and if a pathway is completely contained within

a community, its pathway-wise specificity (with respect to that community) is 1. We computed these two specificity measures by using the biochemical pathways of *E. coli* obtained from the KEGG database [176] (left panel of Fig. 5.9). Three patterns are worth observing in this figure. The top right corner has no points, an indication that there is no 1-1 correspondence between pathways and communities. This conforms to our intuition that biochemical pathways are very sparse graphs, whereas communities correspond, roughly, to dense subgraphs. Second, the bottom left corner is very dense, further supporting the lack of a 1-1 correspondence; however, it is important to notice that the points in this corner are all small, reflecting very small overlap between pathways and communities. Third, the pathways and communities with high specificities have relatively large overlaps. These three trends combined indicate that a few pathways are between 50%-80% contained within communities, very few communities are contained within pathways and the majority of pathways are fragmented across communities.

We studied the Gene Ontology (GO) annotation of the genes that transcribe the enzymes in the *E.coli* network using GS^2 [263], a measure that quantifies the similarity of GO terms among a group of genes. In order to tell whether enzymes inside the same community have a similar ontology, we ran GS^2 on genes that are annotated to transcribe enzymes belonging to the same community. We find that genes inside the same community have a higher similarity of GO annotations than the same number of genes but randomly selected from the gene pool of the organism (right panel of Fig. 5.9). Following Bauer et al. [264], we test whether a community is functionally significant by whether there is a significant enrichment of any GO term. The GO specificity is calculated by dividing the extent of overlap between the GO term and the community by the total number of genes that have that GO term in *E.coli*. The community specificity is calculated by dividing the extent of overlap between the GO

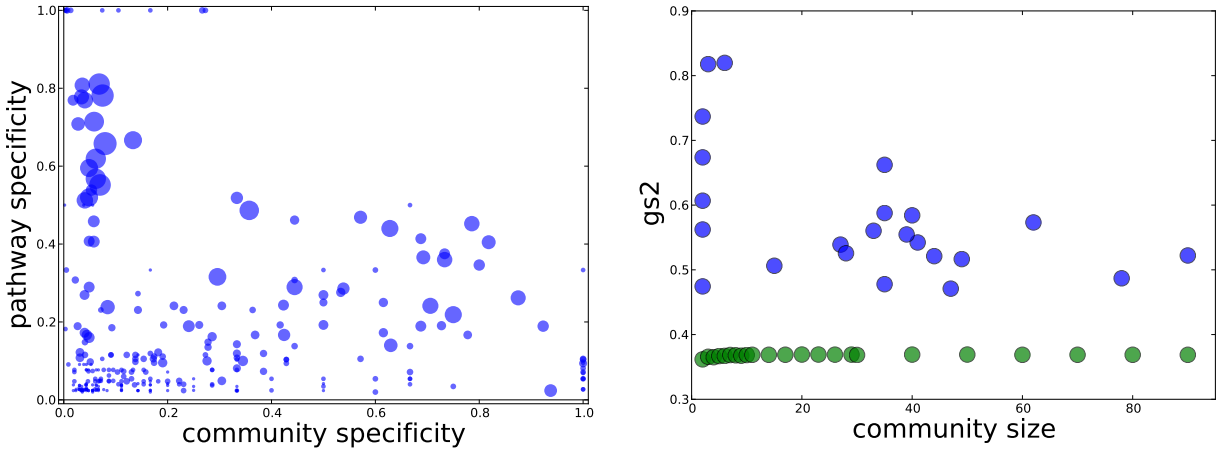

Figure 5.9: **Biological meaning of communities detected.** (Left) The community-pathway overlap in *E. coli*'s metabolic network. Each point corresponds to a community-pathway pair, where the size of a point reflects the size of the overlap between the community and pathway. Similar results are observed on compound networks with currency metabolites deleted (Supplementary material Fig. 7). (Right) Gene Ontology Enrichment of Communities. The similarity of genes inside each community detected from *E.coli* (blue) against the similarity of genes randomly selected (green).

term with the community by the number of genes that transcribe the enzymes in the community. GO-community pairs where the GO term significantly annotates the community are isolated (tested against the hypergeometric distribution with Bonferroni correction for multiple comparisons, $\alpha = 4.7 \times 10^{-5}$ [265]). In spite of many GO-community annotations with significant p-values, no clear 1-1 correspondence between GO terms and community structures is seen (Supplementary Material Fig. 8). This suggests that the GO similarity among genes inside the same community might result from their closer distance on the network, assuming genes inside a community are closer on the network and nodes closer on the network are more likely to share GO annotations.

5.6 Weighted community detection for mining functional information

One way of utilizing a metabolic network's connectivity map is to mine it for functional information. As we discussed above, community structure of a metabolic network is believed to reflect functional categorization [22]. By manually curating a hierarchical compound classification for *U. urealyticum* (see Supporting Information Figure 8), and comparing it to communities detected on the different networks of this organism, we studied the correspondence between community structure and functional categorization. Community structure is detected using our algorithms. Mutual information and uncertainty coefficients (see Materials and Methods) are used to evaluate the match level between the compound classification and the detected communities.

Among all the unweighted networks considered, pathway combination has the best match with the compound classification (top left panel in Fig. 5.10). Manual curation has a better match with functional categories than the hub removal approach, possible due to the fact that the latter approach results in the removal of all edges incident with a hub (see below). Although removing more hubs improves the match, it never reaches the value obtained through other methods.

Similar trends are observed in terms of the uncertainty coefficients (see Supporting Information Figure 9). These results raise the question of why it is hard to get a perfect match between functional categorization and community structure. We believe this has to do with the incongruent nature of the community defintion and biological functions.

First, a module does not have to be necessarily very densely connected to correspond to certain function. Actually, most well-known metabolic pathways are very

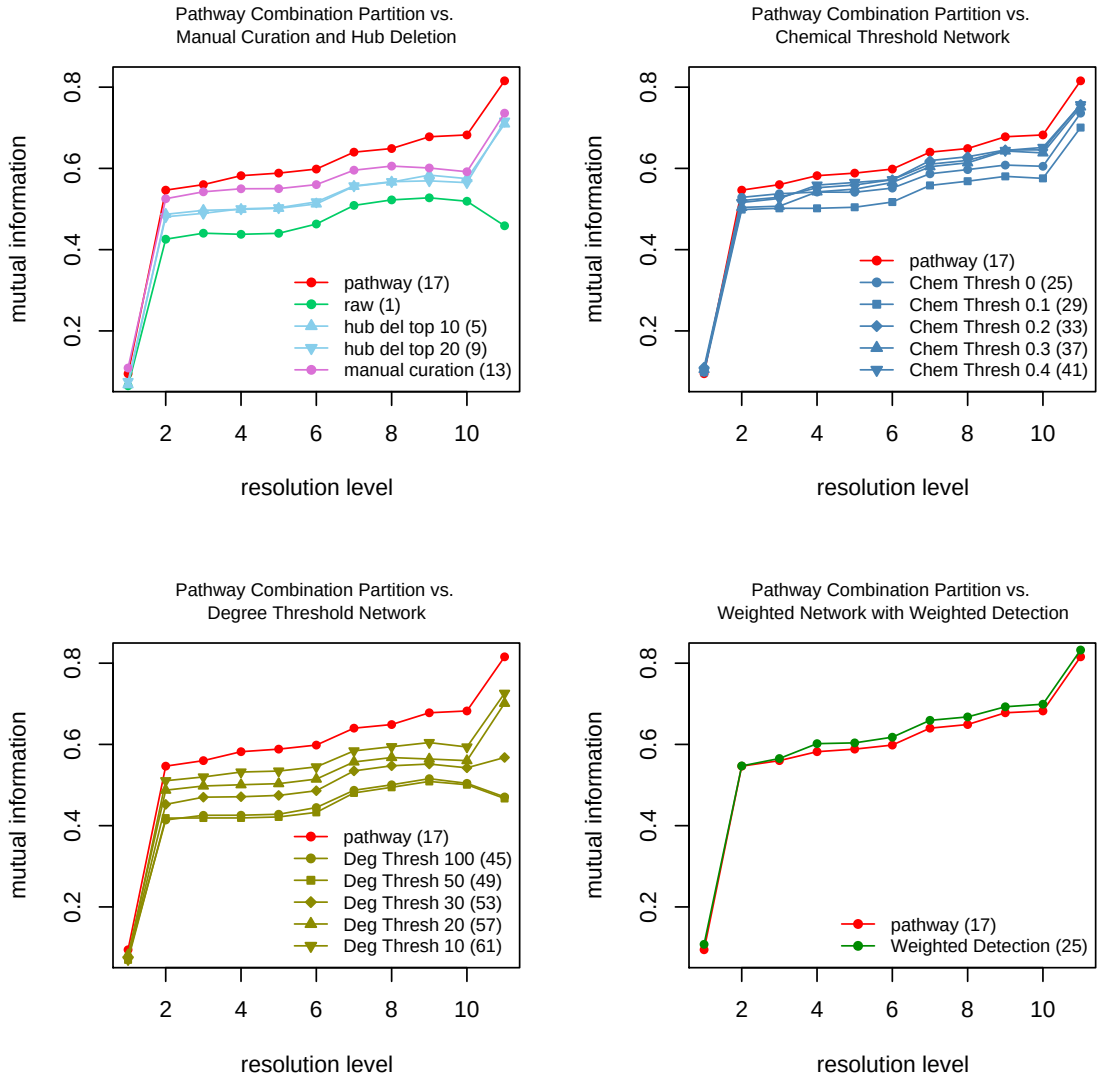

Figure 5.10: Mutual Information between a hierarchical compound classification of *U. urealyticum* and the community structure detected on various representations of the organism's metabolic network. Higher mutual information score suggests a better matching between the two partitions. Mutual information 1 means a perfect match. The x-axis corresponds to the 11 different categories in the classification, and the numbers in parentheses are indices of network representations (see Supporting Information Table 4).

sparsely connected, including Kreb's cycle and the glycolysis pathway. This is consistent with the observation in Figure 5.10 where pathway combination networks have the best match (among all unweighted networks) with the functional categorization,

even though it is very linear as is implied by a low clustering coefficient (see Supporting Information Figure 7,9). There are many explanations for a biological module not to be dense. Theoretically, the information entropy in a completely connected network is as low as that of a network with no connections. If structual complexity is a requirement for the metabolic function, a densely connected network may not always be the optimal choice, let alone that structural complexity may not be favored by selection. Biologically, metabolic networks are context-sensitive networks. Each edge has its unique specific significance to the organism, structually similar integrations of these edges may assume completely different logic. Such functional diversity is missing in analyzing the pure connectivity. The richness in connection does not mean richness in real functionality. If a biologically functional and meaningful module is sparse, it confuses the comunity structure on a simple graph representation since central to the idea of community detection through the optimization of modularity is the notion that the intra-community connectivity is statistically denser than the inter-community connectivity.

Moreover, the resolution of a functional categorization may vary and the one that the community structure is really coding for is hard to pinpoint. Taken to an extreme, every metabolite can legitimately be considered a module by itself say in a even finer molecular view. In that case, a community structure that matches this categorization perfectly can be trivially obtained by partitioning the network into all single-node modules, which is what most community detection algorithms try to avoid. In order to increase our chance of capturing matches at various levels of functionalities, we prepared a hierarchical compound classification. Other technical reasons including the drift effect of community detection algorithms [258] and an intermingle of different edge semantics (see Supporting Information) may also contribute to such imperfection in matching.

A question, then, is raised as to whether it is possible to devise a network representation that yields a better community-functionality correspondence. To answer this question, we propose a weighted representation of metabolic networks, from which a class of unweighted networks can be derived by simply specifying a threshold value. Then, communities can be detected on the unweighted networks, or by a newly developed algorithm that operates directly on the weighted networks. The threshold value has a biochemical implication in measuring the strength of chemical causality. Note that weighted representation without targeting the chemical transformation is not novel in the field. For example, Croes et al. weighted metabolites by their degrees in the graph [42]. Our approach is different in that we weight each reactant pair with chemical causality by the fraction of common chemical moiety shared in the reaction out of the total amount of chemical content of the larger compound. Here, chemical content is measured by the number of atoms excluding hydrogen. In this case, unweighted networks can be obtained by removing edges whose weights are below a certain threshold. We observe that when low-weight edges are deleted, networks gradually reduce to consistent linear topologies (in the sense of low clustering coefficient, high characteristic path length and a similar degree distribution shape) that are conceptually expected in the pathway combination network (see Figure 5.11).

As shown in Fig. 5.12, naturalizing unweighted networks results in highly fragmented networks with a major component with poor coverage of the original networks. However, we observe that raising the threshold in chemically weighted networks results in very slow degradation of the size of the major component (see Supporting Information Figure 15), thus leaving it to provide a good coverage of the original network. Further, we observe that the subgraphs that are "shed" from the major component as the threshold is raised correspond to different functions, indicating that pathway assignments appear at different levels of thresholding. For example, from

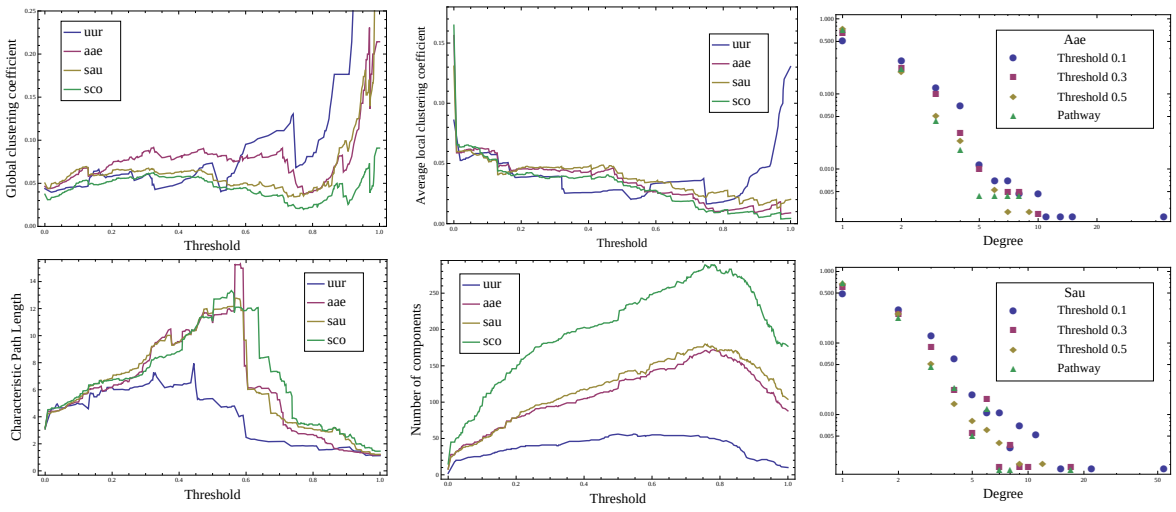

Figure 5.11: Different unweighted networks are obtained by thresholding the compound network weighted by chemical causality. As the threshold increases, the networks' global clustering coefficient (**upper left**), averaged local clustering coefficient (**upper middle**), characteristic path length (**lower left**) and the number of connected components (**lower middle**) are computed. The degree distribution of networks with threshold set on 0.1, 0.3 and 0.5, along with the pathway combination network are plotted for *A. aeolicus* (**upper right**) and *S. aureus* (**lower right**). Numbers shown in the legends are the threshold value.

comparing these subgraphs that are shed at different threshold values of the weighted network of *S. coelicolor* and the pathways in KEGG (see Supporting Information Table 1), we observe that xenobiotic pathways, secondary metabolite pathways and cofactor metabolism pathways appear at relatively low threshold values (< 0.3); nucleotide metabolism pathways appear at relatively high threshold values (> 0.7). In between are amino acids metabolisms. However, fatty acid metabolism pathways appear at low (e.g., 0.228) and high (0.917) threshold values. This suggests a gradual cumulative elongation of lipid instead of combination of chemical moiety of similar sizes.

Alternatively, we can weight the reaction network by the degree of the linking compound underlying each edge. In this case, unweighted networks can be obtained by removing hub nodes using one of the three schemes described above. We observe

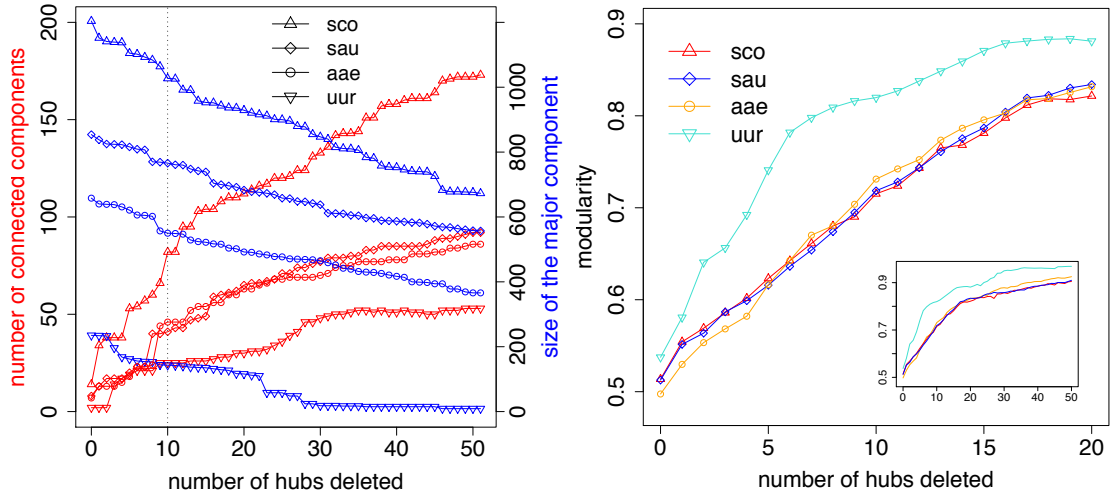

Figure 5.12: **Effects of hub deletion on connectivity patterns.** (Left) The number of connected components and size of major component as hubs are removed. (Right) The network's modularity as hubs are removed. See methods for the calculation of modularity.

that the relative order of the threshold values under different schemes is not conserved across the four organisms (see Supporting Information Figure 16). Particularly for networks that differ greatly in size, different schemes are uniform in either the cutoff degree of hubs (characterizing the strength of flux-coupling) or the hub identities. In this context, we consider three schemes for hub deletions: (1) setting a threshold and removing all nodes whose degrees are higher than the threshold (e.g., [30]), (2) fixing a number of compounds to be removed (e.g., [36]), and (3) fixing a proportion of the compounds to remove. The larger the network, the more likely a compound to be connected to other compounds and hence to have a higher degree. Therefore, when metabolic networks are compared across species, scheme 1 (removing nodes whose degree is higher than a threshold) is uniform in terms of the coupling strength but not in terms of the hub identity (see Supporting Information Figure 17). On the other hand, since most organisms use a roughly similar set of hub metabolites (e.g., ATP,

H_2O, etc.), scheme 2 (removing a specific number of compounds) is more uniform in terms of hub identity but not coupling strength in that the cutoff degree value varies significantly across species. In this sense, scheme 2 is analogous to the deletion of current metabolites where "currency" is associated with the hub status.

Given this tradeoff, a question arises as to what to compromise, the coupling strength or the hub identity. We recommend a consistent combination between edge semantics and naturalization. If coupling causality is concerned, then the naturalization that produces uniform coupling strength is preferred. If chemical causality is considered, the naturalization that produces uniform hub identity is preferred.

Fig. 5.10 shows the match between the hierarchical compound classification and the communities detected on unweighted networks obtained from thresholding the weighted network based on edges (top right panel) and nodes (bottom left panel). The match is clearly better than that of the raw network, the naturalized networks, and very close to the manually curated networks. Yet, the community structure detected on the pathway combination networks still was best in terms of the match. To obtain a better community-functionality correspondence, we have developed an algorithm for detecting communities, directly from weighted networks, by seeking partitions that concentrate the weight in each community. We achieved this by first devising a community detection algorithm on unweighted networks that improves upon existing algorithms by attempting to overcome the "resolution limit" [257], and then adapting the algorithm to weighted networks (See Methods and Materials).

In terms of the weighted algorithm performance, the communities it detects match the functional classification even better than those of the pathway combination networks (bottom right panel in Figure 5.10). Figure 5.13 provides a visual description of the communities identified by our algorithm on the weighted network of *U. urealyticum*; the communities match very well with biological functional assignments

(see Supporting Information Table 2).

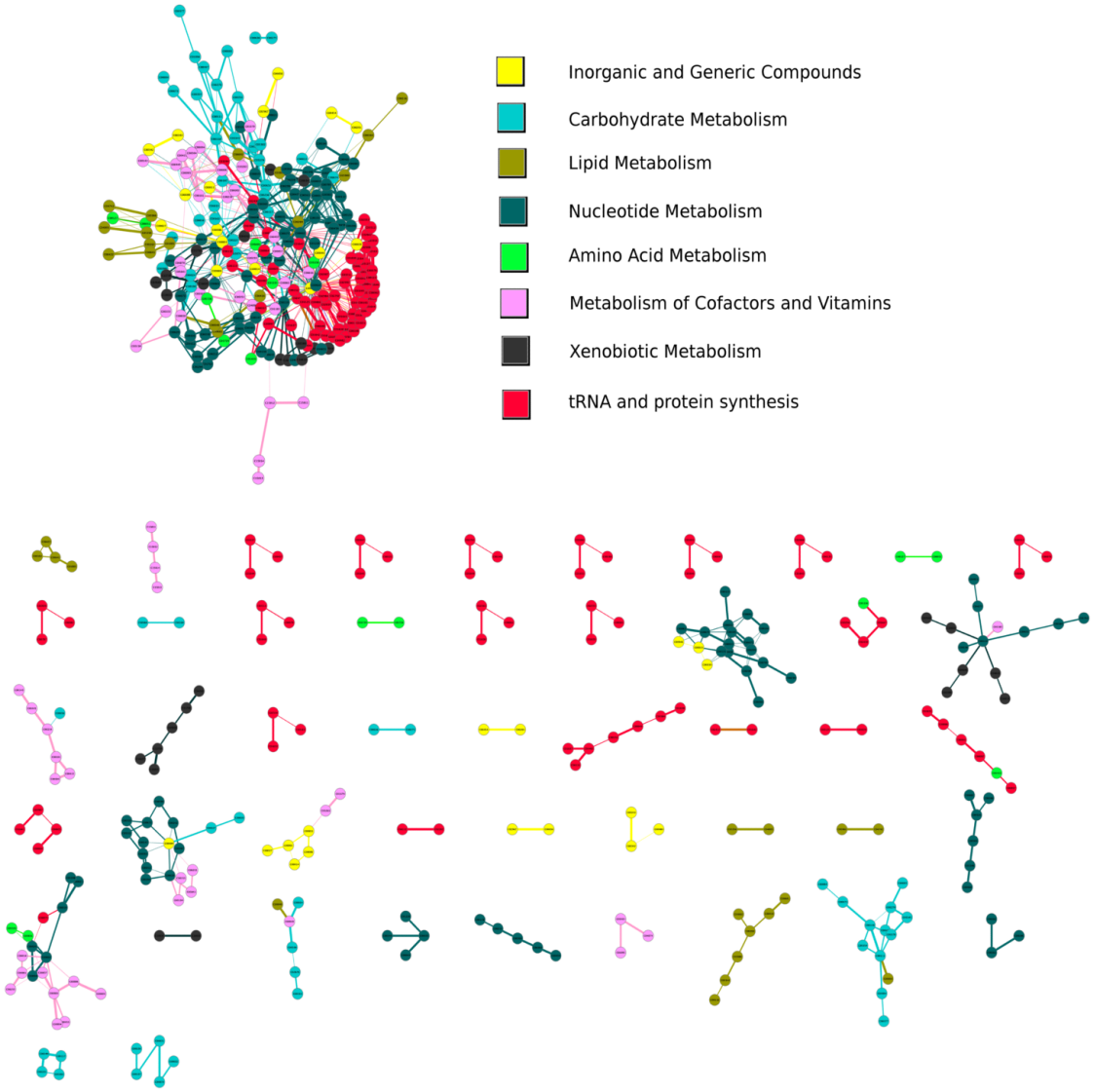

Figure 5.13: **The communities detected using our algorithm on the weighted metabolic network of *U. urealyticum*.**

There are at least two advantages to this weighted community detection on weighted graphs. First, this strategy is able to generate communities with varying degree of connectivity, from linear pathways to high-density subgraphs, as indicated by the results in Figure 5.10 and Figure 5.13. Second, as discussed above, thresholding weighted networks results in fewer fragments than naturalizing unweighted networks, keeping

the major component intact and a good representative of the entire metabolome, which is a desired property for studies that analyze this major component (e.g., [29, 30]). In this weighted representation, any reaction can be present as long as it has a reasonable weight. Further, algorithms that target weights in communities are capable of distinguishing generic reactant pairs, such as [NTP,NDP], owning to an undefined connection between these compounds, and more realistic ones. Last but not least, since the weighting is a ratio of chemical content, the absolute number of atoms involved can vary in a wide range enabling a multi-scale resolution.

In concluding this section, we would like to comment on the term "function". It seems that we have been seeking a community that matches the function. But on the other hand, without molecular experiment or biological insights, we are never clear what a function truly is . If biology was only composed of a simple networks, maybe we could confidently say that a function is equivalent to more interconnection. This scenario can be viewed as the case when all the edges are equally valued in biology. But unfortunately it is never this simple. When all the edges are weighed differently in biology, a functionally relatively independent module is far from being the same as a community structure where all of the edges are equally treated. All that we do in our study is to either exploit the naturalization method or weighting to *simulate* this functional coding under various edge semantics. We are not trying to *predict* functions based purely on simple networks as widely and intuitively recognized in many previous studies.

5.7 Conclusions and prospects

In this chapter, we conducted an evolutionary analysis of metabolic network modularity in order to explore whether it is the network modularity or the community structure on which the modularity score is based, is the unit of selection. We showed

that modularities undergo convergent evolution via different community structures. Further we revisited the association of the modularity score to environmental variability and extended the analysis to other aspects of microbial life styles. We found that on enzyme networks, the number of enzymes, which is also the size of the network and could also indicate the size of the metabolome, might be a determinant of the observed association between modularity and environmental variability. Further, we identified a strong association between network modularity and the microbe's temperature and oxygen requirements. We also found that modularity-based community structure does not correspond to biological functional classifications and is conserved only at the kingdom level.

An important confounding factor with metabolic network analysis is the network semantics, or what the nodes of the network represent and how the network is reconstructed. Previous studies have been based on different reconstructions and network semantics; for example, Parter et al. [29] considered networks with nodes representing metabolites while Kreimer et al. [30] considered networks with nodes representing enzymes. In order for the results to be comparable, we considered in this work four different alternatives (see Data). We found that the same analysis on different network reconstructions can lead to qualitatively different conclusions. For example, the correlation of modularity to the number of enzymes is only true for enzyme networks (Fig. 5.5 and Supplementary Material Fig. 9) but not for compound networks (Supplementary Material Fig. 10 and Fig. 11). For compound networks, we find a significant difference in normalized modularity among different groups but no clear association between modularity and habitat variability (Supplementary Material Fig. 12 and 13) in contrast with enzyme networks (Supplementary Material Fig. 14). We cannot repeat the association of network modularity to the environmental variability on compound network with currency metabolites deleted, as reported

in Parter et. al. [29]. Our result is consistent with a more recent analysis on an Archaean data set where no association was found either [233]. Discrepancy might result from the differences in the network reconstruction, algorithm used to optimize modularity or data used (due to different database releases). Despite different network semantics, it remains consistent that normalized modularity is significantly different among the groups classified by temperature requirements while not as significantly different among the groups classified by the oxygen requirements (Supplementary Material Fig. 15, 16 and 17) and that modularity scores are achieved via distinct community structures (Supplementary Material Fig. 18, 19 and 20).

Our work calls for more biologically meaningful definitions of the modularity for metabolic networks. Modules under such definition might not be graph-theoretically intuitive. Density-based definitions do not describe well pathways and sparse graphs which seem to be ubiquitous in biological systems (e.g., a biochemical pathway may be very sparse and does not fit the definition of a graph-theoretic module). Another drawback from defining modularity as a graph-theoretic concept in metabolic networks is that metabolic systems are inherently hypergraphs instead of standard graphs [266]. Adopting the graph-theoretic definition of modularity imposes a graph representation onto the metabolic system. Thus our work also calls for more careful scrutiny on the recent results related to the adaptive roles on modularity scores and their association with biological phenotypes. Adaptive roles should be explained under specific network reconstruction and care should be taken when one makes generalized conclusions.

Chapter 6

Introgressive descent of metabolic networks

6.1 Introduction

Evolutionary events responsible for biological genome evolution can happen on genomic stretches of various lengths, ranging from one single nucleotide [267], a fragment of several nucleotides [268, 269], a gene [270, 153], or an operon [271, 272, 273, 274], to the entire chromosome [275]. Evolutionary events that happen on the gene scale are of particular interest due to the central role of genes as fundamental functional units of the genome. These evolutionary events include mutational events such as duplication [119, 120], horizontal gene transfer [127, 128] and loss [123, 124, 125, 126]. Also included are non-mutational scenarios such as incomplete lineage sorting [276]—a phenomenon where different alleles of a gene in the population fail to completely sort during speciation, thus confounding the inference of gene genealogy.

In order to understand when and where these events take place and how they contribute to the organismal fitness under different environments, much work has been done to achieve an accurate reconstruction of these evolutionary events on a genome scale, a central topic of phylogenomics [277]—the study of gene and species phylogeny using genomic data. Extending phylogenetic analysis to the genome scale

not only allows for more comprehensive coverage of the genetic information of the organism and more accurate reconstruction of the species phylogeny [278, 279, 280], but also provides a more unequivocal inference of evolutionary events (e.g., loss can not be confirmed without the examination of the whole genome). Various models and methods were proposed to identify subsets, if not all [281], of these mutational events and to sort out the incongruence between the gene phylogeny and the species phylogeny. For example, in the case of horizontal gene transfer, which is believed to be ubiquitous in prokaryotic organisms [282, 283], the presence and absence of genes in each orthology family can inform gene gain (mostly due to horizontal transfer) and loss under the assumption of parsimony [284, 285]. Given more detailed information about the gene phylogeny, horizontal transfer can also be inferred by searching for its Maximum Parsimony (MP) [141], Maximum Likelihood (ML) [286] and Maximum Statistical Agreement Forest (MSAF) [287] reconciliation with the species tree (see [288] for an overview). In case of duplication and loss, parsimony-based methods [139] and Bayesian methods [289, 290] have been proposed.

Despite the long-standing recognition that most gene-scale evolutionary events can also happen to just fragments of genes (or domains) [291], most analyses of gene evolution are based on the concept of the *gene family*—the set of protein-coding genes that evolved from a common ancestor varying only in the local composition of each gene's sequence. This is because in practice it is much easier to assume the granuality of a gene and study the binary operations of creating, deleting or duplicating the gene than it is to account for the gene's finer building blocks, assembly or disassembly. In phylogenetics, the concept of the gene family allows the tree-like representation of a gene's genealogy. But if subgene scale evolutionary events primarily happen to gene fragments and sub-gene scale recombination is rampant, the gene family assumption, as well as the analyses based thereon, becomes problematic. If that is

the case, gene evolution defies a tree-like representation, and instead, resembles more closely a network representation where non-genealogical reticulation edges stand for introgressive descent [292].

Given a set of taxa, I partitioned genes into fragments such that each fragment is the largest stretch of DNA in the gene which either holds full or no homology in all the other genes from all taxa. I refer to these gene fragments as *modules* (following the terminology used in [293]). Fig. 6.1 illustrates this definition. Based on this definition, the module has not undergone further fission into smaller fragments, nor has it ever been broken in the middle so that it is partially recombined to form a different gene (see Fig. 6.1). In other words, modules can be taken to reflect the lowest level of granularity. The gene family assumption corresponds to the cases where each gene has only one module which covers the full length of the gene.

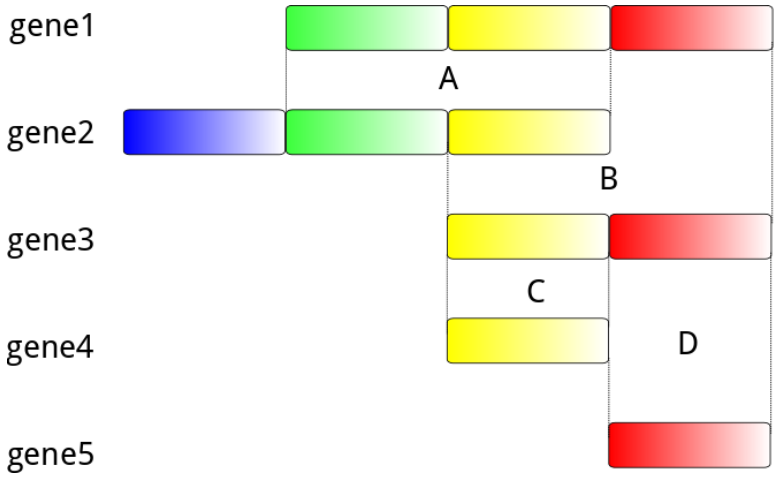

Figure 6.1: **Illustration of module definitions.** Color reflects homology.

Early methods have used protein structure information to identify domains . Protein domains are related to, but different from, the modules under my definition. Protein domains are defined based on knowledge of the 3-D structures, which are very rare [294] compared to the number of protein sequences annotated. The gene modules, on the other hand, are defined based on evolutionary conservation. Although

it is very likely that domains are evolutionarily conserved, there is no necessary 1-1 mapping between the two. In addition to structure-defined domains, most of the other analyses relied on sequence similarity results obtained by performing all-against-all Blast to detect sub-gene level modules [291, 295, 296, 293]. Starting from all the pairwise comparison of sequences, the detection of modules amounts to partitioning the *sequence space graph*—a graph where nodes represent sequences and nodes are connected through alignments—into densely connected subgraphs [297, 298, 299, 291]. For example, recent analyses by Wu et al. [293] relied on the ADDA algorithm—a method which optimizes the likelihood that two sequences align due to sharing a common domain—for domain decomposition and domain family clustering [291]. In addition, Leonard and Richards [296] detected fusion and fission events in fungi by looking for different numbers of hits from "blasting" sequences from opposite directions (the fdfBLAST pipeline). The identified modules are further validated by comparing against curated domain family databases such as Pfam [300, 301, 302] and CDD [303, 304].

Once modules are detected and clustered, methods for identifying evolutionary events on the gene scale can easily be extended to the module scale. Virtually all evolutionary events including horizontal transfer, duplication, loss and incomplete lineage sorting have been previously studied on a whole-gene scale, but can happen on the module scale. In addition, the physical concatenation and split of modules defines two additional mutational events—gene fusion and fission. Recently, novel phylogenomic methods have been developed to identify gene fusion and fission in eukaryotic organisms Drosophila [293] and fungi [296]. Wu et al. [293] reconstructed, for a Drosophila dataset, the subgene level evolutionary history of gain, loss, duplication events as well as gene fusion and fission. Their reconstruction accounts for module synteny (or "architecture" under their terminology). Several simplifications

were performed such as the collapse of tandom repeats into one module and the exclusion of very large, or promiscuous, module families [293]. Leonard and Richards [296] identified 63 gene fusion events from 9 fungal genomes. For these putative gene fusions, they reconstructed the state—whether genes are fused or not—on the phylogenetic tree encompassing 115 fungal genomes using the maximum likelihood method implemented in the Mesquite software [305].

Although many analyses of gene-scale evolutionary events have focused on the adaptive role of these events in shaping the prokaryotic genome by looking for synapomorphy[1] and homoplasy[2], it is intriguing also to not only ask whether the gene family assumption still holds in the cases of whole genes, but also what the adaptive role the module organization of genes play. Microbial metabolic genes are ideal targets of study due to the richness in functional knowledge about metabolism as well as the existing methods for modeling the metabolic system. Microbial metabolism is known for its diversity. It enables microbes to colonize virtually all habitable environments on earth [306, 307]. Understanding the metabolic system sheds light on the system's engineering, which has great application in medicine, energy and environmental industry [308, 309]. Some initial observations on the adaptive role of the horizontal transfer of metabolic genes have been reported. For example, Pal et al. [2] found, by analysing the metabolic gene family and the metabolic network of *E.coli*, that microbial metabolism evolves through transfer of genes responsible for transporting external metabolites. They also found that the unit of transfer is usually on a supragenic level—genes responsible for reactions whose fluxes are coupled in the metabolic network tend to be transferred together in an operon structure [2].

To the best of my knowledge, no pre-existing analysis has used phylogenomic methods to analyze the adaptive role of subgene evolution in large prokaryotic datasets

[1]similarity in genotype due to inheritance
[2]similarity in genotype due to convergent evolution

that span several subphyla. In contrast to [293] and [296] which study eukaryotic organisms, I analyzed the subgene evolutionary history of bacteria, or, more specifically, proteobacteria. Unlike eukaryotic organisms, horizontal gene transfer is rampant in bacteria [2] and archaea [310, 311] (where the Dollo parsimony[1] [312, 313] might not hold). The size of my dataset is beyond the population level in constrast to the closely related taxa set used in [293]. This challenges the identification of module families since a module family can contain up to millions of components.

In order to investigate the role of the module organization of genes, I define mosaicity as the number of modules in a gene minus 1. Genes with only one module, or mosaicity 0, are non-mosaic. Focusing on mosaicity of genes in *E.coli*, I found that most single-copy genes are non-mosaic. Around one half of the *E.coli* protein-coding genes are mosaic. Metabolic genes are more mosaic than non-metabolic genes. Among the metabolic genes, genes responsible for outer membrane transporters are the most mosaic. For non-metabolic genes, genes responsible for proteins that bind DNA or ions are the most mosaic. No overall association between the age of transfer and mosaicity is seen.

6.2 Methods

6.2.1 Data

Fig. 6.2 shows the dataset I investigated. 62 taxa are sampled in the proteobacteria phylum following [2]. There are 12 α-proteobacteria, 7 β-proteobacteria, 34 γ-proteobacteria, 3 δ-proteobacteria and 6 ϵ-proteobacteria. The 6 ϵ-proteobacteria are treated as outgroups following [2]. The gene sequences and gene family (COG) information for these 62 proteobacteria were retrieved from STRING v9 [314]. There

[1] Mutations are irreversible

are 3785 orthology families present in the chosen data set and 3655 of them have at least 2 sequences (130 families are ORFans [152, 315]). Each sequence may correspond to an entire gene or just a segment of the gene.

Altogether, there are 3520 *E.coli* (NCBI taxon id: 511145) genes in the data set. These genes include both metabolic and non-metabolic ones. 886 out of 904 metabolic genes (98%) in the iJR904 model [60] find a gene family assignment. 158 metabolic gene families have more than one sequence, a sign of gene duplication. 804 metabolic genes of *E.coli* find orthologs in *Y. Pestis* (NCBI taxon id: 214092), while 860 find orthologs in *S. enterica* (NCBI taxon id: 99287) and 788 in both (which is similar to what was reported in [2]).

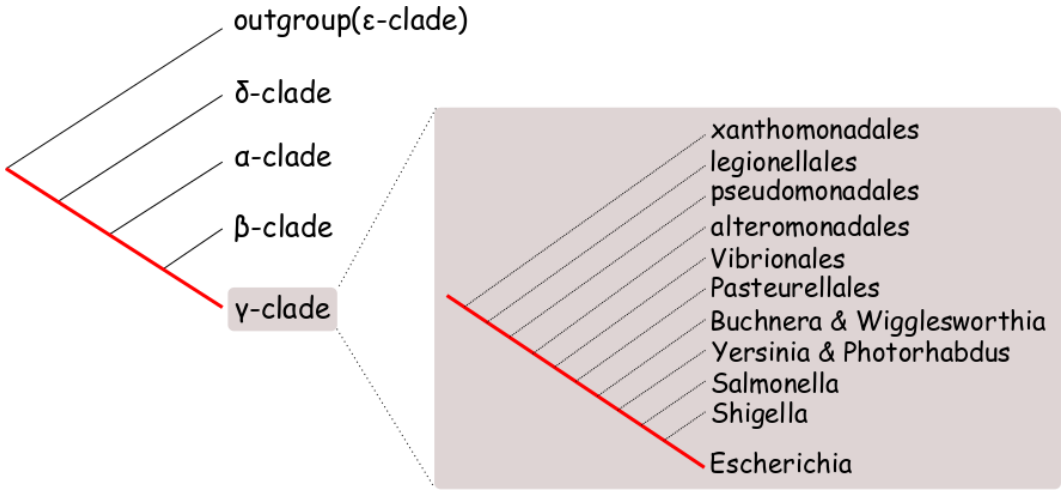

Figure 6.2: **Schematic diagram of the proteobacteria dataset.** Only the branches leading to *E.coli* are shown. All of the other leaves are collapsed clades.

6.2.2 Gene tree inference

Sequences in each gene family are aligned using MUSCLE v3.8 [316] with a default setting. Two gene families (COG3319 and COG2931) which MUSCLE does not accomodate are aligned using ClustalW v2.1 [317] with a default setting. To build gene trees, I first remove 111 gene families with only two taxa. I infer the maximum likeli-

hood tree with 100 bootstraps on 3408 gene families using RAxML v7.3.1 [318] with the JTT substitution model and 20 inferences to avoid local optima. On 100 gene families with three taxa, I use PhyML v3.0 [319] with its default setting. Trees of the remaining 36 gene families with thousands of taxa are inferred using RAxML but with a reduced number of inferences and number of bootstraps. I also run PhyML with 100 bootstraps on 2407 gene families with fewer than or equal to 50 taxa. I compare the gene trees inferred via RAxML and via PhyML. Much incongruence is found.

6.2.3 Species tree inference

Species trees are obtained using 3 methods: 1) Concatenation tree: I concatenate the sequence alignments of 44 gene families with one single-copy from each of the 62 taxa. The species tree is inferred from the concatenated alignment using RAxML with 100 bootstraps. 2) Minimizing Deep Coalescence (MDC) tree: the MDC tree is computed from 3543 gene trees (excluding families with only two taxa and COG1028) using PhyloNet [135]. 3) Minimizing Duplication and loss (MDL) tree: the MDL tree is built from 3543 gene trees using DupTree [139]. I also compare the inferred species tree with the NCBI taxonomy tree. The MDC tree slightly differs from the concatenation tree. The concatenation tree is congruent with the tree published by Pal et al. [2]. The MDC tree does not identify the outgroups assumed, which are Camphylobacters, Helicobacters and Wolinella. Meanwhile, even though Bdellovibrio is not an assumed outgroup, it is grouped together with the outgroup members in the MDC tree. The number of extra lineages is not high for Bdellovibrio. Nor are the numbers of extra lineages for the two Campylobacters. Other incongruences exist too. The MDC tree has a lower number of extra lineages (300355 vs. 345870) than the concatenation tree, ruling out the possibility of non-optimality in searching the

tree space for the MDC tree.

6.2.4 Flux balance analysis

Flux balance analysis [54, 320, 321, 322, 323] is a powerful framework. It utilizes the metabolic reaction stoichiometry to estimate the steady state reaction fluxes (balanced in the sense that the in-flux of a metabolite equals the out-flux under steady state) based on constraints on the reaction fluxes and an objective function, which is typically a function that defines the growth of the cell. Typically the gene-enzyme-reaction relationship, the metabolic reaction stoichiometry, thermodynamic constraints and biomass composition formula are collected in a procedure called metabolic reconstruction [324, 55]. I studied the metabolic reconstruction model iAF1260 [64]. For revisiting the results of Pal et al. [2] and for backward compatibility, I also looked at the earlier iJR904 model [60]. I have used the COBRA toolbox [325] and FASIMU [326] for the task of optimizing flux-balance models. The flux coupling analysis was done using the software F2C2 [327]. The gene knock-out phenotype was also evaluated using the Minimization of Metabolic Adjustment (MOMA) framework [328], where one seeks a solution with minimal euclidean distance to the wildtype flux distribution instead of optimizing the growth objective function. The quadratic programming problem is solved using the IBM ILOG CPLEX optimizer®.

6.3 Whole gene analysis

6.3.1 Microsynteny of horizontal transfer: a case study of threonyl-tRNA synthetase

Threonyl-tRNA synthetase (ThrRS) was originally reported to have been transferred from marine γ-proteobacteria to *P. marinus*, a cyanobacterium [128]. In this study,

I take a closer look at this case by comparing the microenvironment of ThrRS in the source and in the target. I specifically choose Prochlorococcus marinus str. natl1a (taxid: 167555) as the target and Hahella chejuensis kctc 2396 (taxid: 349521) as the source of the transfer. I sampled 100 taxa randomly from bacteria, keeping the source and target species, and, meanwhile, keeping only one species under each genus. Additionally, I merged 229 γ-proteobacteria and 37 cyanobacteria, all including the source species and target species chosen.

Protein family assignments were retrieved from the eggNOG database v3.0 [329]. Species were built by the multiple sequence alignment of 16s rRNA sequences. Multiple 16s rRNA might exist. Only the longest sequence was used. Variable segments of the alignment were filtered using Gblock [330]. The maximum likelihood tree was obtained using the PhyML software [319].

In Fig. 6.3, I plotted the micro-synteny of ThrRS in both the clade of γ-proteobacteria and cyanobacteria. Similarity is measured by the Jaccard coefficient of the intersection of the neighbor configuration 3 genes upstream and 3 genes downstream of ThrRS. Orthology assignment is extracted from the HOGENOM database [331]. The genes nearby are very different between the original copy and horizontally transferred copy. Similarity of microsynteny only extends to very closely related species. I scrutinized the start and end region of the ThrRS gene copy in the two species. I found that at this scale, not only the genes nearby are completely different, but the base pairs next to the starting and terminating codons share no similarity either.

6.3.2 Inference of gain, loss and duplication

Following [285], I infer parsimonious scenario of gene gain and loss from the concatenation tree, and 3785 gene trees using PAUP [332] with the AccTran option. States are encoded in two ways with states of the tree node encoded in: 1) presence/absence

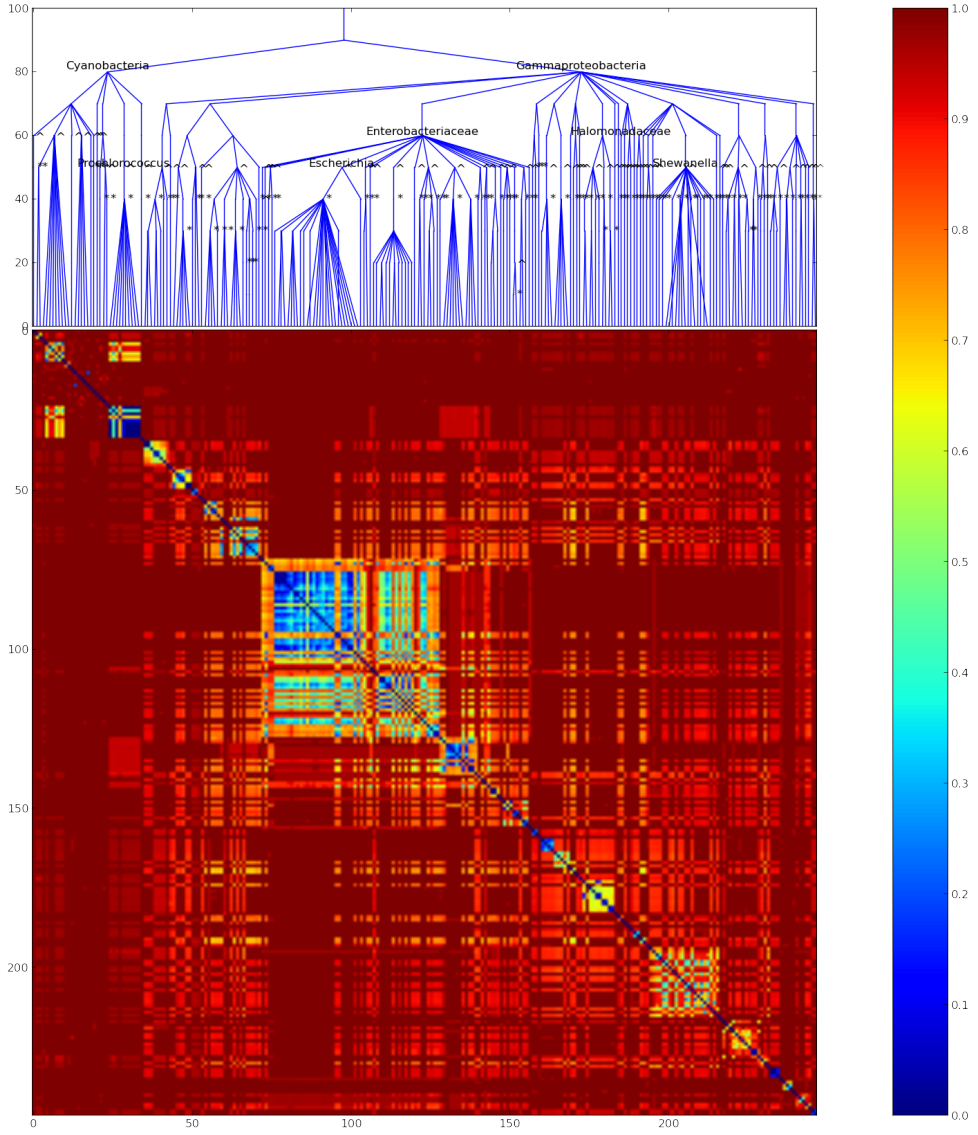

Figure 6.3: **Microenvironment of ThrRS in γ-proteobacteria and cyanobacteria.**

(binary), or 2) copy number of the genes in the family. In the binary encoding, I consider two major classes of mutational events: gene gain and gene loss. Gene gains

are mostly attributable to horizontal gene transfer, as *de novo* innovation of a new gene is rare in microbial evolution after the archaen expansion [333, 334]. Under the binary encoding, duplication events cannot be inferred. The cost ratio

$$\text{gain} : \text{loss} = 2 : 1$$

following [335, 284]. The PAUP block in the inference looks like the following,

```
BEGIN PAUP;
      log start file=dlt.log;
      exe charactermatrix.nxs;
      exe costmatrix.nxs;
      exe tree.nxs;
      ctype GeneCopy:All;
      pset Opt=AccTran;

      taxset out = t85963 t85962 t273121 t195099 t235279 t192222;
      outgroup out;
      set root = outgroup outroot=monophyl;
      root root = outgroup;

      describetrees 1 /plot=phylogram Xout=both ApoList=Yes
          ChgList=yes; [brLens=yes;]
      savetrees from=1 to=1 format=ALTNEX root=yes
          file=relabeled.tre;
END;
```

From the parsimonious inference of gain and loss, I have identified 6537 gains (4753 are unambiguous, around 72.7%) and 15452 losses (13431 are unambiguous, around 86.9%). This compares to 2808 gains as reported by [2]. These results are reasonable considering the growth in the coverage of the orthologous family data to 3785 as compared to 2325 as is used in [2]. There are 8 orthologous families with 8 independent gains inferred. I validated one orthologous family—COG0830 (urease accessory protein, UreF) where 8 independent gains are inferred. I blasted the UreF sequence against the protein sequences from the STRING v9.0 database. I ruled out the possibility that these gains are a result of missing orthologs in the orthologous family.

Under the encoding of gene copy numbers, a taxon with n ($0 \leq n < 20$) copies in a gene family is in the state of n. Taxa with greater than or equal to 20 copies in a gene family are regarded as one state. The cost ratio is:

$$\text{complete gain} : \text{complete loss} : \text{duplication} : \text{othergain/loss} = 50 : 25 : 5 : 1.$$

Complete gains and complete losses are defined as transitions from and to the state 0. Other increases or decreases in copy number are regarded as partial gains and partial losses. The only exception is that the increase of the copy number from 1 to 2 is considered duplication. In other words, duplication events that happened when the copy number is larger than 2 could not be unequivocally identified. Such expansion in the copy number is considered a partial gain, as opposed to a complete gain whose start state is 0. Note that under my encoding duplication is indistinguishable from horizontal transfer without replacement.

6.3.3 Genes gained along the *E.coli* evolutionary history

More than 200 horizontal transfer events have taken place in *E.coli* after its split from Salmonella around 100 million years ago [336]. These horizontal transfers have affected around 17% of the total number of ORFs.

There are, altogether, 769 gain events inferred from 763 gene families and 572 loss events inferred from 561 gene families (some gene families are gained or lost more than once at different times of the *E.coli* history). Gains and losses can be ambiguous due to ties in parsimony score. I only consider unambiguous events (which is about 72.7% for gains and 86.9% for losses). I plot the number of gene gain/loss events along the history of *E.coli*. Each time point labels the clade that gets split from the *E.coli* lineage.

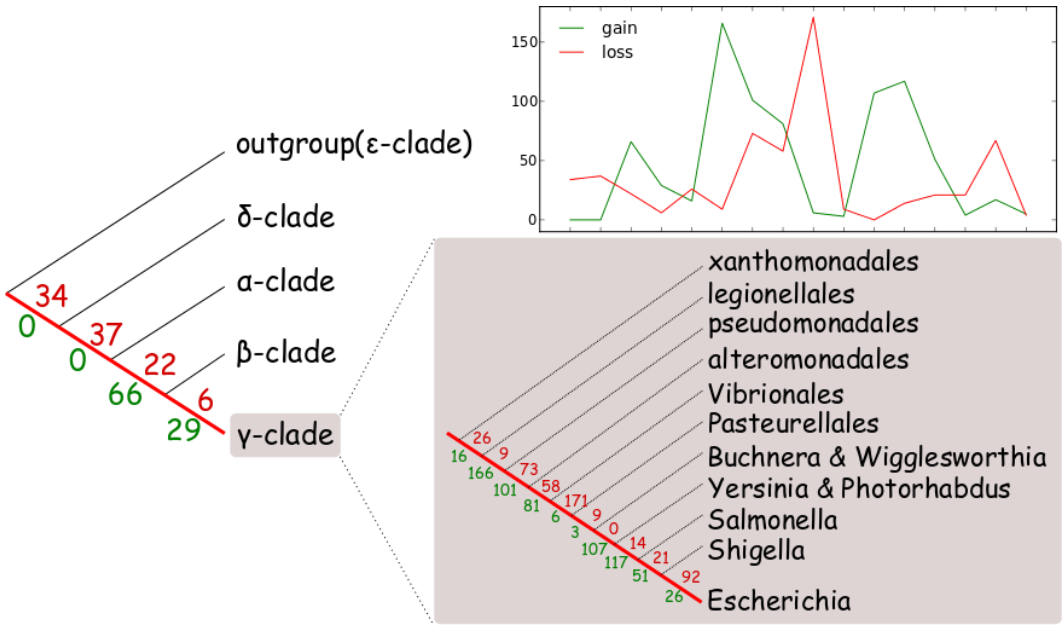

Figure 6.4: **Gene gain and loss in the *E.coli* history.**

I identified three peaks of gains: 1) between the split with α-proteobacteria and the split with β-proteobacteria; 2) between the split with legionellales and the split with pseudomonadales; 3) from the split with Buchnera to the split with Yersinia and

to the split with Salmonella. There is also one peak of losses: from Alteromonadales to Vibironales and sustain until the split with Pasteurellales. 77 gene families have been gained since the split with Salmonella, 194 gained since the split with Yersinia and 301 gained since the split with Buchnera.

6.3.3.1 Annotation analysis of gained gene.

It has been reported that recent horizontal gene transfer preferrably takes place in genes responsible for transporting external nutrients [2]. I studied the meaning of these recent gains by analyzing their annotation. I took the set union of all the words in all the gene annotations from the STRING database [314]. I tabulate the 50 most frequently seen words in the annotations, weeding out the words with no specific biological meanings. The filtered words are: 'protein', 'putative', 'hypothetical', 'of', 'the', 'family', 'predicted', 'binding', 'subunit', 'in', 'conserved', 'like', 'and', 'domain', 'a', 'component', 'system', 'for', 'related', 'containing', 'dependent', '1', '2', '3', '4', '5', '6','A', 'B', 'type', 'involved', 'to', 'is', 'beta', 'alpha', 'peptide', 'By', 'function', 'May', 'Involved', 'chain', 'similarity', 'by', 'role', 'L', 'D', 'cell', 'that', 'it', 'from', 'The', 't', 'two', 'activity', 'gene', 'be', 'with', 'unknown', 'an', 'Part', 'I', 'II', 'III', 'c', 'EC', 'fused' and 'Required'.

The genes that are gained are classified into two groups based on the gain time (i.e, before and after the split with Buchnera). The most frequently seen words in gains are: "membrane" (141), "transport" (96), "transporter" (91), "inner" (82),"transmembrane" (75), "regulator" (65), and "DNA" (56). The word profiles from gained genes and non-gained genes are significantly different ($\chi^2 = 239.8$ $df = 50$, $p < 2.2e-16$). The word profiles of genes gained at early history and later history are significantly different ($\chi^2 = 110.4$ $df = 50$, $p = 1.88e-6$). Therefore I concluded that horizontal gene transfers have functional preferences and genes gained at the early stage and the

recent stage are functionally different. My conclusion is consistent with Pal et al. [2] in that most observed words are related to cross-membrane transport.

6.3.3.2 Ontology of recent gains

I then move on to a more formal analysis of gains in gene ontology (GO [262]). The GO association is retrieved from the genbank and GO consortium [262]. Enrichment analysis is done using goatools (`https://github.com/tanghaibao/goatools/`). Statistical significance is evaluated using threshold of p = 0.05. Anaerobic respiration, phosphotransferase-related, carbohydrate-transport (especially carboxylic acid) transport-related, ion-transport related genes and genes for generating precursor metaboites are susceptible to transfer. I have also identified transposition-related genes as being susceptible to transfer. Biosynthesis-related genes (e.g., translation) and heterocyclic compound binding-related genes are resistant to transfer.

6.3.3.3 Gene gain vs. COG functional category

Using the functional classification from COG, I observe from Fig. 6.5 that the function of many gained genes are poorly characterized. The most gained functional categories are: 1) RNA processing and modification; 2) intracellular trafficking, secretion, and vesicular transport; 3) signal transduction mechanism; 4) secondary metabolites biosynthesis, transport and catabolism; 5) carbohydrate transport and metabolism; 6) signal transduction mechanism. As is also shown in the bottom row of Fig. 6.5, genes related to defense mechanisms are relatively recently gained. Genes from categories “cell motility”, “intracellular trafficking, secretion and vesicular transport”, “nucleotide transport and metabolism” and “coenzyme transport and metabolism” are anciently gained.

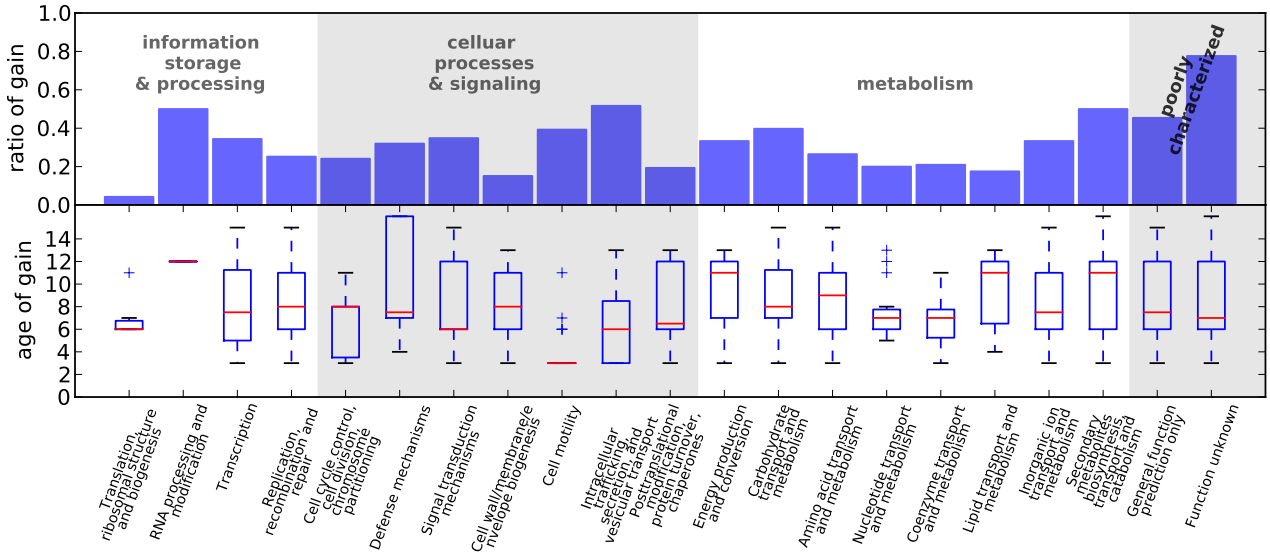

Figure 6.5: **Ratio and age distribution of transferred genes in each COG functional category.** Only genes in the *E.coli* history are plotted.

6.3.3.4 Gain vs. reaction subsystems

In the iJR904 model [60], reactions are classified into 28 subsystems, excluding "Putative" and "Unassigned". Using the gene-protein-reaction relationship, I studied 942 cases where metabolic genes were assigned to one subsystem. I found that the four subsystems with the highest rate of gain are: 1) putative Transporters, 40% (8 out of 20) genes were gained; 2) Pyruvate Metabolism, 38% (8 out of 21) genes were gained; 3) Transport, Extracellular, 30% (55 out of 181) genes were gained (the counting is consistent with the network position analysis above); 4) Alternate Carbon Metabolism, 30% (38 out of 128) genes were gained. A detailed analysis of the pyruvate metabolism case is given below.

The 12 subsystems with the lowest ratio of gains are: Glutamate metabolism (0%), Folate Metabolism (0%), Methylglyoxal Metabolism (0%), Histidine Metabolism (0%), Anaplerotic reactions (0%), Cell Envelope Biosynthesis (3%), Purine and Pyrimidine

Biosynthesis (5%), Valine, Leucine, and isoleucine Metabolism (5%), Membrane Lipid Metabolism (7%), Pentose Phosphate Pathway (8%), Glycine and Serine Metabolism (8%) and Cysteine Metabolism (10%).

6.3.3.5 Transposon/virus-related genes are more susceptible to gain

I also studied the transposon/virus-related genes using the annotation from Ecocyc v14.1 [337]. I identified 81 virus/transposon genes out of 4450 *E.coli* genes. As shown in Fig. 6.6, a higher fraction of virus/transposon related transfers are found in recent branches. However, this is not true for all of the recent branches and the branch with highest fraction of virus/transposon-related transfers is the branch from its split from the Pasteurellales clade to its split from Buchnera and Wigglesworthia clade. This is different from the clear trend of reduction in the fraction of virus/transposon-related transfers reported in [2]. But it still holds that ancient transposon/virus-related gains are extremely rare (Fig. 6.6).

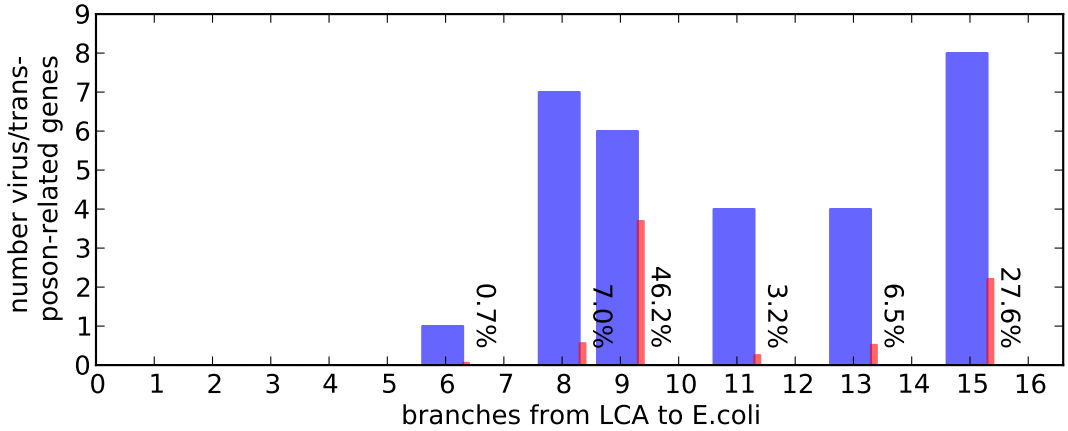

Figure 6.6: **Frequency of virus/transposon-related genes in the transfer.** The red bar is proportional to the fraction of virus/transposon-related genes in all of the genes gained. The fraction value is labeled on the right side of the red bar. The closer the branch is to the right, the closer it is to *E.coli* (and further from the LCA).

6.3.3.6 Placement of gained genes in the metabolic network

Following [2], I divided genes into: 1) genes for transporter proteins; 2) genes for first reactions after transport; 3) medium genes; and 4) biosynthetic genes. I identified 136 genes for 181 transport reactions where 88 major nutrients are transported. I

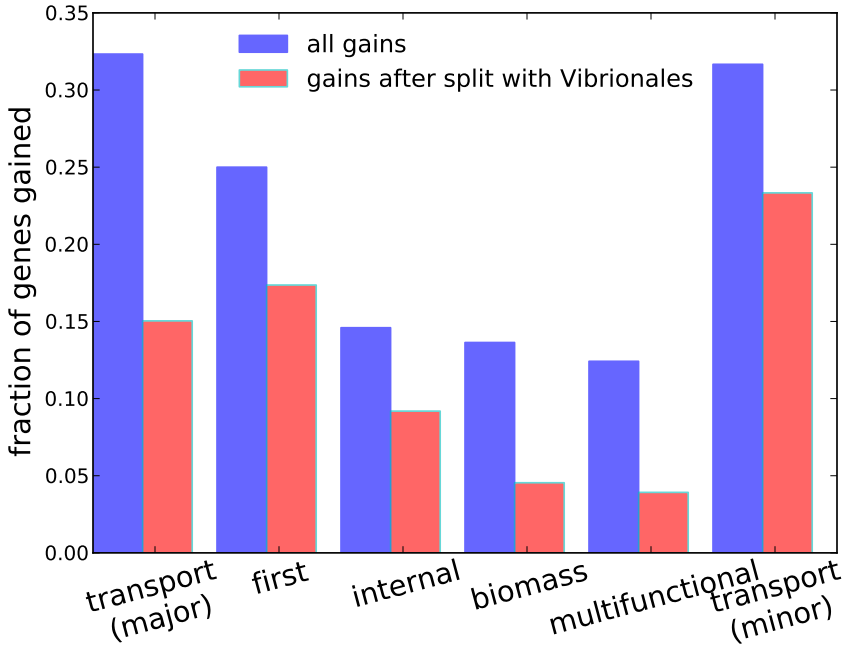

Figure 6.7: **Position of gene gained in the** *E.coli* **network.** blue) Genes are gained at all stage in the *E.coli* evolutionary history. red) Genes are gained after the split with Vibrionales.

found that the fractions of horizontally transferred genes at different positions of the network are significantly different ($\chi^2 = 21.9669, df = 5, p = 0.000531$). In Fig. 6.7, I plot, for each of the four categories, the fraction of genes that were transferred as well as the gene that were involved in more than one category (labeled "multifunctional") or involved in the transfer of non-major nutrients (labeled "transport (minor)"). I found that genes involved in the transportation of nutrients, either major or non-major, are most often transferred. Genes involved in biomass synthesis are the least often transferred. Genes responsible for first reactions after transport get transferred

more often than genes responsible for other internal reactions. This is consistent with [2]. In addition, multifunctional genes are resistant to horizontal gene transfer, which makes sense in light of the complexity hypothesis [154].

6.3.4 Gene gain and loss in metabolic pathways

I studied the ratio of the number of gains, the number of losses and the two combined over the total number of genes in 361 metabolic pathways as curated in the EcoCyc database [337]. I found that the six most gained pathways (with respect to the total number of genes in the pathway) are: 1) formaldehyde oxidation II (glutathione-dependent); 2) galactitol degradation; 3) 4-hydroxybenzoate biosynthesis II (bacteria and fungi); 4) asparagine biosynthesis I; 5) homoserine biosynthesis; 6) 4-aminobutyrate degradation I. The six most lost pathways are: 1) pyruvate oxidation pathway; 2) galactitol degradation; 3) polymyxin resistance; 4) p-aminobenzoate biosynthesis; 5) pyrimidine ribonucleotides interconversion; 6) formaldehyde oxidation II (glutathione-dependent); Interestingly, formaldehyde oxidation II pathway and galactitol degradation pathway appear in both lists, indicating that they involve constant gains and losses. In Fig. 6.8, I plotted the distribution of the number of gains and losses and the ratio of these two events over the total number of genes in each metabolic pathway curated in the EcoCyc database [337]. I found that out of 361 pathways, about 200 pathways involve gain and loss of genes. About 50 pathways involve more than 5 events. Most of these pathways involve both gains and losses. Some pathways underwent heavy horizontal gene transfer but for most pathways, each genes undergoes 2 gain/loss on average.

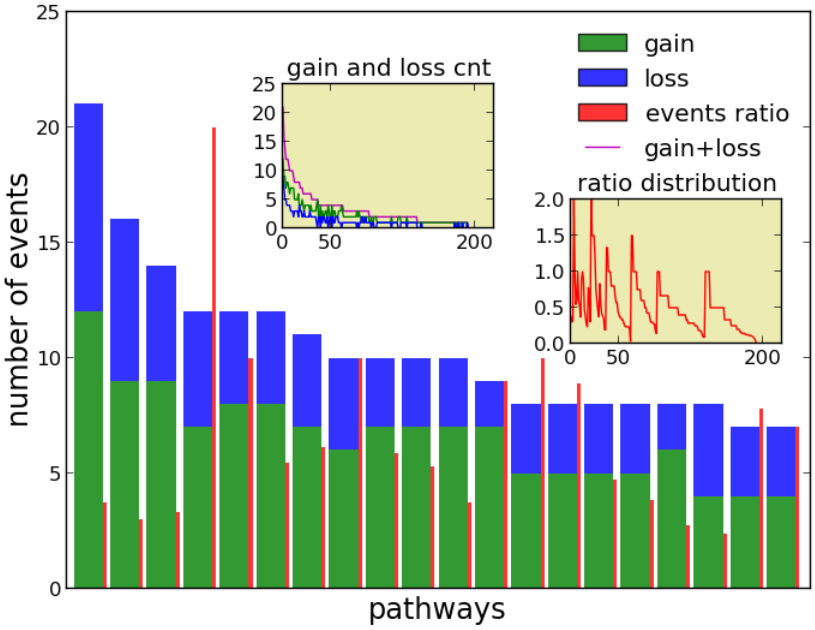

Figure 6.8: **Gain and loss of genes in metabolic pathways.** Pathways are curated in the Ecocyc database. Events ratio is the ratio of the sum of gain and loss events (which can happen to any gene in the pathway) over the total number of genes in the pathway. Pathways are reversely ordered by the sum of the numbers of gain and loss events.

6.3.5 Fitness contribution vs. gain and loss

I measured the fitness contribution of each metabolic gene by optimizing the biomass accumulation using constraint-based modeling. The formula for biomass composition follows [60]. I computed the biomass accumulation under any combination of 136 growth conditions and the single knock-out of 904 genes. Each growth condition is characterized by: 1) a carbon source and 2) the availability of oxygen. A scenario can be lethal if the ATP maintenance requirement is not met (and the model becomes infeasible).

The fitness contribution $f(g, c)$ of a gene g under a specific growth condition c is defined as the reduction in biomass accumulation flux after the gene gets knocked

out. The contribution of the gene to fitness is significant if the knock-out of the gene reduces the fitness by more than 1% or leaves the model infeasible. In cases where the model remains feasible after knock-out, I define the contribution as 1 - [percentage of reduction in biomass accumulation after knock-out]. I compared the fitness calculation from the earlier iJR904 reconstruction and the new iAF1260 reconstruction for each combination of growth condition and gene knockout (out of 888 common genes). The fitness has a bimodal distribution (left panel of Fig. 6.9). In the middle and the right panel of Fig. 6.9, I plotted the comparison in fitness contribution of genes under the iJR904 model and the iAF1260 model. I found most of the fitness contribution calculations are in agreement with the two models. Under only 174 (out of 120768) settings are the fitness calculation different.

I further define the global fitness contribution $F(g)$ of a gene by the number of growth conditions under which the gene has a significant fitness contribution, i.e.,

$$F(g) = |\{c \mid f(g, c) \text{ is significant}\}|.$$

As is consistent with [2] (see Fig. 6.10), I found a significant correlation between the number of gains and losses and the environmental specificity (ANOVA $F = 8.69, p = 9.7e - 15$).

Because of the bimodal distribution of the f, I also plotted the fraction of genes that contribute to fitness in at least 1% of the growth conditions, all growth conditions, and the average number of growth conditions of all genes gained/lost for a specific number of times. Again, the trends in these tree plots are very similar as shown in Fig. 6.11.

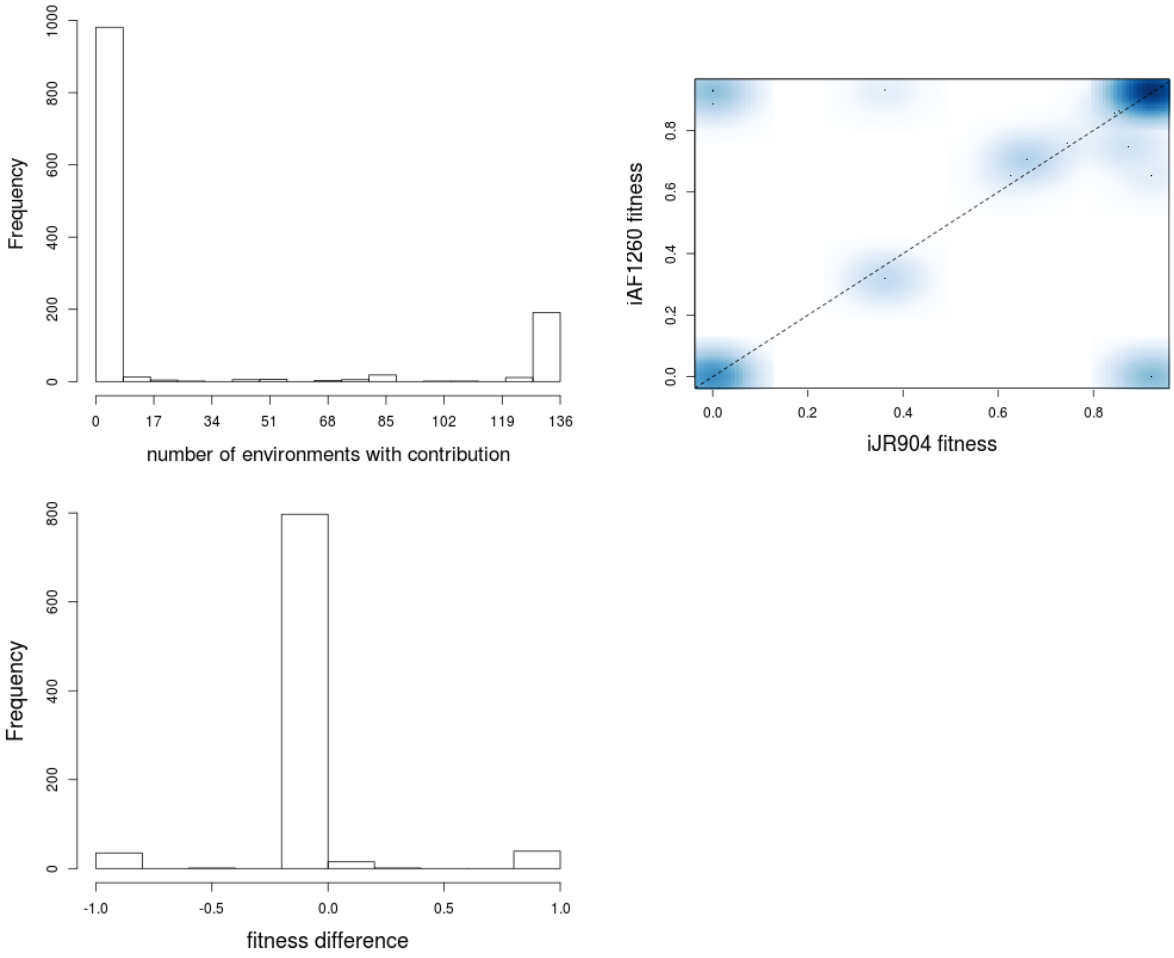

Figure 6.9: **Distribtuion of fitness contribution and model comparison.** Top left panel depicts the distribution of the fitness contribution. Top right panel and the bottom left panel correspond to comparison of between fitness contribution calculated under iJR904 and iAF1260. The fitness contribution of 888 genes under aerobic condition with glucose as the sole carbon source.

6.3.6 Conclusion

In conclusion, results from whole gene analysis based on the more comprehensive metabolic reconstruction (iAF1260) are qualitatively the same as the earlier results reported by Pal et al. [2]. Genes that contribute to fitness in a small number of environments are more susceptible to transfer. Genes that are responsible for transporters are frequently gained or lost. Biosynthesis-related genes and heterocyclic compound binding-related genes are resistant to transfer.

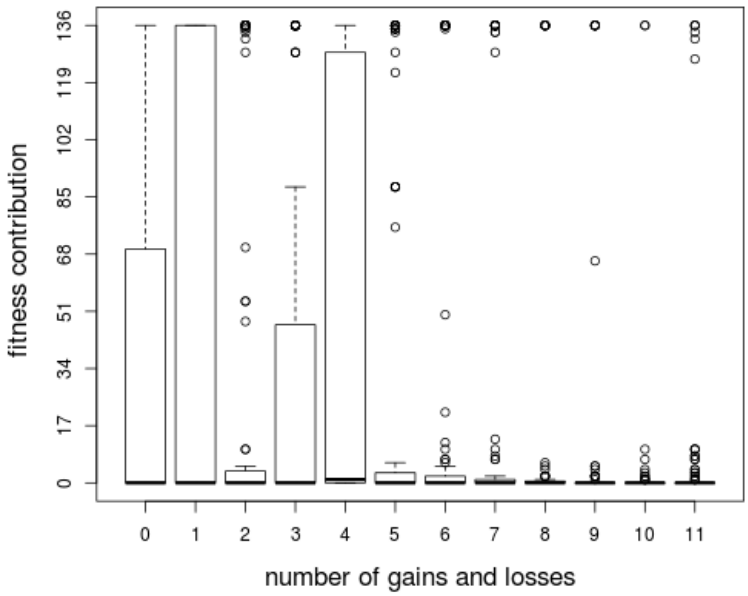

Figure 6.10: **Overall fitness contribution vs. the number of gains and losses I.** The number of gains and losses are counted over the tree of the entire dataset.

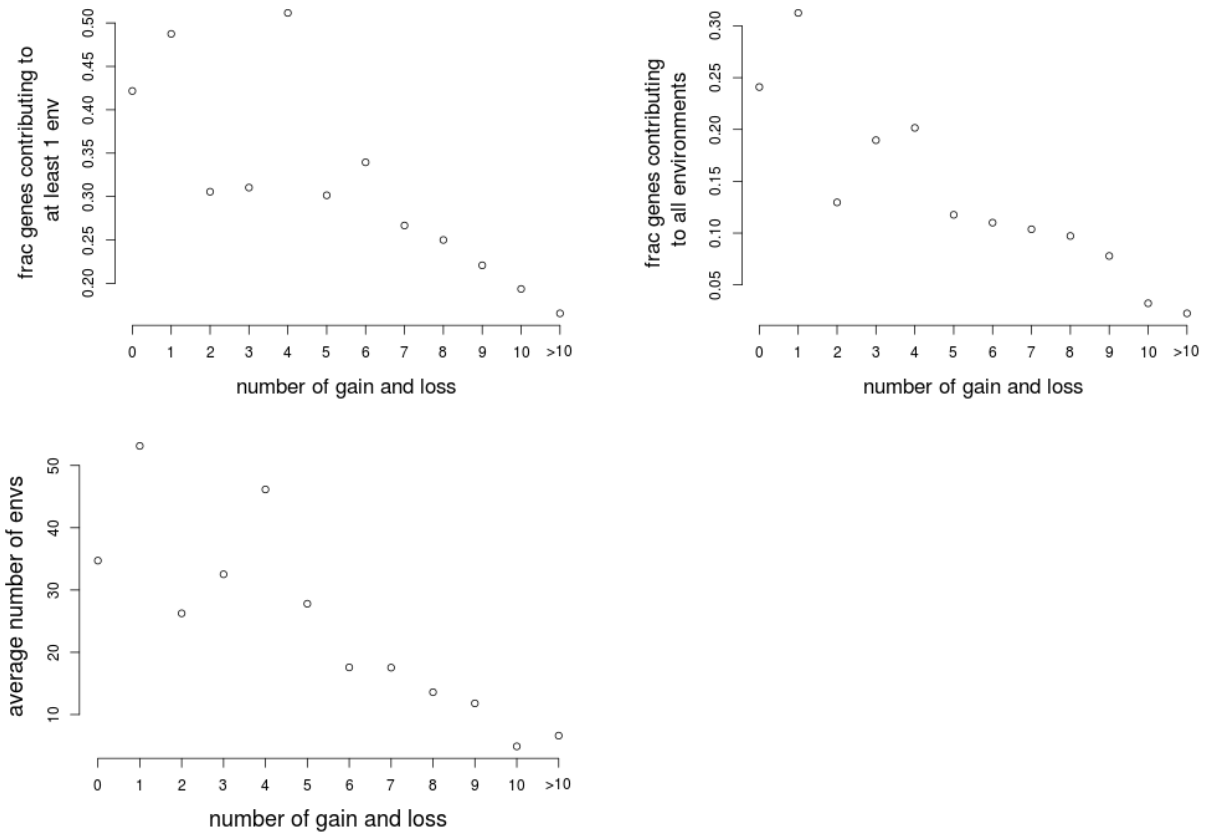

Figure 6.11: **Overall fitness contribution vs. the number of gains and losses II.** Left) The fraction of genes that contribute to fitness under at least 1% of the growth conditions. Middle) The fraction of genes that contribute to fitness under all the growth conditions. Right) The average number of growth conditions (of all the genes that are gained and lost for a certain number of times) under which the gene confer a significant contribution to the fitness.

6.4 Mosaicity of *E.coli* genes

6.4.1 Module detection (target database: 56 proteobacteria)

Gene fusion and fission were first systematically studied by Snel et al. [338], who found that gene fusion was more likely to occur than gene fission in prokaryotic organisms [338]. Moreover, high temperature favors gene fission in thermophiles, which results in lower error in transcription, translation and protein folding [338]. Most microbial organisms harbor up to 20% segmentally variable genes [339]—genes with long, highly variable regions together with other well-conserved regions. Previous attempts to identify domain structures in protein sequences include SCOP [340, 341], CDD [303, 304], SMART [342, 343], EVEREST [344], Pfam [300, 301, 302], and IntroPro [345, 346]. The module detection in my analysis is based purely on sequence similarity. I choose to start from scratch to avoid coverage problems in any existing databases. Each annotated protein-coding gene of E.coli is compared against protein sequences from 56 proteobacteria using blastp [347]. For each protein in *E. coli*, I built a HSP-graph where nodes represent HSP and two HSPs are connected only when the aligned regions of the protein in the two HSPs overlap by 70% (of both alignments). For each gene's HSP graph, I identified its connected components and regard them as conserved evolutionary units (referred to as "modules"). I define the mosaicity of a protein as the number of its modules minus 1. Genes with no blast result (ORFans) are regarded as mosaicity 0.

6.4.1.1 Distribution of mosaicities of *E.coli* protein-coding genes

I calculated the distribution of mosaicity of all the protein-coding genes of *E.coli* (the number of proteins with a specified number of connected components in their HSP graph). As is shown in Fig. 6.12, around one-half of *E.coli* protein-coding genes

contain only one module, which means that they evolve as singletons. The rest of the *E.coli* protein-coding genes consist of multiple modules. In other words, subgene level introgressive descent happened in around 2000 genes.

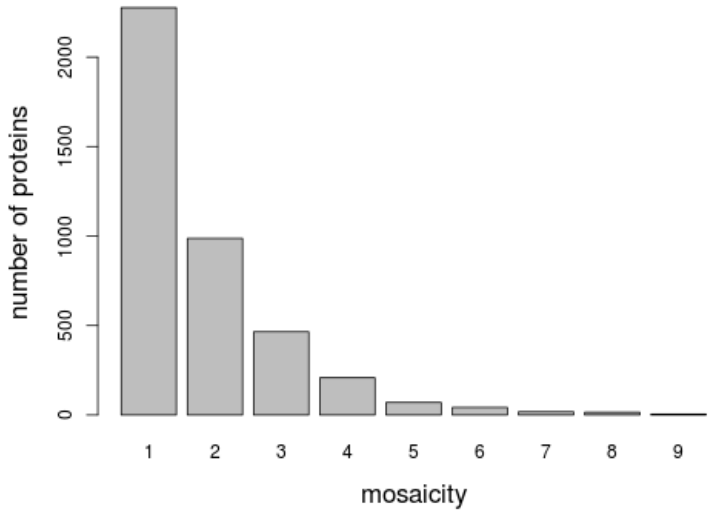

Figure 6.12: **Distribution of mosaicities of *E.coli* protein-coding genes.**

6.4.1.2 Formate metabolism and hyc operon

As a case study, I looked at the enzymes involved in the pyruvate metabolism with a focus on the gain of genes that encode the transformation of formate—an alternative electron acceptor under anaerobic or microaerobic conditions. Two enzymes are involved in the metabolism of formate–Pyruvate Formate Lyase and Formate Hydrogen Lyase. Pyruvate formate lyase is responsible for the cleavage of pyruvate into formate and acetyl-CoA [348]. Reconstruction of ancestral states of the genes that are responsible for the two enzyme complexes revealed that they were gained at two different branches in the evolutionary history of *E.coli* (see Fig. 6.13).

PFL complexes are homodimer consisting of a *PFL activating enzyme* and a *formate acetyltransferase* [349]. In *E.coli*, there are two combinations of the two PFL

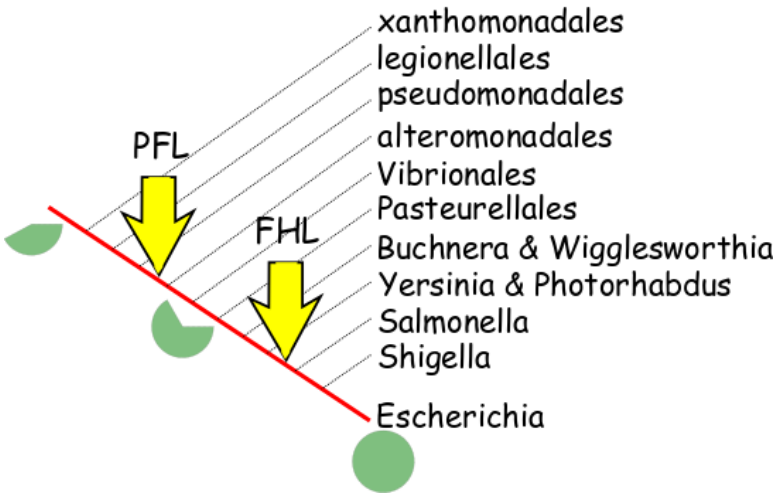

Figure 6.13: **The gain of Formate Hydrogen Lyase and Pyruvate Formate Lyase.**

subunits: 1) pflA with pflB, and 2) pflC with pflD. In addition, the gene tdcE from the tdc operon (which transcribes enzymes in the threonine-degradative pathway) can serve as a functional alternative to PFL in mutants that lose the function of pfl genes [350]. From my reconstruction, the formate acetyltransferase parts (i.e. pflB and pflD, and tdcE) were all gained right before the split of *E.coli* with alteromonadales. The three genes are not close in terms of chromosomal location and are not likely to have been organized in the same operon structure (in fact tdcE is in a completely different operon tdcABCDEFG).

Genes in the hyc operon which encodes the hydrogenase subunit of formate-hydrogen lyase (FHL) complex (which also has a dehydrogenase subunit fdhF [351]), an enzyme complex responsible for the further decomposition of formate into dihydrogen and carbon dioxide under the anaerobic condition in *E.coli* [352]. The existence of FHL helps the cytoplasm from acidification due to formate [353]. Accumulation of formate retards the growth of *E.coli* [354, 355]. FHL plays a central role in the formate regulon [356]. Hyc operon consists of 9 component genes hycABCDEFGHI [357, 358, 359, 360], five of which (hycBCDEG) are included in the orthology data

from STRING v9 database [314]. These five genes were all gained at the branch after the split of *E.coli* with Yersinia and Photorhabdus but before its split with Salmonella.

The HSPs of hycB protein have the following distribution of the aligned regions in hycB (Fig. 6.14).

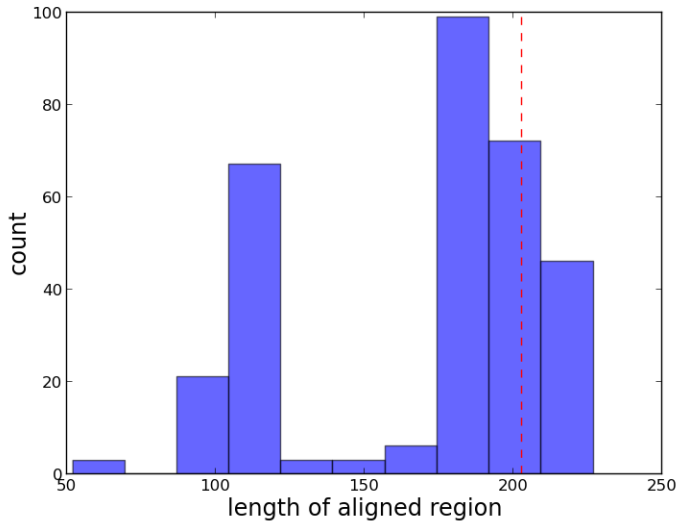

Figure 6.14: **The distribution of length of the aligned region of hycB.** The length of hycB is 203 amino acids. Alignment can be longer than 203 due to insertion.

6.4.1.3 How mosaic are single-copy genes?

The conventional method for reconstructing species phylogeny is based on the concatenation of single-copy genes [248]. Despite the assumption that single-copy genes are resistant to gene level mutational events such as duplication, loss, and horizontal gene transfer, it remains unknown whether these single-copy genes indeed have a tree-like evolutionary history. As shown in Table 6.1, I studied the mosaicity of the single-copy genes. I found out of 44 single-copy genes in the proteobacteria dataset, 31 genes are non-mosaic. Nine genes have mosaicity 2, and three genes have mosaicity 3 and only one gene has mosaicity 6 (see Table 6.1). This means that although the

majority of single-copy genes are non-mosaic. The assumption that single-copy genes must follow tree-like evolution is not true and requires more careful examination.

mosaicity	cnt
1	31
2	9
3	3
6	1

Table 6.1: **Mosaicity of single-copy genes.**

Under a more stringent definition of modules where only connected components of the HSP graph with a size larger than 5 are considered, similar conclusion remains (see Table 6.2).

number of major modules	cnt
1	37
2	5
3	2

Table 6.2: **Number of major modules for single-copy genes.** Major models are defined as connected components in the HSP graph of size 5 or larger.

6.4.1.4 Functional preference of mosaicity

Using the COG functional categories, I studied the mosaicity of genes that belonging to different functional categories (not necessarily metabolism) (see Fig. 6.15). No apparent preference of mosaicity over three main categories is seen. The mean mosaicity of genes differ only slightly among the functional categories (by at most 1 in the medium).

6.4.1.5 How mosaic are metabolic genes?

I compared the mosaicity between metabolic genes and nonmetabolic genes in terms of both the number of connected components in the HSP graph (see left panel of

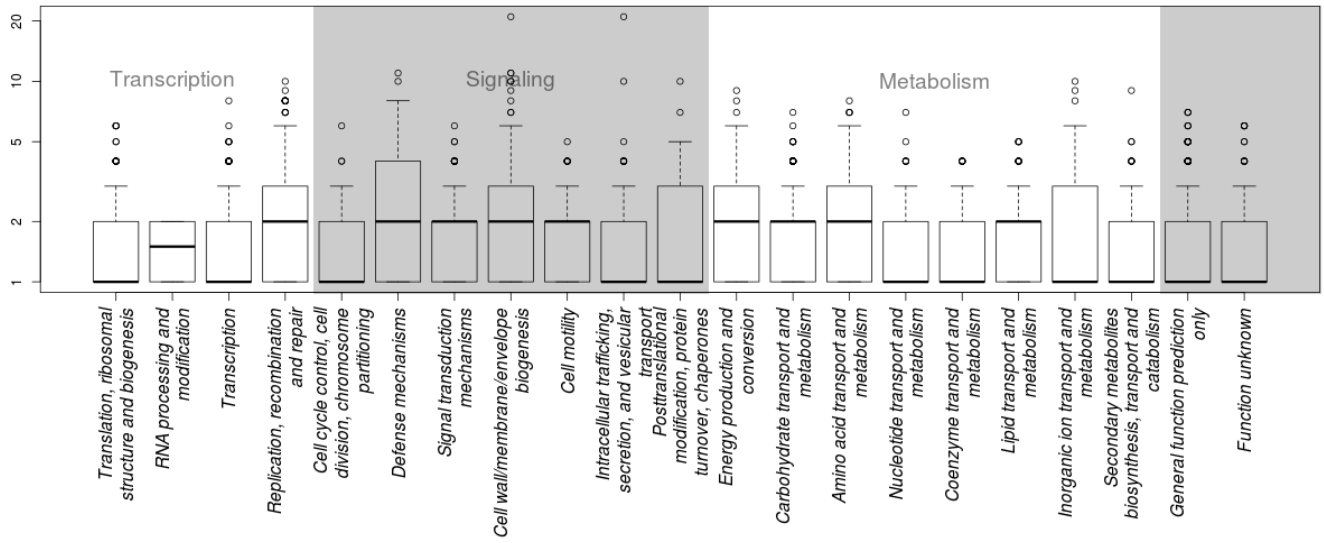

Figure 6.15: **Gene mosaicity of different functional categories.** Functional classifications are retrieved from COG database.

Fig. 6.16) and the major connected components (with more than 5 nodes, see right panel of Fig. 6.16). Metabolic proteins are not very mosaic in general (median has 2 modules and 1 major modules). However, metabolic proteins contain significantly more modules than nonmetabolic genes (metabolic genes are more mosaic/patchy) (Wilcoxon rank sum, $p = 7.602e - 13$ for the number of modules and $p < 2.2e - 16$ for the major modules).

6.4.1.6 Mosaicity vs. metabolic subsystems

I investigated 38 subsystems retrieved from the iAF1260 model. Sorted by the average mosaicity of all the constituent enzymes, the subsystems with the most complex enzymes are:

1. Transport porin through outer membrane
2. other transporter through outer membrane
3. glutamate metabolism
4. tRNA charging and murein biosynthesis

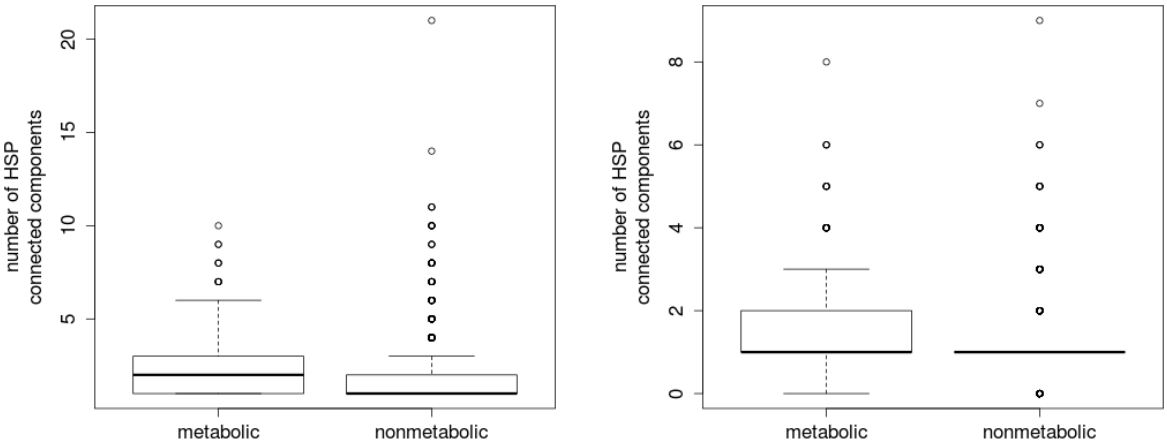

Figure 6.16: **Mosaicity: metabolic vs. nonmetabolic proteins.** Left) Mosaicity counted in the number of any modules. Right) Mosaicity counted in the number of major modules (connected components of HSP graph of size larger than 5).

5. membrane lipid metabolism.

Outer membrane transporters are the most mosaic components. Distinct from the outer membrane transporter, the inner membrane transporter is not very complex (mean=1.944). The subsystems that involve most the non-mosaic enzymes are:

1. methylglyoxal metabolism
2. lipopolysaccharide biosynthesis / recycling
3. unassigned
4. tyrosine metabolism
5. tryptophan metabolism
6. phenylalanine metabolism
7. cysteine metabolism.

6.4.1.7 Which pathway is more mosaic?

I investigated the 361 ecocyc pathways. There are 58 pathways with mean mosaicity 0. They are resistant to being mosaic. The most mosaic pathways (in terms of the mean mosaicity of the constituent genes) are:

1. homoserine biosynthesis (5.250000)
2. glutamate dependent acid resistance (5.000000)
3. glycine betaine biosynthesis I (Gram-negative bacteria) (5.000000)
4. acetyl-CoA biosynthesis I (pyruvate dehydrogenase complex) (4.666667)
5. glutamate biosynthesis I (4.500000)
6. UDP-N-acetyl-D-glucosamine biosynthesis I (4.333333)

6.4.2 Module detection (target database: 635 prokaryotes)

It should be noted that the number of Blast results depend on the taxa sampled in my target Blast database. In order to see whether my results on the module identification are sensitive to the choice of the target database, I validated my results by using a larger Blast database comprising of 635 prokaryotic organisms. A merged microbial dataset was formed by merging 327 proteobacteria, 40 bacteroidetes, 76 archaea, 76 actinobacteria, 40 cyanobacteria and 76 firmicutes. Each *E.coli* protein was blasted against this dataset using blastp [347].

6.4.2.1 Distribution of mosaicities of *E.coli* genes

Fig. 6.17 shows the comparison between the gene mosaicity estimated from blasting against the proteobacteria dataset (small) and from blasting against the merged prokaryotic dataset (large). There is a significant but weak correlation between the two results (Spearman's rank test, $\rho = 0.36, p < 2.2e-16$). From the right panel of Fig. 6.17, it becomes clear that having more taxa could result in higher or lower number of modules. The larger blast database means more HSPs for a given *E.coli* protein. Additional nodes in the HSP graph could lead to the merger of two connected components, hence reducing the number of modules and mosaicity. Additional nodes could also create additional connected components to the HSP graph if the newly

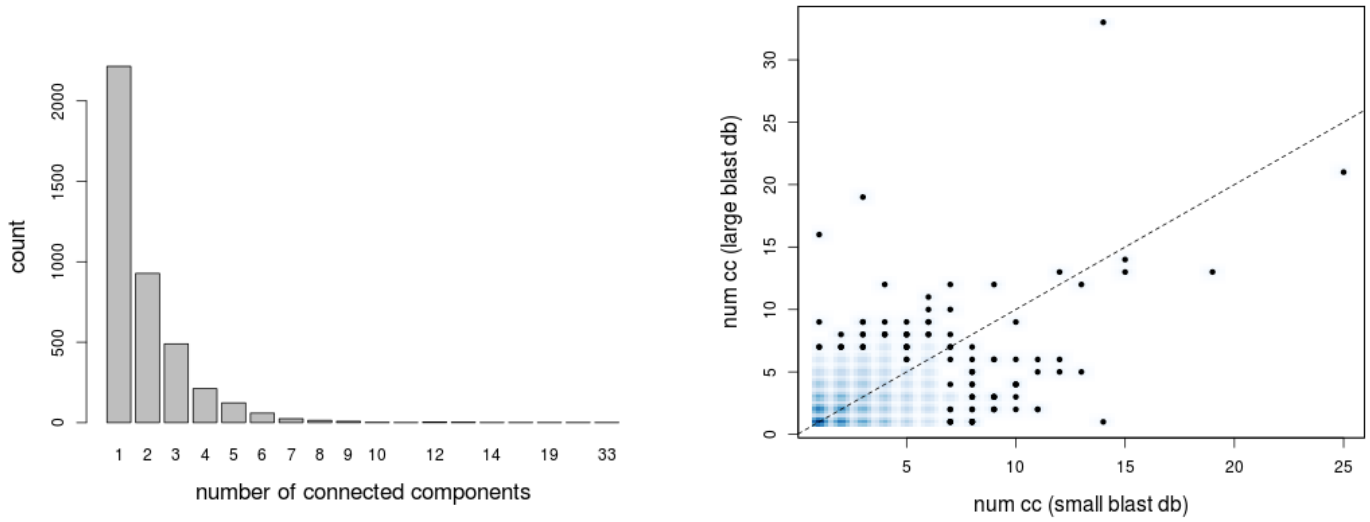

Figure 6.17: **Distribution of mosaicity of *E.coli* genes.** Left) The distribution of mosaicity of *E.coli* genes estimated from blasting the merged prokaryotic database (large). Right) the comparison of mosaicity calculated using the large blast dataset and the small blast dataset. The x-axis (y-axis) corresponds to the number of connected components of the HSP graph built based on small (large) Blast database (see text).

added nodes are not connected to the existing connected components. Both cases are seen in the right panel of Fig. 6.17.

6.4.2.2 Mosaicity vs. age of gene gain

I have compared the mosaicity of a gene with the age of the horizontal transfer as measured by the depth of the branch on which the gene was gained. As shown in Fig. 6.18, no overall association between the age of transfer and mosaicity is seen. On certain branches (e.g., 9, 14 and 15) the gained genes are more mosaic.

6.4.2.3 Mosaicity vs. protein length

Fig. 6.19 compares the length of a protein with its mosaicity. From this figure, I observe that longer genes seem to be more mosaic. This makes sense if the mosaic gene arises from fusion of non-mosaic genes and the rate that fusion and fission occur

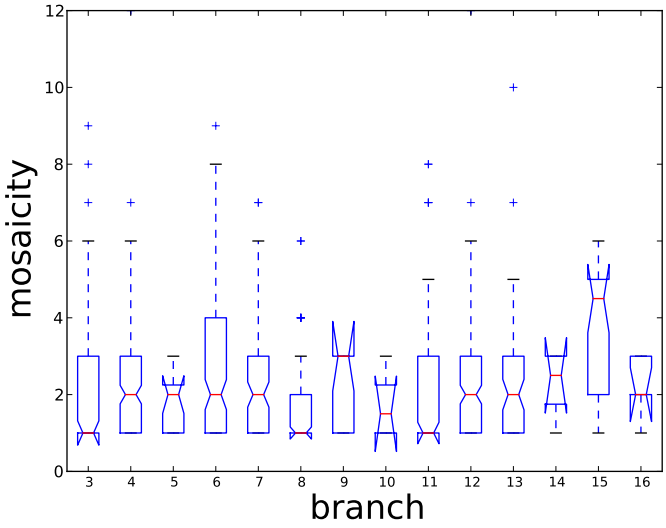

Figure 6.18: **Gene mosaicity vs. the time the gene was gained.** From left to right, each branch corresponds to the branches along the evolutionary history of *E.coli*

per site is relative constant across genes.

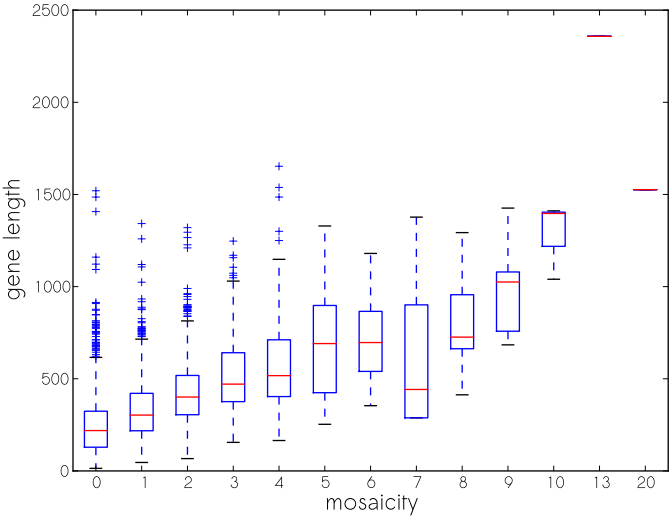

Figure 6.19: **Protein length vs. mosaicity.**

6.4.2.4 Ontology of mosaic genes

I studied the ontology of mosaic genes by looking at GO enrichment. Mosaic genes are defined as genes with more than 4 modules. I found ontologies enriched in mosaic genes are: DNA binding, ion binding and transposon related. Ontology depleted from mosaic genes are: transcriptional regulation of biosynthetic processes. This may be due to the problem of excluding the non-protein coding genes, which are RNA-related and hence involved in transcriptional regulation.

6.4.3 Pathway mosaicity

Fig. 6.20 shows how mosaic metabolic pathways are in terms of the mosaicity of the constituent enzymes. 361 metabolic pathways are retreived from the EcoCyc database. The larger the pathway, the more mosaic enzymes it has. 58 out of 361 pathways are composed of only non-mosaic enzymes. Most pathways have low level of mosaicity. For these pathways, gene fusion and fission do not contribute much to their evolution. The most mosaic pathways are related to biosynthesis of homoserine, glycine betaine, acetyl-CoA, glutamate, and UDP-acetyl-glucosamine. In the evolution of these pathways, fusion and fission might have played adaptive roles.

6.4.4 Mosaicity vs. fitness

Comparison between mosaicity and the fitness contribution of metabolic enzymes reveals that enzymes which contribute to fitness in a small number of environments are more likely to be mosaic. The fitness contribution of enzymes are computed using flux-balance analysis optimizing the flux of biomass accumulation. It is measured by the number of growth conditions under which the knock-out of the enzymes will either leave the model infeasible, or significantly reduce the biomass accumulation. This result is consistent with our previous observation that transporter genes are

GO term	type	description	r_s	r_p	p
0006259	e	DNA metabolic process	42	237	5.98e-08
0006310	e	DNA recombination	23	98	1.65e-06
0004803	e	transposase activity	12	35	0.000197
0004386	e	helicase activity	11	30	0.000302
0003824	e	catalytic activity	147	1992	0.000726
0006313	e	transposition, DNA-mediated	12	42	0.00186
0032196	e	transposition	13	50	0.00219
0006260	e	DNA replication	15	69	0.00402
0043167	e	ion binding	98	1205	0.00676
0008094	e	DNA-dependent ATPase activity	8	20	0.00679
0004553	e	hydrolase activity, hydrolyzing O-glycosyl compounds	10	34	0.0109
0080090	p	regulation of primary metabolic process	7	453	0.0218
0031323	p	regulation of cellular metabolic process	7	452	0.0218
0019219	p	regulation of nucleobase-containing compound metabolic process	6	416	0.029
0051171	p	regulation of nitrogen compound metabolic process	6	417	0.029
0015473	e	fimbrial usher porin activity	5	8	0.0309
0009889	p	regulation of biosynthetic process	7	436	0.0445
0031326	p	regulation of cellular biosynthetic process	7	436	0.0445
0010556	p	regulation of macromolecule biosynthetic process	7	435	0.0447
0043169	e	cation binding	60	658	0.0483

Table 6.3: **Enriched and depleted GO terms in mosaic genes.** Mosaic genes are genes with more than 4 modules. p: depleted; e: enriched; p-value is after bonferroni correction. The study sample size is 243 and the population size is 4450. r_s: ratio in study; r_p: ratio in population.

more mosaic because most transporters are not essential under most of the growth conditions and are typically needed only in a certain type of environment. The fact that the adaptation of *E.coli* to these growth conditions involve the recruitment of mosaic genes raises the possibility that these recruitment might have taken the form

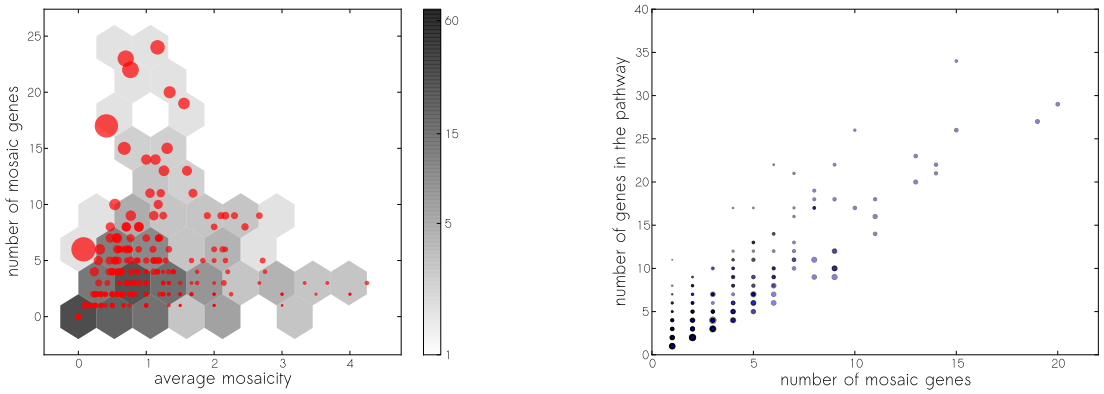

Figure 6.20: **Mosaicity of enzymes in metabolic pathways.** Left) the average mosaicity vs. number of mosaic enzymes in the pathway. Each dot corresponds to a pathway. The size of the dot is proportional to the size of the pathway (in terms of the number of enzymes). The darkness of the hexbin corresponds to the number of observations in the bin. Right) The number of the genes in the pathway vs. the number of mosaic ones.

of sub-gene level mutational events such as fusion, fission and partial horizontal gene transfer.

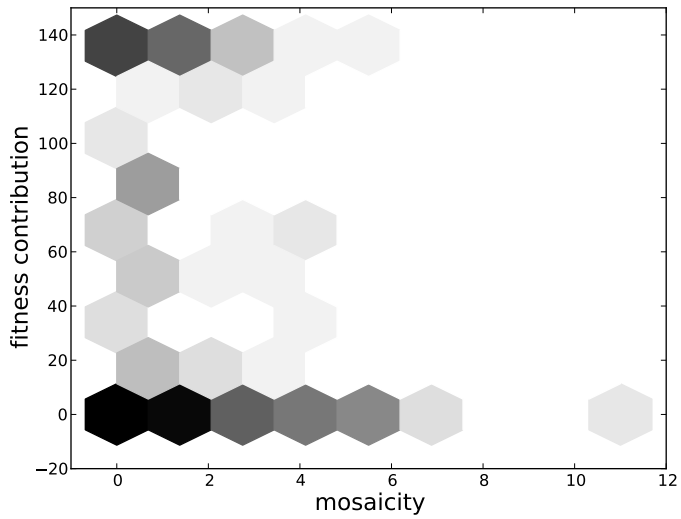

Figure 6.21: **Mosaicity of metabolic enzymes vs. their contribution to the organismal fitness.** The x-axis corresponds to the mosaicity of the enyzme and the y-axis corresponds to its contribution to the fitness (in terms of the number of growth conditions under which the protein contributes to the biomass accumulation). The darkness of each hexbin is porportional to the number of observations.

6.4.5 Conclusion

Mosaicity is characterized by the number of modules a protein has and, hence, how often fission and fusion have taken place. We found that the longer the protein, the more mosaic it is. About one-half of the proteins in *E.coli* are mosaic. Metabolic proteins tend to be more mosaic than non-metabolic proteins. Proteins that are most mosaic have the property of binding to DNA and ion, while biosynthetic enzymes are not mosaic. Most pathways have low level of average mosaicity. But enzymes that contribute to fitness in a small number of environments are more likely to be mosaic, indicating that sub-gene level mutational events contributed to the adaptation of microbial organisms to different environments.

6.5 Module family analysis

6.5.1 Module family reconstruction (56 proteobacteria)

205,177 annotated protein sequences from 56 taxa were downloaded from NCBI genbank. I conducted all-against-all blastp of these sequences. The resulting high-scoring segment pairs (HSPs) were collected and filtered by an E-value threshold 1e-4. There are 19,377,549 HSPs. An HSP graph was assembled in a way similar to the detection of modules in *E.coli*, except that different HSPs can be connected based on genes in any of the species in the proteobacteria dataset. 17,562,439,322 connections are made in the HSP graph. 19,806 connected components of the HSP graph were then identified. The three largest connected components had more than 1 million nodes (having 1,013,222, 2,307,875 and 3,307,785 nodes respectively). Around 7,563 connected components had only 2 nodes (smallest size).

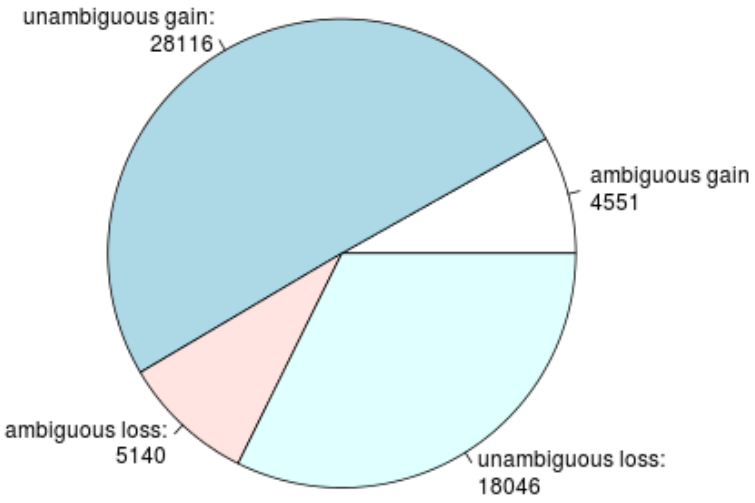

Figure 6.22: **Gain and loss of modules inferred.**

6.5.1.1 Gain and loss of module families

I treated δ-proteobacteria as an outgroup. They are: 1) Desulfovibrio vulgaris str. Hildenborough (taxid: 882) 2) Geobacter sulfurreducens PCA (taxid: 243231); 3) Bdellovibrio bacteriovorus HD100 (taxid: 264462). I encoded the states of the nodes on the species tree in binary variables with 1 indicating presence and 0 absence. I inferred gains and losses from state transitions ($0 \rightarrow 1$ indicates gain and $1 \rightarrow 0$ indicates loss). Most of the events identified were unambiguous (86.1% for gain and 77.8% for loss). For unambiguous events, gains are more common than losses (see Fig. 6.22).

Fig. 6.23 shows the distribution of the number of gains and losses of modules by branch and by module family. Most branches have had fewer than 2 gains and losses. Some branches could harbor up to one thousand gains and losses. Most module families has 1 gain and no losses. There were more module families where gains occurred than where no gain was observed.

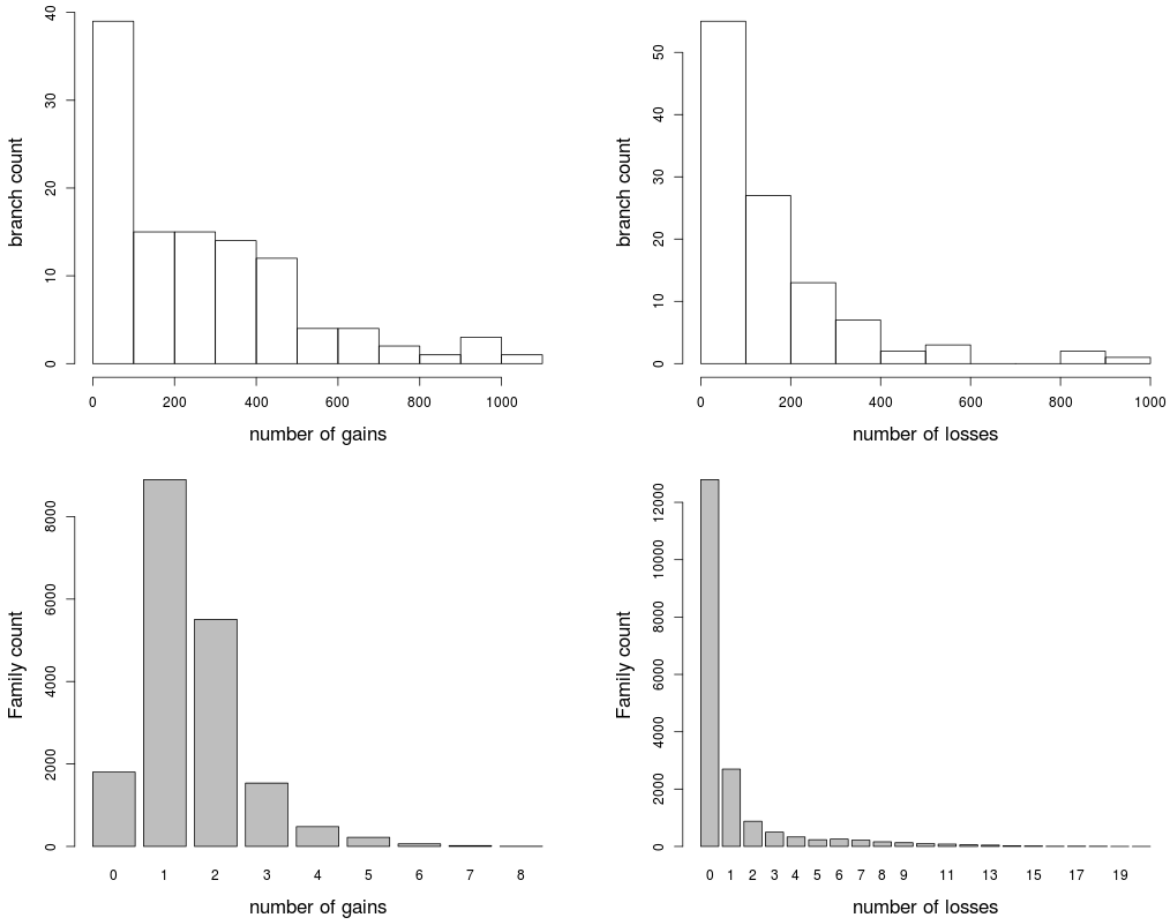

Figure 6.23: **Distribution of gains and losses of modules.** Top left) distribution of module gains by branch; Top right) distribution of module losses by branch; Bottom left) distribution of module gains by module family; Bottom right) distribution of module losses by module family. Only unambiguous gains and losses are plotted.

6.5.1.2 A comparison of module inference with whole-protein inference

In Fig. 6.24, I compared on each branch the number of gains and losses of proteins as inferred from whole protein analysis and of modules as is inferred from the module analysis. More module gains were observed than gene gains because modules are co-gained if the whole gene is gained in one shot. However, the number of modules lost is comparable to the number of genes lost. This indicates that genes frequently undergo partial loss when some of their constituent modules are lost.

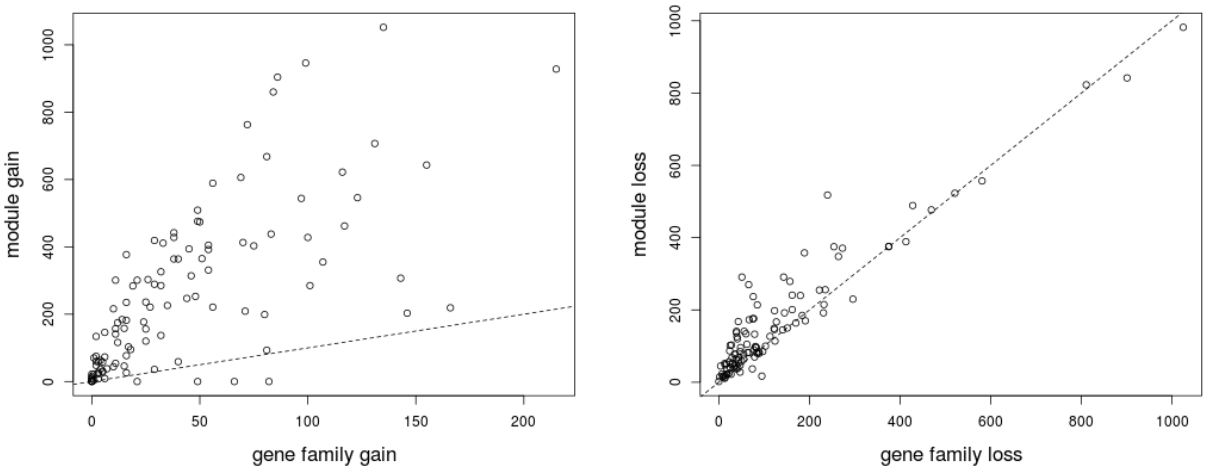

Figure 6.24: **Module vs. whole protein: number of gains and losses on each branch.** The branches are from the species tree inferred from the proteobacteria databaset. X-axis corrisponds to the gain (loss) of gene families and y-axis indicates the gain (loss) of module families.

6.5.2 Module family reconstruction (62 proteobacteria)

215,618 annotated protein sequences from 62 taxa were downloaded from the NCBI genbank. Each protein sequence is used as a query to blast against all the other sequences from the dataset. The blast returns 20,625,378 High-scoring Segment Pairs (HSPs) with E-values greater than 1e-4 and 11,686,942 HSPs with E-value greater than 1e-20. Fig. 6.25 shows the relationship between E-value and alignment length. When E-values are lower, the minimum length of the alignment is longer. However, when the E-value is not very small (larger than 1e-10), this trend is not very obvious. Most alignments are longer than 50 amino acids.

6.5.2.1 HSP graphs and their connected components

An HSP graph is assembled where nodes of the graph are HSPs and two nodes are connected if two alignments involve the same protein and the aligned regions of that protein overlap by 70%. The resulting HSP graph has 18,524,311,298 (18 billion) edges. I identified 20,980 connected components of this graph. To identify

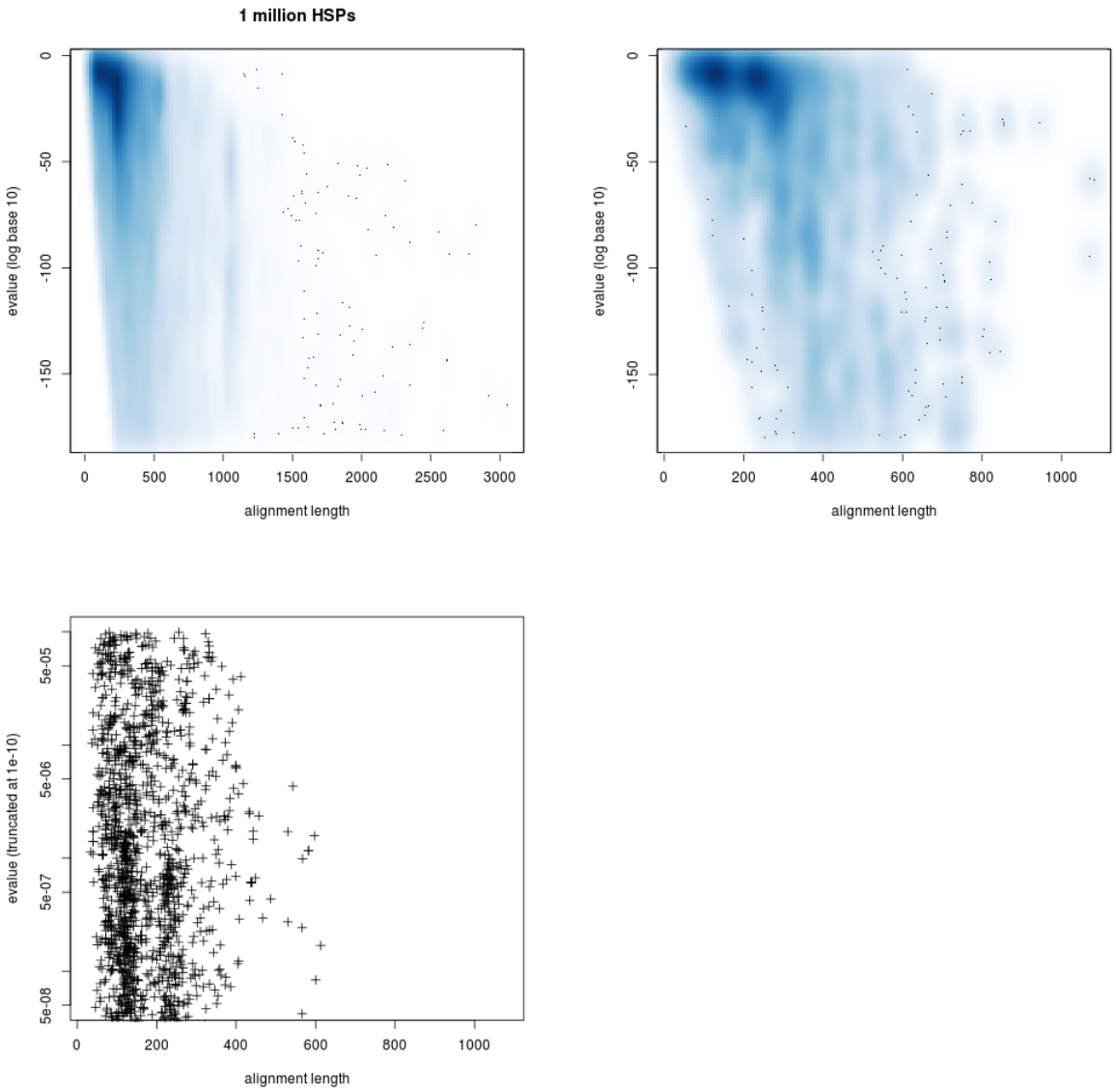

Figure 6.25: **Length of the alignment vs. E-value** Top left) the first 1,000,000 HSPs are plotted (sorting is arbitrary); Top right) the first 10,000 HSPs are plotted; Bottom left) the first 10,000 HSPs but only with E-value higher than 1e-8 are plotted.

the connected components, I passed the edge list twice—the first time to assign a temporary label and establish the label equivalency, the second to reassign the nodes to the label equivalence class. There are three connected components with their size larger than 1 million (having 1,014,605, 2,389,959 and 3,480,428 nodes respectively). The largest connected component has 3,480,428 (3 million) nodes and 2,049,173,689

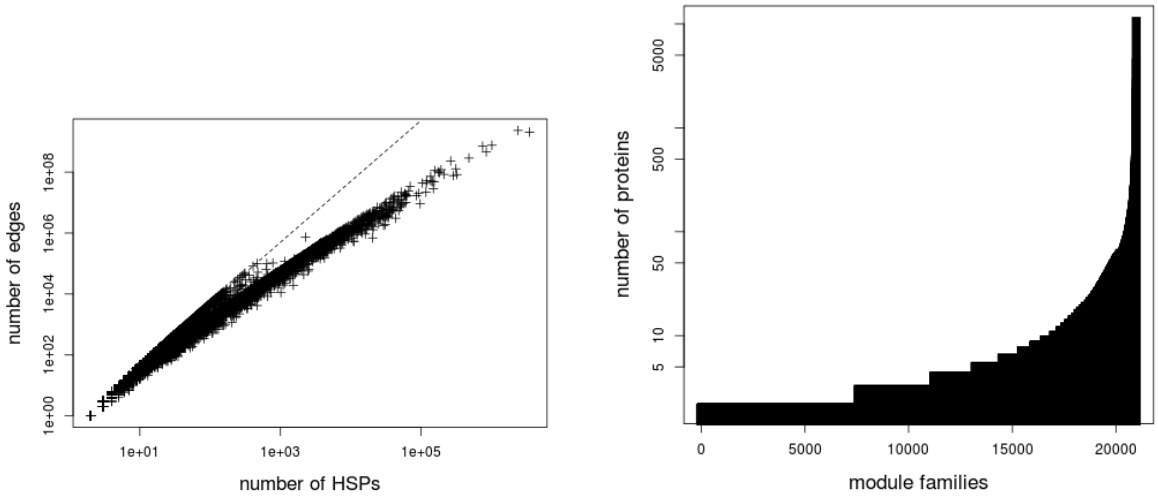

Figure 6.26: **From connected components of HSP graph to module families.** Left) number of nodes vs. number of edges of all the connected components of the HSP graphs. The dashed line corresponds to the line $y = x(x-1)/2$, or the largest number of edges can exist given the number of nodes. Right) the number of proteins that members in each module family belong to.

(2 billion) edges. There are 8,168 connected components that have only 2 HSPs (smallest size). 240 connected components have more than 1 million edges. Shown in the left panel of Fig. 6.26 is the number of nodes and edges for each connected component of the HSP graph.

I regard each connected component as a module family. The aligned region on the protein is regarded as a module identified for that protein. I mapped the connected components to the proteins that bear these alignments. Shown in the right panel of

rank	id	number of proteins	annotation
1	16	10574	chemotaxis, two-component system
2	191	4215	ABC transporter, ATPase
3	434	2648	ABC transporter
4	96	2220	acyl-ACP reductase
5	142	2211	lysR family, a transcriptional regulator

Table 6.4: **The five largest module families.** The annotations are extracted manually from the annotations of the member proteins.

Fig. 6.26, there is a large number of module families with a small number of proteins and small number of module families that appear in a large number of proteins. The five module families that involve the most proteins are shown in Table 6.4.

There is a weak but significant positive correlation between the size of the module family and the lengths of the modules in the family (Spearman's rank $\rho = 0.16$ and $p < 2.2e - 16$). The largest module families are of around 200 amino acids in length.

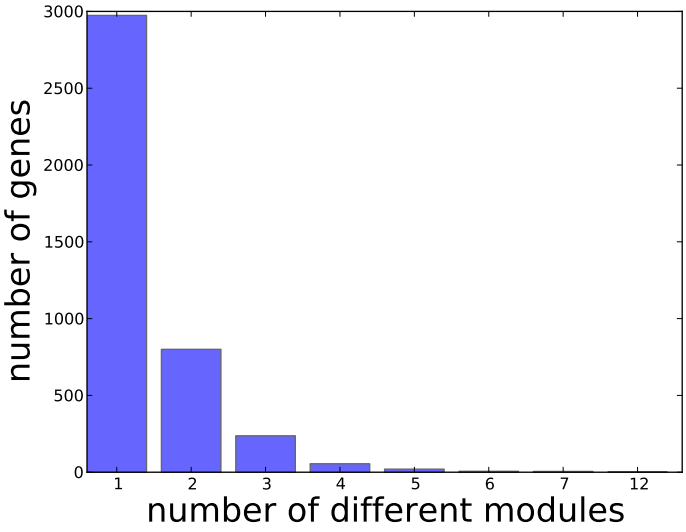

Figure 6.27: **The distribution of the number of modules from different families of *E.coli* genes.**

In Fig. 6.27, I plotted the number of modules that come from different module families for each *E.coli* gene. Compared with the distribution of mosaicity of *E.coli* genes, I found that the distributions of different module families are significantly more skewed towards the lower-end. Genes contain fewer modules from different families than the modules themselves. This implies that *E.coli* genes could contain modules of the same ancestry, which could arise from duplication or horizontal transfer without replacement.

6.5.2.2 Density of connected components of HSP graphs

Ideally, module homology results in a clique in the HSP graph. Therefore, I studied the density of the identified HSP graph. Shown in Fig. 6.28, most connected components are dense, but a significant portion are sparse. 4162 out of 20890 (19.8%) connected components have density lower than 0.5. Low density connected components are larger (right panel of Fig. 6.28). Deviation of the clique structure, or the loss of connections in the connected components of the HSP graph, can be a consequence of two major reasons. One is the transition of error which comes from spurious blasting results. For example, protein A aligns to protein B and B aligns to C and C aligns to D. But the 70% overlap of the segment of B in HSP1 is different from the 70% overlap of the segment of C in HSP2. The alternative explanation is that the sparsity might be due to the intrinsic diversity of the module family. Protein sequences that do not align well might still have evolved from a common ancestor and be homologous, despite that they have lost sequence similarity over many years of evolution.

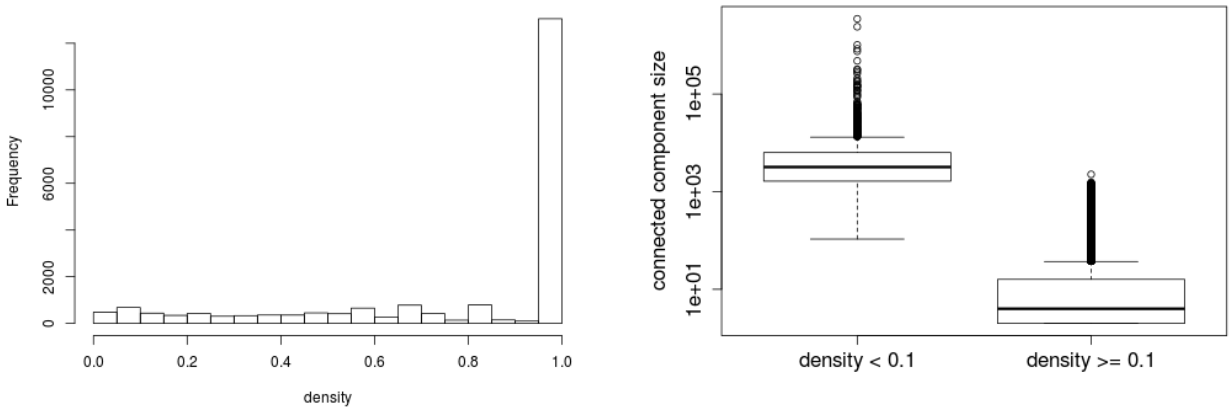

Figure 6.28: **Density of HSP graph connected components.** Left) the distribution of the density of the HSP graph. Right) Comparison in size between dense connected components and sparse connected components.

6.5.2.3 Overlap between module families and COG

I computed the Shannon information entropy of COG assignments for proteins carrying modules from the same module family. Different module families might contain member modules located in the same protein. Fig. 6.29 shows that most module families have member modules located at proteins with the same COG annotation (low entropy). Few module families, on the other hand, have a high entropy (up to around 5). There is a module appearing in over 9000 proteins and these proteins belong to more than 100 COGs. Modules from high entropy module families exist in a larger number of proteins (top right panel of Fig. 6.29). One possible explanation is that more COG annotations are expected by chance (which also leads to higher COG entropy, see bottom left of Fig. 6.29) if the module family is involved in more proteins. Moreover, the more COGs observed, the higher the entropy, which is also expected (bottom right of Fig. 6.29). Among 3785 COG gene families, 1401 COG families contain genes with only 1 module. The rest of the genes are not consistent in the modules (e.g., 983 COG families have both genes with 1 module and 2 modules).

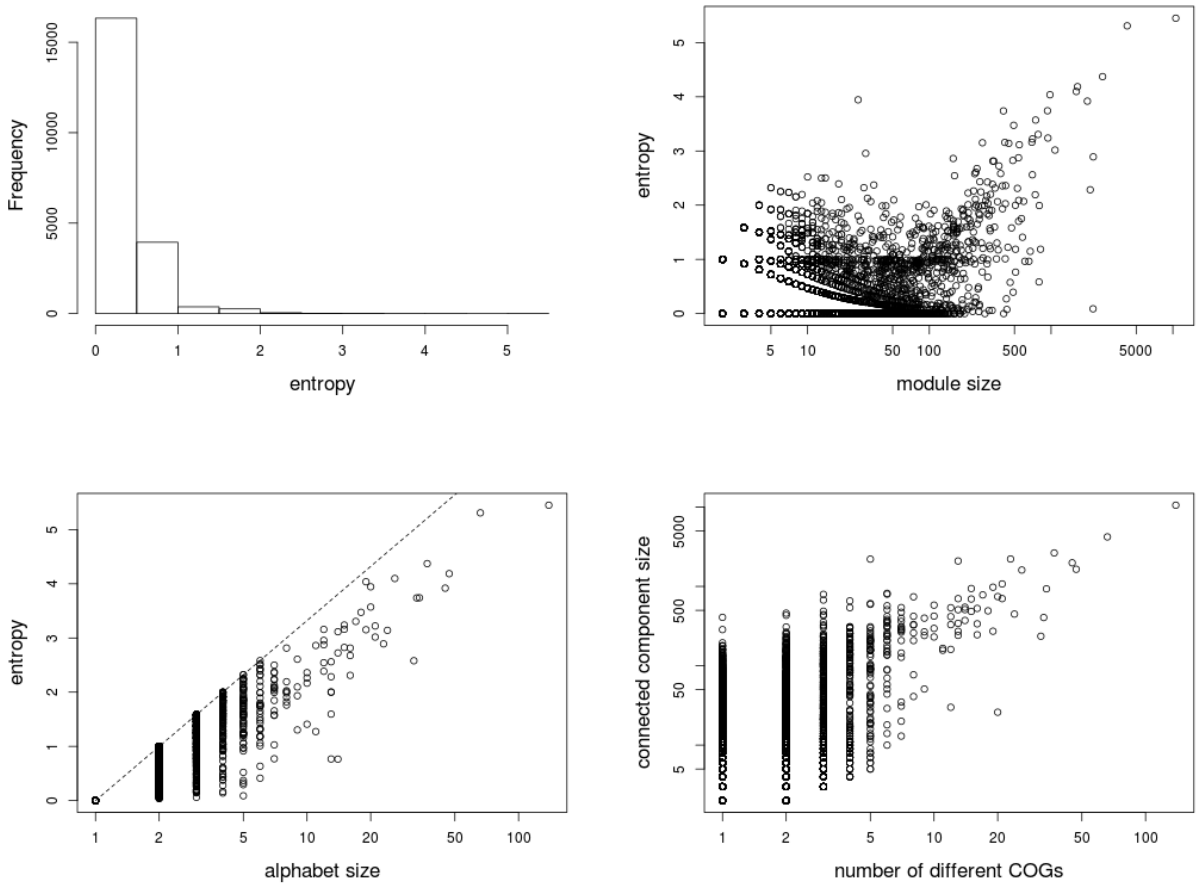

Figure 6.29: **Module family vs. COG.** Top left) Shannon entropy of family assignment in cc. Family assignments are the COG assignments. Top right) family entropy vs. the size of the connected component. Bottom left) family entropy vs. number of families seen in the connected component. The dashed line is the maximum entropy one can get given the size of the alphabet (or the number of COGs seen for that connected component) (log(n), n is the size of the alphabet, base is 2). Some connected components that are equally splitted in annotation. Bottom right) size of the connected components vs. the number of COG annotation.

6.5.2.4 Gain, loss and duplication of module families

I treated the ϵ-proteobacteria as an outgroup. They are: 1) C. jejuni NCTC 11168; 2) C. jejuni RM1221 57899; 3) H. hepaticus ATCC 51449; 4) H. pylori 26695; 5) H. pylori J99; 6) W. succinogenes DSM 1740. For each of the 20980 module families, I coded the state of the taxa and the internal nodes in the copy number of the module

family. The ancestral states were reconstructed using the PAUP software [332] (see Method section for the protocol). Shown in Fig. 6.30, more than half of the module families had only one copy in each taxon. Some module families were extremely promiscuous, having up to 471 copies in one taxon.

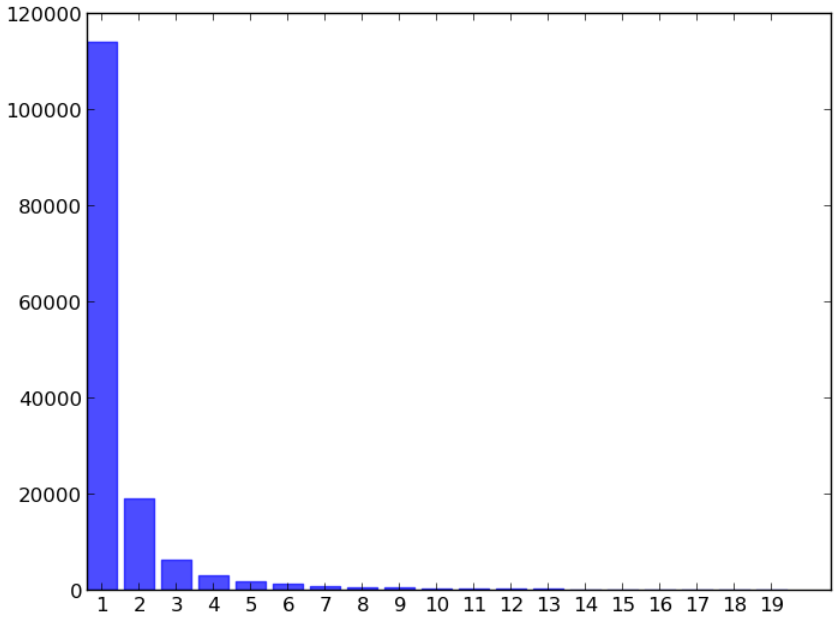

Figure 6.30: **The distribution of copy number of modules in the proteobacteria dataset.** The axis is non-exaustive. The maximum copy number is 471.

As shown in Fig. 6.31, ambiguous changes are relatively rare compared with unambiguous changes. Complete gains and losses are more common than incomplete gains and losses. Duplication events are relatively rare compared with gains and losses.

The distribution of complete gains and losses (Fig. 6.32) resembles the distribution of gains and losses from the binary inference (Fig. 6.23). Most module families do not contain any incomplete gains or losses.

The distribution of complete gain and losses (Fig. 6.33) also resembles the one from binary inference (Fig. 6.23). Most branches have fewer than 100 incomplete

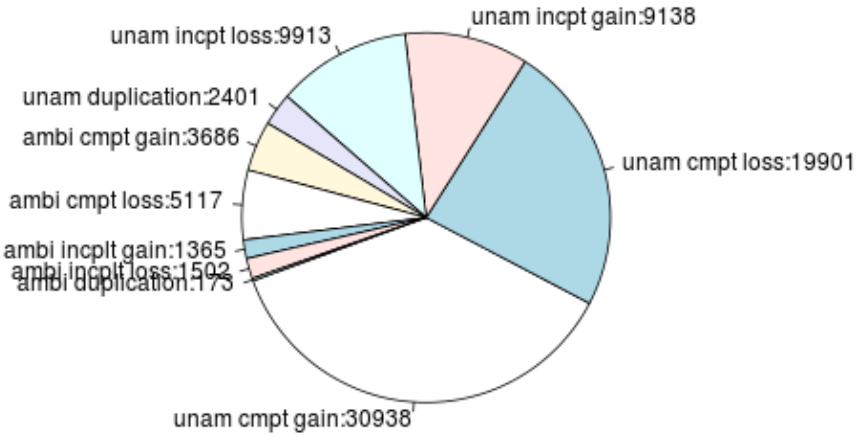

Figure 6.31: **Piechart of gains, losses and duplications.** The gains are classified into complete gains where the source state is zero and incomplete gains where the source state is nonzero. The losses are also classified into complete and incomplete losses depending on whether the target state is zero or not. "unam" and "ambi" are short for "unambiguous" and "ambiguous". "cmpt" and "incplt" are short for "complete" and "incomplete".

gains or losses. Some branches have up to 600 incomplete gains and 400 incomplete losses.

Based on my inference, duplications are extremely rare (see Fig. 6.34). This is consistent with Pal et al.[2] who state that the recent evolution of bacteria is dominated by horizontal gene transfer instead of duplication.

The peaks and troughs of gains and losses of modules parallel well with the gains and losses of whole genes (Fig. 6.35). One explanation is that when a gene consists of multiple modules and the gain and loss of the gene results in the co-gain and co-loss of all the modules the gene have. Therefore, the number of gains and losses are amplified when one compares evolutionary events on the gene scale with ones on the module scale. Most modules from the same protein are gained and lost at the same time.

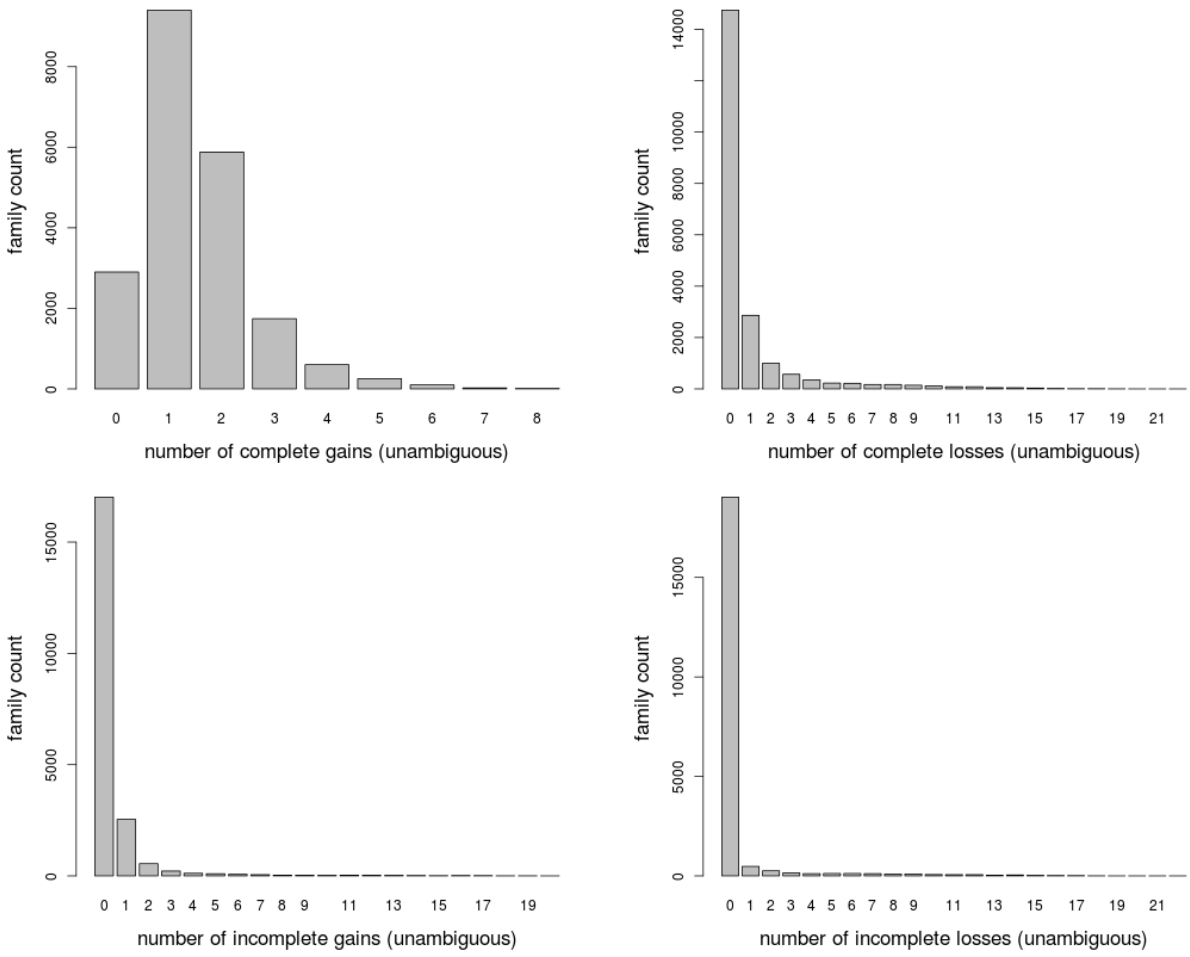

Figure 6.32: **Distribution of gain and loss by module family.** Top left) the distribution of the number of complete gains; Top right) the distribution of the number of complete losses; Bottom left) the distribution of the number of incomplete gains; Bottom right) the distribution of the number of incomplete losses.

6.5.3 Conclusion

Module families are identified based on sequence similarity and HSP graph of blast results from protein sequences of the whole dataset. Most module families belong to genes of similar COG annotation. The inference of gains and losses of module families suggests that most modules from the same protein are gained and lost at the same time and module-level mutations are relatively rare compared to other mutational events such as gain and loss.

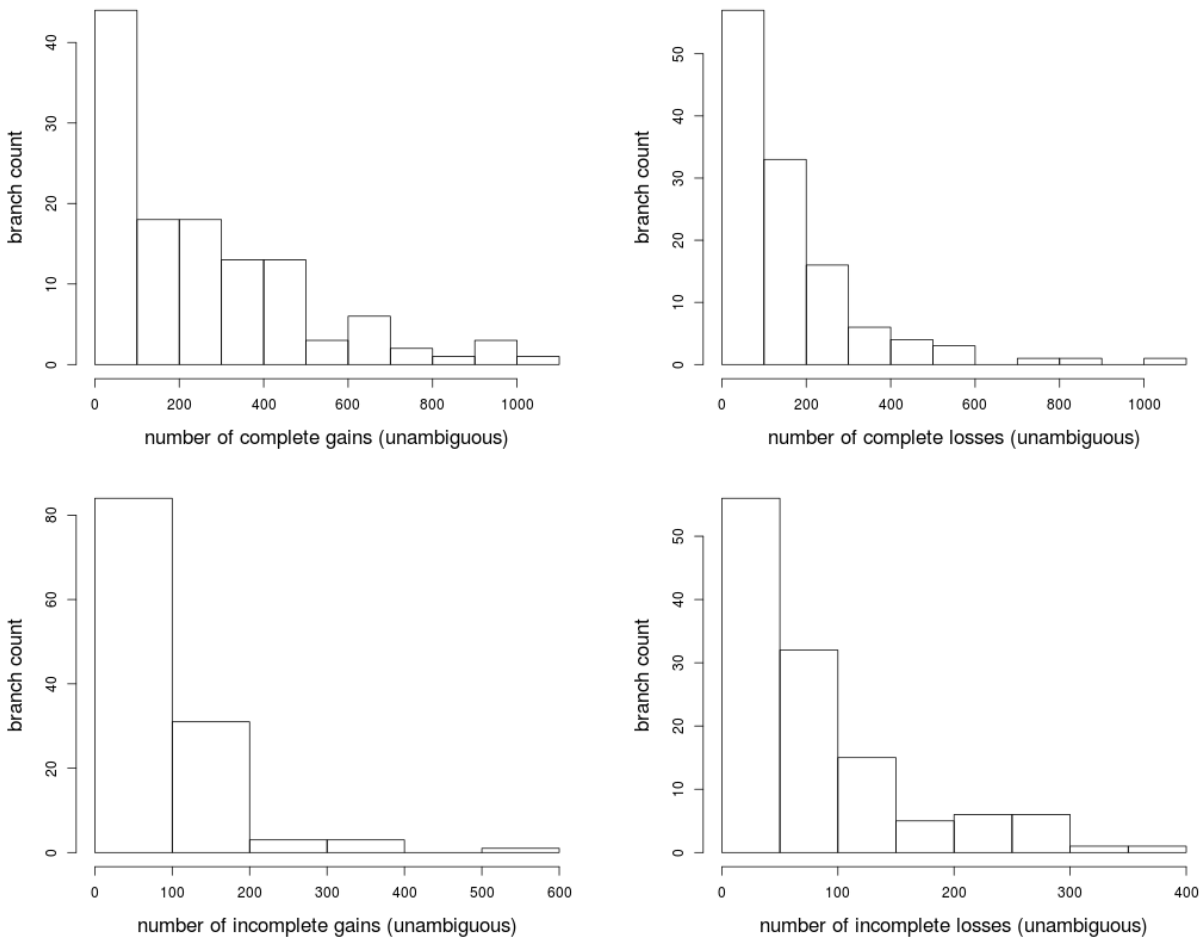

Figure 6.33: **Distribution of gain and loss by branch.** Top left) the distribution of the number of complete gains; Top right) the distribution of the number of complete losses; Bottom left) the distribution of the number of incomplete gains; Bottom right) the distribution of the number of incomplete losses.

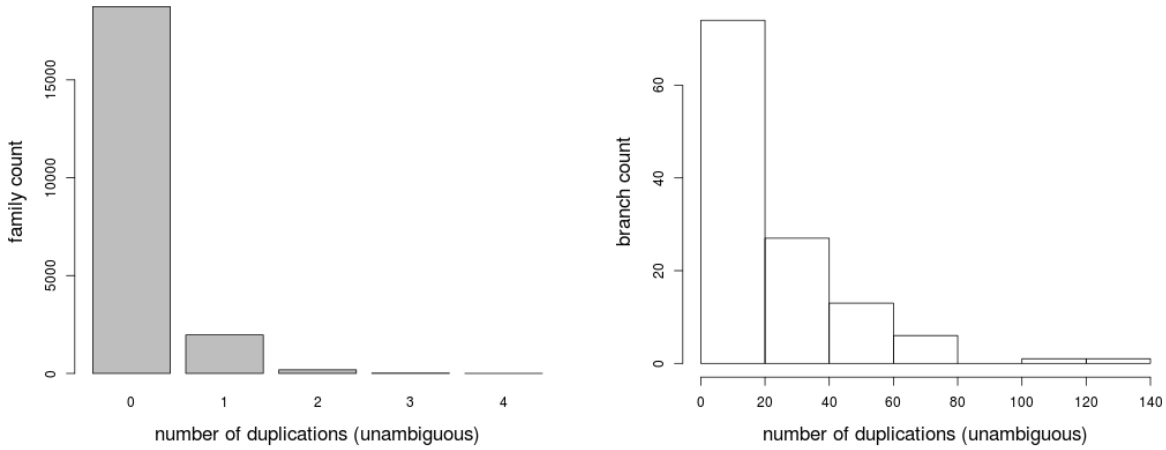

Figure 6.34: **Distribution of the number of duplications.** Left) by module family; Right) by branch.

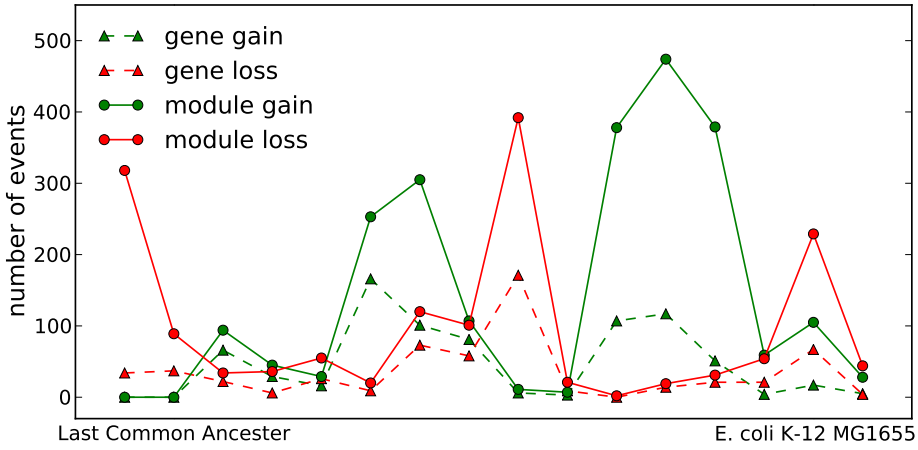

Figure 6.35: **Gain and loss of module families along the *E.coli* evolutionary history** The gains and losses of whole genes is also plotted for comparison. Only unambiguous complete changes are considered.

Chapter 7

Conclusion and future directions

The following conclusions can be made from the work presented in this dissertation. First, accounting for the hypergraph nature of the metabolic network helps construct better null models for testing organising principles, such as the scale-freeness and hierarchical structure of metabolic networks. Second, SCL is a systematic criterion for pruning metabolic network connectivity based on the chemical content conserved among reactants in the metabolic reaction. SCL can guide efficient path-searching algorithms which lead to biochemically relevant pathways. Third, alternative tracings that arise from chemical graph symmetry could potentially alter the results of pathway inference and isotopomer transition patterns. The detection of symmetry can be automated through a graph automorphism formulation. Decomposition of chemical graph symmetry allows more compact storage and selective reassembly of automorphic mappings of chemical graphs. Alternative mappings can be enumerated by composing any valid atom mapping of the reaction with the identified reaction symmetries. Fourth, metabolome size, a signature of the organism's living style, is a crucial determinant for the modular structure identified in the organism's metabolic networks. Metabolic network modularity can be achieved through multiple underlying community structures. The current modularity-based quantification of the community structure of

metabolic networks is not a good reflection of the ontology of biochemical pathways. Finally, introgressive descent of microbial genes challenges the *gene family* concept and, hence, the tree-like view of evolution long held by researchers. Mutational events that occur on the module level including horizontal gene transfer, duplication, loss and gene fusion and fission play pivotal roles in the emergence of metabolic pathways and mosaic genes in microbial organisms. As phylogenomic methods continue to be extended to the sub-gene scale, these mutational events on gene fragments can be identified. Their relative contribution and role in shaping metabolic genes as well as operon architecture can be more thoroughly revealed.

As modeling metabolism continues to hold a central position in biomedical research and in guiding bioengineering practices, more sophisticated modeling techniques are required. The next-generation metabolic modeling techniques are pushing in three major dimensions: 1) models that incorporate finer chemical structures, such as atom tracing, but still keep the scope of the whole metabolome; Finer tracing of atoms in the metabolic systems to help scientists peek into the intricate cellular systems and give more accurate estimates of the metabolic reaction fluxes as well as the activity of the enzymes that control those reactions. 2) models that bridge metabolic components over a wider temporal and spatial scale. For example, the modeling of several different cell types in a multi-cellular organism or the co-habitation of different bacterial species in a metagenomic study. 3) models that interface well with non-metabolic components of the cell, such as signal transduction and transcriptional regulation. Integration and crosstalk are vital in understanding the control of metabolism beyond small-molecule chemical processes. Whichever direction innovation in the metabolic modeling takes, emphasis must be placed on using available experimental data either to constrain the model or to validate the model semantics and parameters.

In addition to modeling endeavors, more powerful algorithms for the phylogenomic reconstruction of mutational as well as non-mutational evolutionary events must be produced. These new algorithms must take into consideration the architecture of both gene organization in modules and operon organization in genes. Methods that apply to a large number of remotely-related species should be developed to shed light on ancient and long distance mutational events. More suitable null model construction for metabolic pathways and network evolution, together with more biochemically relevant definitions for topological features such as the modular structure, must be devised to re-evaluate whether an observation made of the microbial genome evolution is an outcome of adaptation to the environments or neutral forces. Assumptions must be carefully evaluated in simplifying studies where outcomes from mutiple evolutionary events are intertwined.

References

[1] A. M. Feist, J. C. M. Scholten, B. O. Palsson, F. J. Brockman, and T. Ideker, "Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri," *Molecular Systems Biology*, vol. 2, 2006.

[2] C. Pál, B. Papp, and M. J. Lercher, "Adaptive evolution of bacterial metabolic networks by horizontal gene transfer." *Nature genetics*, vol. 37, no. 12, pp. 1372–5, Dec. 2005.

[3] D. Segrè, A. Deluna, G. M. Church, and R. Kishony, "Modular epistasis in yeast metabolism." *Nature genetics*, vol. 37, no. 1, pp. 77–83, Jan. 2005.

[4] H.-G. Holzhütter, "The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks." *European journal of biochemistry / FEBS*, vol. 271, no. 14, pp. 2905–22, Jul. 2004.

[5] B. O. Palsson, "The challenges of in silico biology," *Nat Biotech*, vol. 18, no. 11, pp. 1147–1150, Nov. 2000.

[6] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar, "Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae," *Bioinformatics*, vol. 18, no. 2, pp. 351–361, Feb. 2002.

[7] S. Klamt and J. Stelling, "Combinatorial Complexity of Pathway Analysis in Metabolic Networks," *Molecular Biology Reports*, vol. 29, no. 1, pp. 233–236, Mar. 2002.

[8] S. J. Wiback, I. Famili, H. J. Greenberg, and B. O. Palsson, "Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space," *Journal of Theoretical Biology*, vol. 228, no. 4, pp. 437–447, Jun. 2004.

[9] V. Acuña, F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M.-F. Sagot, and L. Stougie, "Modes and cuts in metabolic networks: complexity and algorithms." *Bio Systems*, vol. 95, no. 1, pp. 51–60, Jan. 2009.

[10] 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." *Journal of theoretical biology*, vol. 203, no. 3, pp. 229–48, Apr. 2000.

[11] P. Mendes, "GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems." *Computer applications in the biosciences : CABIOS*, vol. 9, no. 5, pp. 563–71, Oct. 1993.

[12] M. Tomita, K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and A. C. Hutchinson, "E-CELL: software environment for whole-cell simulation," *Bioinformatics*, vol. 15, pp. 72–84, 1999.

[13] E. L. King, K. Tuncay, P. Ortoleva, and C. Meile, "In Silico Geobacter sulfurreducens Metabolism and Its Representation in Reactive Transport Models," *Appl. Environ. Microbiol.*, vol. 75, no. 1, pp. 83–92, 2009.

[14] Z. Szallasi, *System modeling in cell biology : from concepts to nuts and bolts*. Cambridge Mass.: MIT Press, 2006.

[15] K. Voss, M. Heiner, and I. Koch, "Steady state analysis of metabolic pathways using Petri nets," *In Silico Biology*, vol. 3, no. 3, pp. 367–387, 2003.

[16] I. Zevedei-Oancea and S. Schuster, "Topological analysis of metabolic networks based on Petri net theory," *In Silico Biology*, vol. 3, no. 3, pp. 323–345, 2003.

[17] M. Heiner, I. Koch, and J. Will, "Model validation of biological pathways using Petri nets–demonstrated for apoptosis," *Bio Systems*, vol. 75, no. 1-3, pp. 15–28, Jul. 2004.

[18] I. Koch, B. H. Junker, and M. Heiner, "Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber," *Bioinformatics*, vol. 21, no. 7, pp. 1219–1226, Apr. 2005.

[19] M. Heiner and I. Koch, "Petri Net Based Model Validation in Systems Biology," in *Applications and Theory of Petri Nets 2004*. Springer Berlin Heidelberg, 2004, pp. 216–237.

[20] D. A. Fell and A. Wagner, "The small world of metabolism," *Nature Biotechnology*, vol. 18, no. 11, pp. 1121–1122, Nov. 2000.

[21] A. Wagner and D. A. Fell, "The small world inside large metabolic networks." *Proceedings. Biological sciences / The Royal Society*, vol. 268, no. 1478, pp. 1803–1810, Sep. 2001.

[22] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, Oct. 2000.

[23] R. Tanaka, "Scale-Rich Metabolic Networks," *Physical Review Letters*, vol. 94, no. 16, p. 168101, Apr. 2005.

[24] R. Tanaka, M. Csete, and J. C. Doyle, "Highly optimised global organisation of metabolic networks," *Systems Biology, IEEE Proceedings*, vol. 152, no. 4, pp. 179–184, 2005.

[25] M. Csete and J. C. Doyle, "Bow ties, metabolism and disease," *Trends in Biotechnology*, vol. 22, no. 9, pp. 446–450, Sep. 2004.

[26] H.-W. Ma and A.-P. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, vol. 19, no. 11, pp. 1423–1430, Jul. 2003.

[27] J. Zhao, H. Yu, J.-H. Luo, Z.-W. Cao, and Y.-X. Li, "Hierarchical modularity of nested bow-ties in metabolic networks." *BMC bioinformatics*, vol. 7, no. 1, p. 386, Jan. 2006.

[28] K. Takemoto and T. Akutsu, "Origin of structural difference in metabolic networks with respect to temperature." *BMC systems biology*, vol. 2, p. 82, Jan. 2008.

[29] M. Parter, N. Kashtan, and U. Alon, "Environmental variability and modularity of bacterial metabolic networks." *BMC evolutionary biology*, vol. 7, no. 1, p. 169, Jan. 2007.

[30] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin, "The evolution of modularity in bacterial metabolic networks," *Proceedings of the National Academy of Sciences*, vol. 105, no. 19, pp. 6976–6981, May 2008.

[31] E. Borenstein, M. Kupiec, M. W. Feldman, and E. Ruppin, "Large-scale reconstruction and phylogenetic analysis of metabolic environments," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 482–14 487, Sep. 2008.

[32] O. Ebenhöh, T. Handorf, and R. Heinrich, "Structural analysis of expanding metabolic networks." *Genome informatics. International Conference on Genome Informatics*, vol. 15, no. 1, pp. 35–45, Jan. 2004.

[33] J. Raymond and D. Segrè, "The effect of oxygen on biochemical networks and the evolution of complex life." *Science (New York, N.Y.)*, vol. 311, no. 5768, pp. 1764–7, Mar. 2006.

[34] J. Zhao, G.-H. Ding, L. Tao, H. Yu, Z.-H. Yu, J.-H. Luo, Z.-W. Cao, and Y.-X. Li, "Modular co-evolution of metabolic networks." *BMC bioinformatics*, vol. 8, no. 1, p. 311, Jan. 2007.

[35] W.-c. Liu, W.-h. Lin, A. J. Davis, F. Jordán, H.-t. Yang, and M.-j. Hwang, "A network perspective on the topological importance of enzymes and their phylogenetic conservation." *BMC bioinformatics*, vol. 8, no. 1, p. 121, Jan. 2007.

[36] J. Diaz-Mejia, E. Perez-Rueda, and L. Segovia, "A network perspective on the evolution of metabolism by gene duplication," *Genome Biology*, vol. 8, no. 2, p. R26, 2007.

[37] V. Spirin, M. S. Gelfand, A. A. Mironov, and L. A. Mirny, "A metabolic network in the evolutionary context: Multiscale structure and modularity," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8774–8779, Jun. 2006.

[38] A. J. Greenberg, S. R. Stockwell, and A. G. Clark, "Evolutionary constraint and adaptation in the metabolic network of Drosophila." *Molecular biology and evolution*, vol. 25, no. 12, pp. 2537–46, Dec. 2008.

[39] D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabási, "The implications of human metabolic network topology for disease comorbidity," *Proceedings of the National Academy of Sciences*, vol. 105, no. 29, pp. 9880–9885, Jul. 2008.

[40] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics," *PLoS Computational Biology*, vol. 3, no. 4, p. e59, Apr. 2007.

[41] H. Yu, Y. Xia, V. Trifonov, and M. Gerstein, "Design principles of molecular networks revealed by global comparisons and composite motifs," *Genome Biology*, vol. 7, no. 7, p. R55, 2006.

[42] D. Croes, F. Couche, S. J. Wodak, and J. van Helden, "Inferring Meaningful Pathways in Weighted Metabolic Networks," *Journal of Molecular Biology*, vol. 356, no. 1, pp. 222–236, 2006.

[43] D. Croes, F. Couche, S. J. Wodak, and J. van Helden, "Metabolic PathFinding: inferring relevant pathways in biochemical networks." *Nucleic acids research*, vol. 33, no. Web Server issue, pp. W326–30, Jul. 2005.

[44] Y. Zheng, J. D. Szustakowski, L. Fortnow, R. J. Roberts, and S. Kasif, "Computational Identification of Operons in Microbial Genomes," *Genome Research*, vol. 12, no. 8, pp. 1221–1230, 2002.

[45] 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*, vol. 318, no. 3, pp. 911–932, May 2002.

[46] H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa, "A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters," *Nucleic Acids Research*, vol. 28, no. 20, pp. 4021–4028, Oct. 2000.

[47] A. G. Smart, L. A. N. Amaral, and J. M. Ottino, "Cascading failure and robustness in metabolic networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 36, pp. 13 223–8, Sep. 2008.

[48] S. Klamt, U.-U. Haus, and F. Theis, "Hypergraphs and Cellular Networks," *PLoS Computational Biology*, vol. 5, no. 5, p. e1000385, May 2009.

[49] G. Gallo, G. Longo, S. Pallottino, and S. Nguyen, "Directed hypergraphs and applications," *Discrete Appl. Math.*, vol. 42, no. 2-3, pp. 177–201, 1993.

[50] B. O. Palsson, *Systems biology: Properties of Reconstructed Networks*. Cambridge University Press, 2006.

[51] C. Berge, *Graphs and hypergraphs*, ser. English. Amsterdam, New York: North-Holland Pub. Co., American Elsevier Pub. Co., 1976.

[52] C. Berge, *Hypergraphs*. Elsevier, 1989.

[53] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002.

[54] A. Varma and B. O. Palsson, "Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use," *Bio/Technology*, vol. 12, no. 10, pp. 994–998, Oct. 1994.

[55] A. M. Feist, M. J. Herrgard, J. L. R. Ines Thiele, and B. O. Palsson, "Reconstruction of biochemical networks in microorganisms," *Nat Rev Microbiology*, vol. 7, pp. 129–143, 2009.

[56] C. H. Schilling and B. O. Palsson, "Assessment of the Metabolic Capabilities of Haemophilus influenzae Rd through a Genome-scale Pathway Analysis," *Journal of Theoretical Biology*, vol. 203, no. 3, pp. 249–283, 2000.

[57] J. S. Edwards and B. O. Palsson, "The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities," *Proceedings of the National Academy of Sciences*, vol. 97, no. 10, pp. 5528–5533, May 2000.

[58] J. S. Edwards, R. U. Ibarra, and B. O. Palsson, "In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data," *Nat Biotech*, vol. 19, no. 2, pp. 125–130, Feb. 2001.

[59] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles, "Metabolic network structure determines key aspects of functionality and regulation," *Nature*, vol. 420, no. 6912, pp. 190–193, Nov. 2002.

[60] J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson, "An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)." *Genome biology*, vol. 4, no. 9, p. R54, Jan. 2003.

[61] J. Förster, I. Famili, P. Fu, B. O. . A. Palsson, J. Nielsen, and J. FÃÂűrster, "Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network," *Genome Research*, vol. 13, no. 2, pp. 244–253, Feb. 2003.

[62] I. Famili, J. Förster, J. Nielsen, and B. O. Palsson, "Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 23, pp. 13 134–13 139, Nov. 2003.

[63] J. S. Edwards and B. O. Palsson, "Systems Properties of the Haemophilus influenzae Rd Metabolic Genotype," *J. Biol. Chem.*, vol. 274, no. 25, pp. 17 410–17 416, Jun. 1999.

[64] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. O. Palsson, "A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information." *Molecular systems biology*, vol. 3, p. 121, Jan. 2007.

[65] C. H. Schilling, M. W. Covert, I. Famili, G. M. Church, J. S. Edwards, and B. O. Palsson, "Genome-Scale Metabolic Model of Helicobacter pylori 26695," *J. Bacteriol.*, vol. 184, no. 16, pp. 4582–4593, 2002.

[66] 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." *Journal of bacteriology*, vol. 187, no. 16, pp. 5818–30, Aug. 2005.

[67] N. C. Duarte, M. J. Herrgå rd, and B. O. Palsson, "Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model." *Genome research*, vol. 14, no. 7, pp. 1298–309, Jul. 2004.

[68] L. Kuepfer, U. Sauer, and L. M. Blank, "Metabolic functions of duplicate genes in Saccharomyces cerevisiae." *Genome research*, vol. 15, no. 10, pp. 1421–30, Oct. 2005.

[69] M. J. Herrgå rd, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttilä, E. Klipp, B. O. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen, and D. B. Kell, "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology." *Nature biotechnology*, vol. 26, no. 10, pp. 1155–60, Oct. 2008.

[70] P. Romero, J. Wagg, M. Green, D. Kaiser, M. Krummenacker, and P. Karp, "Computational prediction of human metabolic pathways from the complete human genome," *Genome Biology*, vol. 6, no. 1, p. R2, 2004.

[71] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. O. . A. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data," *Proceedings of the National Academy of Sciences*, vol. 104, no. 6, pp. 1777–1782, Feb. 2007.

[72] I. Yeh, T. Hanekamp, S. Tsoka, P. D. Karp, and R. B. Altman, "Computational Analysis of Plasmodium falciparum Metabolism: Organizing Genomic Information to Facilitate Drug Discovery," *Genome Research*, vol. 14, no. 5, pp. 917–924, May 2004.

[73] T. D. Vo, H. J. Greenberg, and B. O. Palsson, "Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data," *The Journal of Biological Chemistry*, vol. 279, no. 38, pp. 39 532–39 540, Sep. 2004.

[74] I. Thiele, N. D. Price, T. D. Vo, and B. O. Palsson, "Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet." *The Journal of biological chemistry*, vol. 280, no. 12, pp. 11 683–95, Mar. 2005.

[75] S. Tsoka, D. Simon, and C. A. Ouzounis, "Automated metabolic reconstruction for Methanococcus jannaschii," *Archaea*, vol. 1, no. 4, p. 223âĂŞ229, Oct. 2004.

[76] S. A. Becker and B. O. Palsson, "Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation." *BMC microbiology*, vol. 5, no. 1, p. 8, Jan. 2005.

[77] M. Heinemann, A. Kümmel, R. Ruinatscha, and S. Panke, "In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network." *Biotechnology and bioengineering*, vol. 92, no. 7, pp. 850–64, Dec. 2005.

[78] A. P. Oliveira, J. Nielsen, and J. Förster, "Modeling Lactococcus lactis using a genome-scale flux model." *BMC microbiology*, vol. 5, no. 1, p. 39, Jan. 2005.

[79] K. Sheikh, J. Förster, and L. K. Nielsen, "Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus." *Biotechnology progress*, vol. 21, no. 1, pp. 112–21, Feb. 2005.

[80] I. Borodina, P. Krabben, and J. Nielsen, "Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism." *Genome Res*, vol. 15, no. 6, pp. 820–829, Jun. 2005.

[81] R. Mahadevan, D. R. Bond, J. E. Butler, A. Esteve-Nuñez, M. V. Coppi, B. O. Palsson, C. H. Schilling, and D. R. Lovley, "Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling." *Applied and environmental microbiology*, vol. 72, no. 2, pp. 1558–68, Feb. 2006.

[82] B. Teusink, A. Wiersma, D. Molenaar, C. Francke, W. M. de Vos, R. J. Siezen, and E. J. Smid, "Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model." *The Journal of biological chemistry*, vol. 281, no. 52, pp. 40 041–8, Dec. 2006.

[83] G. J. E. Baart, B. Zomer, A. de Haan, L. A. van der Pol, E. C. Beuvery, J. Tramper, and D. E. Martens, "Modeling Neisseria meningitidis metabolism: from genome to metabolic fluxes," *Genome Biology*, vol. 8, no. 7, p. R136, 2007.

[84] D. J. V. Beste, T. Hooper, G. Stewart, B. Bonde, C. Avignone-Rossa, M. E. Bushell, P. Wheeler, S. Klamt, A. M. Kierzek, and J. McFadden, "GSMN-TB: a web-based genome-scale network model of Mycobacterium tuberculosis metabolism." *Genome biology*, vol. 8, no. 5, p. R89, Jan. 2007.

[85] N. Jamshidi and B. O. Palsson, "Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets." *BMC systems biology*, vol. 1, p. 26, Jan. 2007.

[86] T. Y. Kim, H. U. Kim, J. M. Park, H. Song, J. S. Kim, and S. Y. Lee, "Genome-scale analysis of Mannheimia succiniciproducens metabolism," *Biotechnology and Bioengineering*, vol. 97, no. 4, pp. 657–671, 2007.

[87] O. Resendis-Antonio, J. L. Reed, S. Encarnación, J. Collado-Vides, and B. O. Palsson, "Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli." *PLoS computational biology*, vol. 3, no. 10, pp. 1887–95, Oct. 2007.

[88] O. Gonzalez, S. Gronau, M. Falb, F. Pfeiffer, E. Mendoza, R. Zimmer, and D. Oesterhelt, "Reconstruction, modeling & analysis of Halobacterium salinarum R-1 metabolism." *Molecular bioSystems*, vol. 4, no. 2, pp. 148–59, Feb. 2008.

[89] J. Nogales, B. O. Palsson, and I. Thiele, "A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory." *BMC systems biology*, vol. 2, no. 1, p. 79, Jan. 2008.

[90] A. K. Chavali, J. D. Whittemore, J. A. Eddy, K. T. Williams, and J. A. Papin, "Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major," *Molecular Systems Biology*, vol. 4, 2008.

[91] M. A. Oberhardt, J. Puchalka, K. E. Fryer, V. A. P. Martins dos Santos, and J. A. Papin, "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1." *Journal of bacteriology*, vol. 190, no. 8, pp. 2790–803, Apr. 2008.

[92] K. R. Kjeldsen and J. Nielsen, "In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network." *Biotechnology and bioengineering*, vol. 102, no. 2, pp. 583–97, Feb. 2009.

[93] B. CHANCE, G. R. WILLIAMS, W. F. HOLMES, and J. HIGGINS, "Respiratory enzymes in oxidative phosphorylation. V. A mechanism for oxidative phosphorylation," *The Journal of Biological Chemistry*, vol. 217, no. 1, pp. 439–451, Nov. 1955.

[94] B. CHANCE, D. GARFINKEL, J. HIGGINS, and B. HESS, "Metabolic control mechanisms. 5. A solution for the equations representing interaction between glycolysis and respiration in ascites tumor cells," *The Journal of Biological Chemistry*, vol. 235, pp. 2426–2439, Aug. 1960.

[95] D. GARFINKEL and B. HESS, "METABOLIC CONTROL MECHANISMS. VII.A DETAILED COMPUTER MODEL OF THE GLYCOLYTIC PATHWAY IN ASCITES CELLS." *The Journal of biological chemistry*, vol. 239, pp. 971–83, Apr. 1964.

[96] D. Garfinkel, L. Garfinkel, M. Pring, S. B. Green, and B. Chance, "Computer applications to biochemical kinetics," *Annual Review of Biochemistry*, vol. 39, pp. 473–498, 1970.

[97] R. Heinrich and T. A. Rapoport, "A linear steady-state treatment of enzymatic chains. General properties, control and effector strength." *European journal of biochemistry / FEBS*, vol. 42, no. 1, pp. 89–95, Feb. 1974.

[98] R. D. King, S. M. Garrett, and G. M. Coghill, "On the use of qualitative reasoning to simulate and identify metabolic pathways," *Bioinformatics*, vol. 21, no. 9, pp. 2017–2026, May 2005.

[99] K. R. Heidtke and S. Schulze-Kremer, "BioSim–a new qualitative simulation environment for molecular biology." *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 6, pp. 85–94, Jan. 1998.

[100] V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman, "Petri net representations in metabolic pathways." *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 1, pp. 328–36, Jan. 1993.

[101] M. W. Covert, C. H. Schilling, and B. O. Palsson, "Regulation of Gene Expression in Flux Balance Models of Metabolism," *Journal of Theoretical Biology*, vol. 213, no. 1, pp. 73–88, 2001.

[102] M. W. Covert and B. O. Palsson, "Transcriptional regulation in constraints-based metabolic models of Escherichia coli." *The Journal of biological chemistry*, vol. 277, no. 31, pp. 28 058–64, Aug. 2002.

[103] M. W. Covert and B. O. Palsson, "Constraints-based models: regulation of gene expression reduces the steady-state solution space." *Journal of theoretical biology*, vol. 221, no. 3, pp. 309–25, Apr. 2003.

[104] T. Shlomi, O. Berkman, and E. Ruppin, "Regulatory on/off minimization of metabolic flux changes after genetic perturbations." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 21, pp. 7695–700, May 2005.

[105] T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin, "A genome-scale computational study of the interplay between transcriptional regulation and metabolism." *Molecular systems biology*, vol. 3, no. 101, p. 101, Jan. 2007.

[106] J. M. Lee, J. Min Lee, E. P. Gianchandani, J. A. Eddy, and J. A. Papin, "Dynamic analysis of integrated signaling, metabolic, and regulatory networks." *PLoS computational biology*, vol. 4, no. 5, p. e1000086, May 2008.

[107] S. Chandrasekaran and N. D. Price, "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis." *Proc Natl Acad Sci U S A*, vol. 107, no. 41, pp. 17 845–17 850, Oct. 2010.

[108] J. Davies, "Origins and evolution of antibiotic resistance." *Microbiología (Madrid, Spain)*, vol. 12, no. 1, pp. 9–16, Mar. 1996.

[109] T. R. Walsh, "Combinatorial genetic evolution of multiresistance." *Current opinion in microbiology*, vol. 9, no. 5, pp. 476–82, Oct. 2006.

[110] P. P. Peralta-Yahya, F. Zhang, S. B. del Cardayre, and J. D. Keasling, "Microbial engineering for the production of advanced biofuels." *Nature*, vol. 488, no. 7411, pp. 320–8, Aug. 2012.

[111] G. Zhu, Y. Peng, B. Li, J. Guo, Q. Yang, and S. Wang, "Biological removal of nitrogen from wastewater." *Reviews of environmental contamination and toxicology*, vol. 192, pp. 159–95, Jan. 2008.

[112] W. F. Doolittle, Y. Boucher, C. L. Nesbø, C. J. Douady, J. O. Andersson, and A. J. Roger, "How big is the iceberg of which organellar genes in nuclear genomes are but the tip?" *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 358, no. 1429, pp. 39–57; discussion 57–8, Jan. 2003.

[113] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield, "Community structure and metabolism through reconstruction of microbial genomes from the environment." *Nature*, vol. 428, no. 6978, pp. 37–43, Mar. 2004.

[114] Z. Yang and B. Rannala, "Molecular phylogenetics: principles and practice." *Nature reviews. Genetics*, vol. 13, no. 5, pp. 303–14, May 2012.

[115] B. Snel, M. A. Huynen, and B. E. Dutilh, "Genome trees and the nature of genome evolution." *Annual review of microbiology*, vol. 59, pp. 191–209, Jan. 2005.

[116] J. A. Eisen, "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis." *Genome research*, vol. 8, no. 3, pp. 163–7, Mar. 1998.

[117] P. Pamilo and M. Nei, "Relationships between gene trees and species trees." *Molecular biology and evolution*, vol. 5, no. 5, pp. 568–83, Sep. 1988.

[118] P. Bonizzoni, G. Della Vedova, and R. Dondi, "Reconciling Gene Trees to a Species Tree," in *Algorithms and Complexity*, ser. LNCS. Springer Berlin, 2003, vol. 2653, pp. 636 –636.

[119] S. Ohno, *Evolution by gene duplication*. Springer-Verlag, 1970.

[120] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes." *Science (New York, N.Y.)*, vol. 290, no. 5494, pp. 1151–5, Nov. 2000.

[121] M. Long and K. Thornton, "Gene duplication and evolution." *Science (New York, N.Y.)*, vol. 293, no. 5535, p. 1551, Aug. 2001.

[122] J. Zhang, "Evolution by gene duplication: an update," *Trends in Ecology & Evolution*, 2003.

[123] S. G. Andersson and C. G. Kurland, "Reductive evolution of resident genomes." *Trends in microbiology*, vol. 6, no. 7, pp. 263–8, Jul. 1998.

[124] R. Gil, B. Sabater-Muñoz, A. Latorre, F. J. Silva, and A. Moya, "Extreme genome reduction in Buchnera spp.: Toward the minimal genome needed for symbiotic life," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 7, pp. 4454–4458, Apr. 2002.

[125] A. N. Khachane, K. N. Timmis, and V. A. P. Martins dos Santos, "Dynamics of reductive genome evolution in mitochondria and obligate intracellular microbes." *Molecular biology and evolution*, vol. 24, no. 2, pp. 449–56, Feb. 2007.

[126] K. Georgiades, V. Merhej, K. El Karkouri, D. Raoult, and P. Pontarotti, "Gene gain and loss events in Rickettsia and Orientia species." *Biology direct*, vol. 6, p. 6, Jan. 2011.

[127] H. Ochman, J. G. Lawrence, and E. A. Groisman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, no. 6784, pp. 299–304, May 2000.

[128] O. Zhaxybayeva, J. P. Gogarten, R. L. Charlebois, W. F. Doolittle, and R. T. Papke, "Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events." *Genome Res*, vol. 16, no. 9, pp. 1099–1108, Sep. 2006.

[129] J. Felsenstein, *Inferring Phylogenies*. Sunderland, Mass: Sinauer Associates, 2003.

[130] E. M. Jewett and N. A. Rosenberg, "iGLASS: an improvement to the GLASS method for estimating species trees from gene trees." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 19, no. 3, pp. 293–315, Mar. 2012.

[131] E. Mossel and S. Roch, "Incomplete lineage sorting: consistent phylogeny estimation from multiple loci." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 7, no. 1, pp. 166–71, 2010.

[132] L. Liu, L. Yu, and D. K. Pearl, "Maximum tree: a consistent estimator of the species tree." *Journal of mathematical biology*, vol. 60, no. 1, pp. 95–106, Jan. 2010.

[133] L. Liu, L. Yu, D. K. Pearl, and S. V. Edwards, "Estimating species phylogenies using coalescence times among sequences." *Systematic biology*, vol. 58, no. 5, pp. 468–77, Oct. 2009.

[134] B. Alix, D. A. Boubacar, and M. Vladimir, "T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks." *Nucleic acids research*, vol. 40, no. Web Server issue, pp. W573–9, Jul. 2012.

[135] C. Than, D. Ruths, and L. Nakhleh, "PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships." *BMC bioinformatics*, vol. 9, no. 1, p. 322, Jan. 2008.

[136] C. M. Zmasek and S. R. Eddy, "A simple algorithm to infer gene duplication and speciation events on a gene tree," *Bioinformatics*, vol. 17, no. 9, pp. 821–828, Sep. 2001.

[137] K. Chen, D. Durand, and M. Farach-Colton, "NOTUNG: a program for dating gene duplications and optimizing gene family trees." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 7, no. 3-4, pp. 429–47, Jan. 2000.

[138] J.-F. J.-F. Dufayard, L. Duret, S. Penel, M. Gouy, F. F. Rechenmann, and G. Perrière, "Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases," *Bioinformatics*, vol. 21, no. 11, pp. 2596–2603, Jun. 2005.

[139] A. Wehe, M. S. Bansal, J. G. Burleigh, and O. Eulenstein, "DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony." *Bioinformatics (Oxford, England)*, vol. 24, no. 13, pp. 1540–1, Jul. 2008.

[140] R. Chaudhary, M. S. Bansal, A. Wehe, D. Fernández-Baca, and O. Eulenstein, "iGTP: a software package for large-scale gene tree parsimony analysis." *BMC bioinformatics*, vol. 11, p. 574, Jan. 2010.

[141] L. Nakhleh, D. Ruths, and L.-s. Wang, "RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer," in *Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05). LNCS #3595.* Kunming: Springer, 2005, pp. 84—-93.

[142] J. H. Degnan and L. A. Salter, "Gene tree distributions under the coalescent process." *Evolution; international journal of organic evolution*, vol. 59, no. 1, pp. 24–37, Jan. 2005.

[143] Y. Wu, "Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood." *Evolution; international journal of organic evolution*, vol. 66, no. 3, pp. 763–75, Mar. 2012.

[144] L. S. Kubatko, B. C. Carstens, and L. L. Knowles, "STEM: species tree estimation using maximum likelihood for gene trees under coalescence." *Bioinformatics (Oxford, England)*, vol. 25, no. 7, pp. 971–3, Apr. 2009.

[145] A. Tofigh, M. Hallett, and J. Lagergren, "Simultaneous identification of duplications and lateral gene transfers." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 8, no. 2, pp. 517–35, 2011.

[146] M. A. Suchard, "Stochastic models for horizontal gene transfer: taking a random walk through tree space." *Genetics*, vol. 170, no. 1, pp. 419–31, May 2005.

[147] L. Liu, "BEST: Bayesian estimation of species trees under the coalescent model." *Bioinformatics (Oxford, England)*, vol. 24, no. 21, pp. 2542–3, Nov. 2008.

[148] S. Hohna, M. Defoin-Platel, and A. J. Drummond, "Clock-constrained tree proposal operators in Bayesian phylogenetic inference," in *2008 8th IEEE International Conference on BioInformatics and BioEngineering*. IEEE, Oct. 2008, pp. 1–7.

[149] B. R. Larget, S. K. Kotha, C. N. Dewey, and C. Ané, "BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis." *Bioinformatics (Oxford, England)*, vol. 26, no. 22, pp. 2910–1, Nov. 2010.

[150] B. Haegeman and J. S. Weitz, "A neutral theory of genome evolution and the frequency distribution of genes." *BMC genomics*, vol. 13, no. 1, p. 196, Jan. 2012.

[151] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution." *Nature reviews. Microbiology*, vol. 3, no. 9, pp. 679–87, Sep. 2005.

[152] V. Daubin and H. Ochman, "Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli." *Genome research*, vol. 14, no. 6, pp. 1036–42, Jun. 2004.

[153] S. S. Abby, E. Tannier, M. Gouy, and V. Daubin, "Lateral gene transfer as a support for the tree of life." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 13, pp. 4962–4967, Mar. 2012.

[154] R. Jain, M. C. Rivera, and J. A. Lake, "Horizontal gene transfer among genomes: the complexity hypothesis." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 7, pp. 3801–6, Mar. 1999.

[155] J. G. Lawrence and J. R. Roth, "Selfish operons: horizontal transfer may drive the evolution of gene clusters." *Genetics*, vol. 143, no. 4, pp. 1843–60, Aug. 1996.

[156] C. Pál and L. D. Hurst, "Evidence against the selfish operon theory." *Trends in genetics : TIG*, vol. 20, no. 6, pp. 232–4, Jul. 2004.

[157] J. O. Andersson and S. G. Andersson, "Genome degradation is an ongoing process in Rickettsia." *Molecular biology and evolution*, vol. 16, no. 9, pp. 1178–91, Sep. 1999.

[158] N. Moran and J. Wernegreen, "Lifestyle evolution in symbiotic bacteria: insights from genomics." *Trends in ecology & evolution*, vol. 15, no. 8, pp. 321–326, Aug. 2000.

[159] G. C. Conant and K. H. Wolfe, "Turning a hobby into a job: how duplicated genes find new functions." *Nature reviews. Genetics*, vol. 9, no. 12, pp. 938–50, Dec. 2008.

[160] J. A. Papin, T. Hunter, B. O. Palsson, and S. Subramaniam, "Reconstruction of cellular signalling networks and analysis of their properties." *Nat Rev Mol Cell Biol*, vol. 6, no. 2, pp. 99–111, Feb. 2005.

[161] V. Lacroix, L. Cottret, P. Thébault, and M.-F. Sagot, "An introduction to metabolic networks and their structural analysis." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 5, no. 4, pp. 594–617, Dec. 2008.

[162] K. R. Patil and J. Nielsen, "Uncovering transcriptional regulation of metabolism by using metabolic network topology." *Proc Nat Acad Sci USA*, vol. 102, pp. 2685–2689, 2005.

[163] M. Arita, "The metabolic world of Escherichia coli is not small," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 6, pp. 1543–1547, Feb. 2004.

[164] P. Holme, "Model validation of simple-graph representations of metabolism," *Journal of The Royal Society Interface*, vol. 6, no. 40, pp. 1027–1034, 2009.

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

[166] A. Mithani, G. M. Preston, and J. Hein, "Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison." *Bioinformatics (Oxford, England)*, vol. 25, no. 14, pp. 1831–2, Jul. 2009.

[167] U.-U. Haus, S. Klamt, and T. Stephen, "Computing knock-out strategies in metabolic networks," *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, vol. 15, no. 3, pp. 259–268, Apr. 2008.

[168] T. Handorf, O. Ebenhöh, and R. Heinrich, "Expanding Metabolic Networks: Scopes of Compounds, Robustness, and Evolution," *Journal of Molecular Evolution*, vol. 61, no. 4, pp. 498–512, Oct. 2005.

[169] F. Ay, T. Kahveci, and V. DE Crécy-Lagard, "A fast and accurate algorithm for comparative analysis of metabolic pathways." *Journal of bioinformatics and computational biology*, vol. 7, no. 3, p. 389, 2009.

[170] A. Varma and B. O. Palsson, "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110." *Appl. Environ. Microbiol.*, vol. 60, no. 10, pp. 3724–3731, 1994.

[171] Z. Hu, J. Mellor, J. Wu, M. Kanehisa, J. M. Stuart, and C. DeLisi, "Towards zoomable multidimensional maps of the cell," *Nature Biotechnology*, vol. 25, no. 5, pp. 547–554, 2007.

[172] F. Bertault and P. Eades, "Drawing Hypergraphs in the Subset Standard (Short Demo Paper)," in *Graph Drawing*. Springer Berlin Heidelberg, 2001, pp. 45–76.

[173] J. C. Nacher, N. Ueda, T. Yamada, M. Kanehisa, and T. Akutsu, "Clustering under the line graph transformation: application to reaction network." *BMC bioinformatics*, vol. 5, p. 207, Jan. 2004.

[174] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality and clustering in complex hyper-networks," *Physica A: Statistical Mechanics and its Applications*, vol. 364, pp. 581–594, May 2006.

[175] J. C. Nacher, N. Ueda, T. Yamada, M. Kanehisa, and T. Akutsu, "Study on the clustering coefficients in metabolic network using a hierarchical framework," *International workshop on bioinformatics and systems biology*, pp. 34–35, 2004.

[176] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopaedia of Genes and Genomes," *Nucl. Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.

[177] P. Erdös and A. Rényi, "On random graphs," *Publ. Math.*, vol. Debrecen 6, pp. 290–297, 1959.

[178] A.-L. Barabási, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[179] M. Lynch, "The evolution of genetic networks by non-adaptive processes," *Nature Reviews Genetics*, vol. 8, no. 10, pp. 803–813, 2007.

[180] A. Wagner, "Neutralism and selectionism: a network-based reconciliation," *Nature Reviews Genetics*, vol. 9, no. 12, pp. 965–974, 2008.

[181] D. J. Watts and S. H. Strogatz, "Collective dynamics of /'small-world/' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[182] A. Barrat and M. Weigt, "On the properties of small-world network models," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 13, no. 3, p. 19, Mar. 1999.

[183] M. Arita, "Scale-Freeness and Biological Networks," *J Biochem(Tokyo)*, vol. 138, pp. 1–4, 2005.

[184] H. Ma and A.-P. Zeng, "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms," *Bioinformatics*, vol. 19, no. 2, pp. 270–277, 2003.

[185] J. van Helden, L. Wernisch, D. Gilbert, and S. J. Wodak, "Graph-based analysis of metabolic networks." *Ernst Schering Research Foundation workshop*, vol. 38, no. 38, pp. 245–74, Jan. 2002.

[186] K. Faust, D. Croes, and J. van Helden, "Metabolic pathfinding using RPAIR annotation," *J Mol Biol*, vol. 388, no. 2, pp. 390–414, May 2009.

[187] D. Zhu and Z. S. Qin, "Structural comparison of metabolic networks in selected single cell organisms." *BMC bioinformatics*, vol. 6, no. 1, p. 8, Jan. 2005.

[188] T. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. Stringer-Calvert, J. Tenenbaum, and P. Karp, "BioWarehouse: a bioinformatics database warehouse toolkit," *BMC Bioinformatics*, vol. 7, no. 1, p. 170, 2006.

[189] M. G. Poolman, B. K. Bonde, A. Gevorgyan, H. H. Patel, and D. A. Fell, "Challenges to be faced in the reconstruction of metabolic networks from public databases," *IEEE Proceedings - Systems Biology*, vol. 153, no. 5, pp. 379–384, 2006.

[190] T. Blum and O. Kohlbacher, "Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 15, no. 6, pp. 565–76, 2008.

[191] C. Lemer, E. Antezana, F. Couche, F. Fays, X. Santolaria, R. Janky, Y. Deville, J. Richelle, and S. J. Wodak, "The aMAZE LightBench: a web

interface to a relational database of cellular processes," *Nucl. Acids Res.*, vol. 32, no. suppl, pp. D443–448, 2004.

[192] Y. Deville, D. Gilbert, J. van Helden, and S. J. Wodak, "An overview of data models for the analysis of biochemical pathways," *Brief Bioinform*, vol. 4, no. 3, pp. 246–259, 2003.

[193] M. Kotera, M. Hattori, M. Oh, R. T. Yamamoto, T. Komeno, J. Yabuzaki, K. Tonomura, S. Goto, and M. Kanehisa, "RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions," in *Genome Inform*, vol. 15, 2004, p. P062.

[194] S. A. Rahman, P. Advani, R. Schunk, R. Schrader, and D. Schomburg, "Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC)," *Bioinformatics*, vol. 21, no. 7, pp. 1189–1193, 2005.

[195] J. D. Crabtree, D. P. Mehta, and T. M. Kouri, "An open-source java platform for automated reaction mapping." *Journal of chemical information and modeling*, vol. 50, no. 9, pp. 1751–6, Sep. 2010.

[196] P. Ravikirthi, P. F. Suthers, and C. D. Maranas, "Construction of an E. Coli genome-scale atom mapping model for MFA calculations." *Biotechnology and bioengineering*, vol. 108, no. 6, pp. 1372–82, Jun. 2011.

[197] T. Szyperski, "13C-NMR, MS and metabolic flux balancing in biotechnology research." *Quarterly reviews of biophysics*, vol. 31, no. 1, pp. 41–106, Feb. 1998.

[198] T. Hogiri, C. Furusawa, Y. Shinfuku, N. Ono, and H. Shimizu, "Analysis of metabolic network based on conservation of molecular structure," *Biosystems*, vol. 95, no. 3, pp. 175–178, 2009.

[199] F. Mu, R. F. Williams, C. J. Unkefer, P. J. Unkefer, J. R. Faeder, and W. S. Hlavacek, "Carbon-fate maps for metabolic reactions." *Bioinformatics (Oxford, England)*, vol. 23, no. 23, pp. 3193–9, Dec. 2007.

[200] Yukako Tohsato, Yu Nishimura, Y. Tohsato, and Y. Nishimura, "Reaction Similarities Focusing Substructure Changes of Chemical Compounds and Metabolic Pathway Alignments," *Information and Media Technologies*, vol. 4, no. 2, pp. 390–399, 2009.

[201] A. Rivas-Ubach, J. Sardans, M. Pérez-Trujillo, M. Estiarte, and J. Peñuelas, "Strong relationship between elemental stoichiometry and metabolome in plants." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 11, pp. 4181–6, Mar. 2012.

[202] B. M. Trost, "Atom Economy – A Challenge for Organic Synthesis: Homogeneous Catalysis Leads the Way," *Angewandte Chemie International Edition in English*, vol. 34, no. 3, pp. 259–281, Feb. 1995.

[203] A. P. Heath, G. N. Bennett, and L. E. Kavraki, "Finding metabolic pathways using atom tracking," *Bioinformatics*, vol. 26, no. 12, pp. 1548–1555, 2010.

[204] E. Pitkänen, P. Jouhten, and J. Rousu, "Inferring branching pathways in genome-scale metabolic networks." *BMC systems biology*, vol. 3, p. 103, Jan. 2009.

[205] J. Schellenberger, D. C. Zielinski, W. Choi, S. Madireddi, V. Portnoy, D. A. Scott, J. L. Reed, A. L. Osterman, and B. O. Palsson, "Predicting outcomes of steady-state 13C isotope tracing experiments with Monte Carlo sampling." *BMC systems biology*, vol. 6, no. 1, p. 9, Jan. 2012.

[206] W. Wiechert and A. A. de Graaf, "Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments." *Biotechnology and bioengineering*, vol. 55, no. 1, pp. 101–17, Jul. 1997.

[207] T. Szyperski, "Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism." *European journal of biochemistry / FEBS*, vol. 232, no. 2, pp. 433–48, Sep. 1995.

[208] K. Schmidt, M. Carlsen, J. Nielsen, and J. Villadsen, "Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices." *Biotechnology and bioengineering*, vol. 55, no. 6, pp. 831–40, Sep. 1997.

[209] J. Rosenthal and G. Murphy, "Group Theory and the Vibrations of Polyatomic Molecules," *Reviews of Modern Physics*, vol. 8, no. 4, pp. 317–346, Oct. 1936.

[210] D. Harrisand and M. Bertolucci, *Symmetry and spectroscopy: an introduction to vibrational and electronic spectroscopy.* Dover Pubns, 1989.

[211] P. Atkins and J. de Paula, *Physical Chemistry*, 9th ed. W. H. Freeman, 2009.

[212] 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*, vol. 125, no. 39, pp. 11 853–11 865, 2003.

[213] M. Hattori, N. Tanaka, M. Kanehisa, and S. Goto, "SIMCOMP/SUBCOMP: chemical structure search servers for network analyses." *Nucleic acids research*, vol. 38, no. Web Server issue, pp. W652–6, Jul. 2010.

[214] J. D. Crabtree and D. P. Mehta, "Automated reaction mapping," *J. Exp. Algorithmics*, vol. 13, pp. 1.15—-1.29, 2009.

[215] M. Heinonen, S. Lappalainen, T. Mielikäinen, and J. Rousu, "Computing atom mappings for biochemical reactions without subgraph isomorphism." *Journal*

of computational biology : a journal of computational molecular cell biology, vol. 18, no. 1, pp. 43–58, Jan. 2011.

[216] L. G. Shapiro and R. M. Haralick, "Structural Descriptions and Inexact Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, no. 5, pp. 504–519, Sep. 1981.

[217] G. P. Moss, "Basic terminology of stereochemistry (IUPAC Recommendations 1996)," *Pure and Applied Chemistry*, vol. 68, no. 12, pp. 2193–2222, 1996.

[218] E. M. Luks, "Isomorphism of graphs of bounded valence can be tested in polynomial time," in *Foundations of Computer Science, 1980., 21st Annual Symposium on*, 1980, pp. 42–49.

[219] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "An improved algorithm for matching large graphs," in *In: 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen*, 2001, pp. 149–159.

[220] E. Artin, *Galois Theory: Lectures Delivered at the University of Notre Dame by Emil Artin (Notre Dame Mathematical Lectures, Number 2)*. Dover Publications, 1997.

[221] M. R. Antoniewicz, J. K. Kelleher, and G. Stephanopoulos, "Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions." *Metabolic engineering*, vol. 9, no. 1, pp. 68–86, Jan. 2007.

[222] R. Balducci and R. S. Pearlman, "Efficient exact solution of the ring perception problem," *Journal of Chemical Information and Computer Sciences*, vol. 34, no. 4, pp. 822–831, Jul. 1994.

[223] E. Pitkänen, "Computational Methods for Reconstruction and Analysis of Genome-Scale Metabolic Networks," Ph.D. dissertation, University of Helsinki, 2010.

[224] M. L. Blinov, J. R. Faeder, B. Goldstein, and W. S. Hlavacek, "BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains." *Bioinformatics*, vol. 20, no. 17, pp. 3289–3291, Nov. 2004.

[225] A. S. Yang, "Modularity, evolvability, and adaptive radiations: a comparison of the hemi- and holometabolous insects," *Evolution and Development*, vol. 3, no. 2, pp. 59–72, Mar. 2001.

[226] T. F. Hansen, "Is modularity necessary for evolvability?" *Biosystems*, vol. 69, no. 2-3, pp. 83–94, May 2003.

[227] C. K. Griswold, "Pleiotropic mutation, modularity and evolvability." *Evolution & development*, vol. 8, no. 1, pp. 81–93, 2006.

[228] J. F. Y. Brookfield, "Evolution and evolvability: celebrating Darwin 200." *Biology letters*, vol. 5, no. 1, pp. 44–6, Feb. 2009.

[229] A. Hintze and C. Adami, "Evolution of complex modular biological networks." *PLoS computational biology*, vol. 4, no. 2, p. e23, Feb. 2008.

[230] P. Holme, "Metabolic robustness and network modularity: a model study." *PloS one*, vol. 6, no. 2, p. e16605, Jan. 2011.

[231] R. Harrison, B. Papp, C. Pál, S. G. Oliver, and D. Delneri, "Plasticity of genetic interactions in metabolic networks of yeast." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 7, pp. 2307–12, Feb. 2007.

[232] O. S. Soyer and T. Pfeiffer, "Evolution under fluctuating environments explains observed robustness in metabolic networks." *PLoS computational biology*, vol. 6, no. 8, Jan. 2010.

[233] K. Takemoto and S. Borjigin, "Metabolic Network Modularity in Archaea Depends on Growth Conditions," *PLoS ONE*, vol. 6, no. 10, p. e25874, Oct. 2011.

[234] J. M. Peregrín-Alvarez, C. Sanford, and J. Parkinson, "The conservation and evolutionary modularity of metabolism." *Genome biology*, vol. 10, no. 6, p. R63, Jan. 2009.

[235] J. M. Peregrín-Alvarez, X. Xiong, C. Su, and J. Parkinson, "The Modular Organization of Protein Interactions in \textit{Escherichia coli}." *PLoS Comput Biol*, vol. 5, no. 10, p. e1000523, Oct. 2009.

[236] K. H. Ten Tusscher and P. Hogeweg, "Evolution of Networks for Body Plan Patterning; Interplay of Modularity, Robustness and Evolvability," *PLoS Computational Biology*, vol. 7, no. 10, p. e1002208, 2011.

[237] T. Dagan, Y. Artzy-Randrup, and W. Martin, "Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 29, pp. 10 039–44, Jul. 2008.

[238] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 26113, Feb. 2004.

[239] G. Wagner and L. Altenberg, "Complex Adaptations and the Evolution of Evolvability," *Evolution*, vol. 50, no. 3, pp. 967–976, 1996.

[240] M. E. J. Newman, "Modularity and community structure in networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–82, Jun. 2006.

[241] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, p. 66111, Dec. 2004.

[242] J. Duch and A. Arenas, "Community detection in complex networks using Extremal Optimization," *Physical Review E*, vol. 72, no. 2, p. 27104, Aug. 2005.

[243] G. Agarwal and D. Kempe, "Modularity-maximizing graph communities via mathematical programming," *The European Physical Journal B*, vol. 66, no. 3, pp. 409–418, Nov. 2008.

[244] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, Feb. 2010.

[245] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, pp. 298–305, 1973.

[246] A. Pothen, H. D. Simon, and K. P. Liou, "Partitioning Sparse Matrices with Eigenvectors of Graphs," *SIAM Journal on Matrix Analysis and Applications*, vol. 11, pp. 430–452, 1990.

[247] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger, "Hierarchical clustering using mutual information," *Europhysics Letters (EPL)*, vol. 70, no. 2, pp. 278–284, Apr. 2005.

[248] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life." *Science*, vol. 311, no. 5765, pp. 1283–1287, Mar. 2006.

[249] N. Kashtan and U. Alon, "Spontaneous evolution of modularity and network motifs." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13 773–8, Sep. 2005.

[250] R. Sokal and C. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Scientific Bulletin*, vol. 28, pp. 1409 – 1438, 1958.

[251] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," 2001.

[252] K. A. Lawrence, M. W. Jewett, P. A. Rosa, and F. C. Gherardini, "Borrelia burgdorferi bb0426 encodes a 2'-deoxyribosyltransferase that plays a central role in purine salvage." *Molecular microbiology*, vol. 72, no. 6, pp. 1517–29, Jun. 2009.

[253] K. Tilly, P. A. Rosa, and P. E. Stewart, "Biology of infection with Borrelia burgdorferi." *Infectious disease clinics of North America*, vol. 22, no. 2, pp. 217–34, v, Jun. 2008.

[254] F. O. Glöckner, M. Kube, M. Bauer, H. Teeling, T. Lombardot, W. Ludwig, D. Gade, A. Beck, K. Borzym, K. Heitmann, R. Rabus, H. Schlesner, R. Amann, and R. Reinhardt, "Complete genome sequence of the marine planctomycete Pirellula sp. strain 1." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, pp. 8298–303, Jul. 2003.

[255] V. Pérez-Brocal, R. Gil, S. Ramos, A. Lamelas, M. Postigo, J. M. Michelena, F. J. Silva, A. Moya, and A. Latorre, "A small microbial genome: the end of a long symbiotic relationship?" *Science (New York, N.Y.)*, vol. 314, no. 5797, pp. 312–3, Oct. 2006.

[256] A. E. Douglas, "Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera." *Annual review of entomology*, vol. 43, pp. 17–37, Jan. 1998.

[257] S. Fortunato and M. Barthélemy, "Resolution limit in community detection." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 1, pp. 36–41, Jan. 2007.

[258] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On Modularity Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 172–188, 2008.

[259] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," *Physical Review E*, vol. 81, no. 4, p. 46106, Apr. 2010.

[260] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, Mar. 2004.

[261] M. Koyutürk, W. Szpankowski, and A. Grama, "Assessing significance of connectivity and conservation in protein interaction networks." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 14, no. 6, pp. 747–64, 2007.

[262] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M.

Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White, "The Gene Ontology (GO) database and informatics resource." *Nucleic acids research*, vol. 32, no. Database issue, pp. D258–61, Jan. 2004.

[263] T. Ruths, D. Ruths, and L. Nakhleh, "GS2: an efficiently computable measure of GO-based similarity of gene sets." *Bioinformatics (Oxford, England)*, vol. 25, no. 9, pp. 1178–84, May 2009.

[264] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson, "Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration." *Bioinformatics (Oxford, England)*, vol. 24, no. 14, pp. 1650–1, Jul. 2008.

[265] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes." *Bioinformatics (Oxford, England)*, vol. 20, no. 18, pp. 3710–5, Dec. 2004.

[266] W. Zhou and L. Nakhleh, "Properties of metabolic graphs: biological organization or representation artifacts?" *BMC bioinformatics*, vol. 12, no. 1, p. 132, Jan. 2011.

[267] F. Hommais, S. Pereira, C. Acquaviva, P. Escobar-Páramo, and E. Denamur, "Single-nucleotide polymorphism phylotyping of Escherichia coli." *Applied and environmental microbiology*, vol. 71, no. 8, pp. 4784–92, Aug. 2005.

[268] Y. Wang and Z. Zhang, "Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variations in bacterial rRNA genes." *Microbiology (Reading, England)*, vol. 146 (Pt 1, pp. 2845–54, Nov. 2000.

[269] C. X. Chan, A. E. Darling, R. G. Beiko, and M. A. Ragan, "Are protein domains modules of lateral genetic transfer?" *PloS one*, vol. 4, no. 2, p. e4524, Jan. 2009.

[270] I. Matic, C. Rayssiguier, and M. Radman, "Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species." *Cell*, vol. 80, no. 3, pp. 507–15, Feb. 1995.

[271] W. H. Yap, Z. Zhang, and Y. Wang, "Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon." *Journal of bacteriology*, vol. 181, no. 17, pp. 5201–9, Sep. 1999.

[272] N. Igarashi, J. Harada, S. Nagashima, K. Matsuura, K. Shimada, and K. V. Nagashima, "Horizontal transfer of the photosynthesis gene cluster and operon

rearrangement in purple bacteria." *Journal of molecular evolution*, vol. 52, no. 4, pp. 333–41, Apr. 2001.

[273] M. V. Omelchenko, K. S. Makarova, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin, "Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ." *Genome biology*, vol. 4, no. 9, p. R55, Jan. 2003.

[274] S. Castillo-Ramírez, J. F. Vázquez-Castellanos, V. González, and M. A. Cevallos, "Horizontal gene transfer and diverse functional constrains within a common replication-partitioning system in Alphaproteobacteria: the repABC operon." *BMC genomics*, vol. 10, p. 536, Jan. 2009.

[275] Y. Akagi, H. Akamatsu, H. Otani, and M. Kodama, "Horizontal chromosome transfer, a mechanism for the evolution and differentiation of a plant-pathogenic fungus." *Eukaryotic cell*, vol. 8, no. 11, pp. 1732–8, Nov. 2009.

[276] A. C. Retchless and J. G. Lawrence, "Phylogenetic incongruence arising from fragmented speciation in enteric bacteria." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 25, pp. 11 453–8, Jun. 2010.

[277] J. A. Eisen and C. M. Fraser, "Phylogenomics: intersection of evolution and genomics." *Science (New York, N.Y.)*, vol. 300, no. 5626, pp. 1706–7, Jun. 2003.

[278] T. Dagan and W. Martin, "The tree of one percent." *Genome biology*, vol. 7, no. 10, p. 118, Jan. 2006.

[279] K. P. Williams, J. J. Gillespie, B. W. S. Sobral, E. K. Nordberg, E. E. Snyder, J. M. Shallom, and A. W. Dickerman, "Phylogeny of gammaproteobacteria." *Journal of bacteriology*, vol. 192, no. 9, pp. 2305–14, May 2010.

[280] C. P. Andam and J. P. Gogarten, "Biased gene transfer in microbial evolution." *Nature reviews. Microbiology*, vol. 9, no. 7, pp. 543–55, Jul. 2011.

[281] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand, "Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees." *Bioinformatics (Oxford, England)*, vol. 28, no. 18, pp. i409–i415, Sep. 2012.

[282] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal gene transfer in prokaryotes: quantification and classification." *Annu Rev Microbiol*, vol. 55, pp. 709–742, 2001.

[283] Y. Boucher, C. J. Douady, R. T. Papke, D. A. Walsh, M. E. R. Boudreau, C. L. Nesbø, R. J. Case, and W. F. Doolittle, "Lateral gene transfer and the origins of prokaryotic groups." *Annual review of genetics*, vol. 37, pp. 283–328, Jan. 2003.

[284] B. G. Mirkin, T. I. Fenner, M. Y. Galperin, and E. V. Koonin, "Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes." *BMC evolutionary biology*, vol. 3, p. 2, Jan. 2003.

[285] B. Boussau, E. O. Karlberg, A. C. Frank, B.-A. Legault, and S. G. E. Andersson, "Computational inference of scenarios for alpha-proteobacterial genome evolution." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9722–7, Jun. 2004.

[286] S. Linz, A. Radtke, and A. von Haeseler, "A likelihood framework to measure horizontal gene transfer." *Molecular biology and evolution*, vol. 24, no. 6, pp. 1312–9, Jun. 2007.

[287] S. S. Abby, E. Tannier, M. Gouy, and V. Daubin, "Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests." *BMC bioinformatics*, vol. 11, p. 324, Jan. 2010.

[288] J.-P. J.-P. Doyon, V. Ranwez, V. Daubin, and V. Berry, "Models, algorithms and programs for phylogeny reconciliation," *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 392–400, Sep. 2011.

[289] O. Akerborg, B. Sennblad, L. Arvestad, and J. Lagergren, "Simultaneous Bayesian gene tree reconstruction and reconciliation analysis." *Proc Natl Acad Sci U S A*, vol. 106, no. 14, pp. 5714–5719, Apr. 2009.

[290] M. D. Rasmussen and M. Kellis, "A Bayesian approach for fast and accurate gene tree reconstruction." *Molecular biology and evolution*, vol. 28, no. 1, pp. 273–90, Jan. 2011.

[291] A. Heger and L. Holm, "Exhaustive enumeration of protein domain families." *Journal of molecular biology*, vol. 328, no. 3, pp. 749–67, May 2003.

[292] E. Bapteste, P. Lopez, F. Bouchard, F. Baquero, J. O. McInerney, and R. M. Burian, "Evolutionary analyses of non-genealogical bonds produced by introgressive descent." *Proceedings of the National Academy of Sciences of the United States of America*, Oct. 2012.

[293] Y.-C. Wu, M. D. Rasmussen, and M. Kellis, "Evolution at the subgene level: domain rearrangements in the Drosophila phylogeny." *Molecular biology and evolution*, vol. 29, no. 2, pp. 689–705, Feb. 2012.

[294] D. Vitkup, E. Melamud, J. Moult, and C. Sander, "Completeness in structural genomics." *Nature structural biology*, vol. 8, no. 6, pp. 559–66, Jun. 2001.

[295] K. Suhre and J.-M. Claverie, "FusionDB: a database for in-depth analysis of prokaryotic gene fusion events." *Nucleic acids research*, vol. 32, no. Database issue, pp. D273–6, Jan. 2004.

[296] G. Leonard and T. A. Richards, "Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 52, pp. 21 402–7, Dec. 2012.

[297] G. Yona, N. Linial, N. Tishby, and M. Linial, "A map of the protein space–an automatic hierarchical classification of all protein sequences." *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 6, pp. 212–21, Jan. 1998.

[298] W. N. Grundy, "Homology detection via family pairwise search." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 5, no. 3, pp. 479–91, Jan. 1998.

[299] J. Gracy and P. Argos, "Automated protein sequence database classification. II. Delineation Of domain boundaries from sequence similarities." *Bioinformatics (Oxford, England)*, vol. 14, no. 2, pp. 174–87, Jan. 1998.

[300] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer, "The Pfam protein families database." *Nucleic acids research*, vol. 30, no. 1, pp. 276–80, Jan. 2002.

[301] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D281–D288, Jan. 2008.

[302] S. J. Sammut, R. D. Finn, and A. Bateman, "Pfam 10 years on: 10,000 families and still growing." *Briefings in bioinformatics*, vol. 9, no. 3, pp. 210–9, May 2008.

[303] A. Marchler-Bauer, J. B. Anderson, P. F. Cherukuri, C. DeWeese-Scott, L. Y. Geer, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, G. H. Marchler, M. Mullokandov, B. A. Shoemaker, V. Simonyan, J. S. Song, P. A. Thiessen, R. A. Yamashita, J. J. Yin, D. Zhang, and S. H. Bryant, "CDD: a Conserved Domain Database for protein classification." *Nucleic acids research*, vol. 33, no. Database issue, pp. D192–6, Jan. 2005.

[304] A. Marchler-Bauer, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, A. Tasneem, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, and S. H. Bryant, "CDD:

specific functional annotation with the Conserved Domain Database." *Nucleic acids research*, vol. 37, no. Database issue, pp. D205–10, Jan. 2009.

[305] W. P. Maddison and D. Maddison, "Mesquite: a modular system for evolutionary analysis," 2011.

[306] D. M. Downs, "Understanding microbial metabolism." *Annual review of microbiology*, vol. 60, pp. 533–59, Jan. 2006.

[307] C. A. Fewson, "Microbial metabolism of mandelate: a microcosm of diversity." *FEMS microbiology reviews*, vol. 4, no. 2, pp. 85–110, 1988.

[308] U. Wiesmann, "Biological nitrogen removal from wastewater." *Advances in biochemical engineering/biotechnology*, vol. 51, pp. 113–54, Jan. 1994.

[309] G. H. McArthur and S. S. Fong, "Toward engineering synthetic microbial metabolism." *Journal of biomedicine & biotechnology*, vol. 2010, p. 459760, Jan. 2010.

[310] L. Boto, "Horizontal gene transfer in evolution: facts and challenges." *Proceedings. Biological sciences / The Royal Society*, vol. 277, no. 1683, pp. 819–27, Mar. 2010.

[311] S. Garcia-Vallve, "Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes," *Genome Research*, vol. 10, no. 11, pp. 1719–1725, Nov. 2000.

[312] V. A. Albert, *Parsimony, Phylogeny, and Genomics*. Oxford University Press, Mar. 2006.

[313] O. Sakarya, K. S. Kosik, and T. H. Oakley, "Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony." *Bioinformatics (Oxford, England)*, vol. 24, no. 5, pp. 606–12, Mar. 2008.

[314] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." *Nucleic acids research*, vol. 39, no. Database issue, pp. D561–8, Jan. 2011.

[315] D. Cortez, P. Forterre, and S. Gribaldo, "A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes." *Genome Biol*, vol. 10, no. 6, p. R65, 2009.

[316] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic acids research*, vol. 32, no. 5, pp. 1792–7, Jan. 2004.

[317] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic acids research*, vol. 22, no. 22, pp. 4673–80, Nov. 1994.

[318] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." *Bioinformatics (Oxford, England)*, vol. 22, no. 21, pp. 2688–90, Nov. 2006.

[319] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." *Systematic biology*, vol. 59, no. 3, pp. 307–21, May 2010.

[320] K. Lee, F. Berthiaume, G. N. Stephanopoulos, and M. L. Yarmush, "Metabolic flux analysis: a powerful tool for monitoring tissue function." *Tissue engineering*, vol. 5, no. 4, pp. 347–68, Aug. 1999.

[321] 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*, vol. 15, no. 3, pp. 296–303, 1999.

[322] J. S. Edwards, M. Covert, and B. Palsson, "Metabolic modelling of microbes: the flux-balance approach." *Environmental microbiology*, vol. 4, no. 3, pp. 133–40, Mar. 2002.

[323] K. Raman and N. Chandra, "Flux balance analysis of biological systems: applications and challenges." *Briefings in bioinformatics*, vol. 10, no. 4, pp. 435–49, Jul. 2009.

[324] M. A. Oberhardt, B. O. Palsson, and J. A. Papin, "Applications of genome-scale metabolic reconstructions." *Molecular systems biology*, vol. 5, p. 320, Jan. 2009.

[325] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. O. Palsson, and M. J. Herrgard, "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox." *Nature protocols*, vol. 2, no. 3, pp. 727–38, Jan. 2007.

[326] A. Hoppe, S. Hoffmann, A. Gerasch, C. Gille, and H.-G. Holzhutter, "FASIMU: flexible software for flux-balance computation series in large metabolic networks," *BMC Bioinformatics*, vol. 12, no. 1, p. 28, 2011.

[327] A. Larhlimi, L. David, J. Selbig, and A. Bockmayr, "F2C2: a fast tool for the computation of flux coupling in genome-scale metabolic networks," *BMC Bioinformatics*, vol. 13, no. 1, p. 57, 2012.

[328] D. Segrè, D. Vitkup, and G. M. Church, "Analysis of optimality in natural and perturbed metabolic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 23, pp. 15 112–15 117, 2002.

[329] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork, "eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges." *Nucleic acids research*, vol. 40, no. Database issue, pp. D284–9, Jan. 2012.

[330] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." *Molecular biology and evolution*, vol. 17, no. 4, pp. 540–52, Apr. 2000.

[331] S. Penel, A.-M. Arigon, J.-F. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière, "Databases of homologous gene families for comparative genomics." *BMC bioinformatics*, vol. 10 Suppl 6, no. Suppl 6, p. S3, Jan. 2009.

[332] D. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sinauer Associates, 2003.

[333] D. Fischer and D. Eisenberg, "Finding families for genomic ORFans." *Bioinformatics (Oxford, England)*, vol. 15, no. 9, pp. 759–62, Sep. 1999.

[334] L. a. David and E. J. Alm, "Rapid evolutionary innovation during an Archaean genetic expansion." *Nature*, vol. 469, no. 7328, pp. 93–6, Jan. 2011.

[335] B. Snel, P. Bork, and M. A. Huynen, "Genomes in flux: the evolution of archaeal and proteobacterial gene content." *Genome research*, vol. 12, no. 1, pp. 17–25, Jan. 2002.

[336] J. G. Lawrence and H. Ochman, "Molecular archaeology of the Escherichia coli genome." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 16, pp. 9413–7, Aug. 1998.

[337] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro, "The EcoCyc Database." *Nucleic acids research*, vol. 30, no. 1, pp. 56–8, Jan. 2002.

[338] B. Snel, P. Bork, and M. Huynen, "Genome evolution. Gene fusion versus gene fission." *Trends in genetics : TIG*, vol. 16, no. 1, pp. 9–11, Jan. 2000.

[339] Y. Zheng, R. J. Roberts, and S. Kasif, "Segmentally variable genes: a new perspective on adaptation." *PLoS biology*, vol. 2, no. 4, p. E81, Apr. 2004.

[340] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *Journal of molecular biology*, vol. 247, no. 4, pp. 536–40, Apr. 1995.

[341] L. Lo Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia, "SCOP: a structural classification of proteins database." *Nucleic acids research*, vol. 28, no. 1, pp. 257–9, Jan. 2000.

[342] J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting, "SMART, a simple modular architecture research tool: identification of signaling domains." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 11, pp. 5857–64, May 1998.

[343] I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork, "SMART 5: domains in the context of genomes and networks." *Nucleic acids research*, vol. 34, no. Database issue, pp. D257–60, Jan. 2006.

[344] E. Portugaly, A. Harel, N. Linial, and M. Linial, "EVEREST: automatic identification and classification of protein domains in all protein sequences." *BMC bioinformatics*, vol. 7, p. 277, Jan. 2006.

[345] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites." *Nucleic acids research*, vol. 29, no. 1, pp. 37–40, Jan. 2001.

[346] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats, "InterPro: the integrative protein signature database." *Nucleic acids research*, vol. 37, no. Database issue, pp. D211–5, Jan. 2009.

[347] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research*, vol. 25, no. 17, pp. 3389–402, Sep. 1997.

[348] J. Knappe, H. P. Blaschkowski, P. Gröbner, and T. Schmitt, "Pyruvate formate-lyase of Escherichia coli: the acetyl-enzyme intermediate." *European journal of biochemistry / FEBS*, vol. 50, no. 1, pp. 253–63, Dec. 1974.

[349] J. Zhu and K. Shimizu, "The effect of pfl gene knockout on the metabolism for optically pure D-lactate production by Escherichia coli." *Applied microbiology and biotechnology*, vol. 64, no. 3, pp. 367–75, Apr. 2004.

[350] G. Sawers, C. Hesslinger, N. Muller, and M. Kaiser, "The glycyl radical enzyme TdcE can replace pyruvate formate-lyase in glucose fermentation." *Journal of bacteriology*, vol. 180, no. 14, pp. 3509–16, Jul. 1998.

[351] H. Wang and R. P. Gunsalus, "Coordinate regulation of the Escherichia coli formate dehydrogenase fdnGHI and fdhF genes in response to nitrate, nitrite, and formate: roles for NarL and NarP." *Journal of bacteriology*, vol. 185, no. 17, pp. 5076–85, Sep. 2003.

[352] M. J. Axley, D. A. Grahame, and T. C. Stadtman, "Escherichia coli formate-hydrogen lyase. Purification and properties of the selenium-dependent formate dehydrogenase component." *The Journal of biological chemistry*, vol. 265, no. 30, pp. 18 213–8, Oct. 1990.

[353] M. Takács, A. Tóth, B. Bogos, A. Varga, G. Rákhely, and K. L. Kovács, "Formate hydrogenlyase in the hyperthermophilic archaeon, Thermococcus litoralis." *BMC microbiology*, vol. 8, p. 88, Jan. 2008.

[354] C. Kirkpatrick, L. M. Maurer, N. E. Oyelakin, Y. N. Yoncheva, R. Maurer, and J. L. Slonczewski, "Acetate and formate stress: opposite responses in the proteome of Escherichia coli." *Journal of bacteriology*, vol. 183, no. 21, pp. 6466–77, Nov. 2001.

[355] T. Warnecke and R. T. Gill, "Organic acid toxicity, tolerance, and production in Escherichia coli biorefining applications." *Microbial cell factories*, vol. 4, p. 25, Aug. 2005.

[356] R. Rossmann, G. Sawers, and A. Böck, "Mechanism of regulation of the formate-hydrogenlyase pathway by oxygen, nitrate, and pH: definition of the formate regulon." *Molecular microbiology*, vol. 5, no. 11, pp. 2807–14, Nov. 1991.

[357] R. Böhm, M. Sauter, and A. Böck, "Nucleotide sequence and expression of an operon in Escherichia coli coding for formate hydrogenlyase components." *Molecular microbiology*, vol. 4, no. 2, pp. 231–43, Feb. 1990.

[358] M. Sauter, R. Böhm, and A. Böck, "Mutational analysis of the operon (hyc) determining hydrogenase 3 formation in Escherichia coli," *Molecular Microbiology*, vol. 6, no. 11, pp. 1523–1532, Jun. 1992.

[359] J. K. Rosentel, F. Healy, J. A. Maupin-Furlow, J. H. Lee, and K. T. Shanmugam, "Molybdate and regulation of mod (molybdate transport), fdhF, and hyc (formate hydrogenlyase) operons in Escherichia coli." *Journal of bacteriology*, vol. 177, no. 17, pp. 4857–64, Sep. 1995.

[360] W. T. Self, A. Hasona, and K. T. Shanmugam, "Expression and regulation of a silent operon, hyf, coding for hydrogenase 4 isoenzyme in Escherichia coli." *Journal of bacteriology*, vol. 186, no. 2, pp. 580–7, Jan. 2004.