Chapter 2
Metabolic Networks and Their Evolution

Andreas Wagner

Abstract Since the last decade of the twentieth century, systems biology has gained the ability to study the structure and function of genome-scale metabolic networks. These are systems of hundreds to thousands of chemical reactions that sustain life. Most of these reactions are catalyzed by enzymes which are encoded by genes. A metabolic network extracts chemical elements and energy from the environment, and converts them into forms that the organism can use. The function of a whole metabolic network constrains evolutionary changes in its parts. I will discuss here three classes of such changes, and how they are constrained by the function of the whole. These are the accumulation of amino acid changes in enzyme-coding genes, duplication of enzyme-coding genes, and changes in the regulation of enzymes. Conversely, evolutionary change in network parts can alter the function of the whole network. I will discuss here two such changes, namely the elimination of reactions from a metabolic network through loss of function mutations in enzyme-coding genes, and the addition of metabolic reactions, for example through mechanisms such as horizontal gene transfer. Reaction addition also provides a window into the evolution of metabolic innovations, the ability of a metabolism to sustain life on new sources of energy and of chemical elements.

1 Introduction

Metabolic networks are large systems of chemical reactions that serve two main purposes. The first is to convert sources of energy in the environment into forms of energy useful to an organism. The second is to synthesize small molecules needed for cell growth from sources of chemical elements—nutrients—in the

A. Wagner (✉)
Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Y27-J-54, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
e-mail: andreas.wagner@ieu.uzh.ch

O.S. Soyer (ed.), *Evolutionary Systems Biology*, Advances in Experimental Medicine and Biology 751, DOI 10.1007/978-1-4614-3567-9_2,

environment. These small molecules typically comprise the 20 amino acids found in proteins, DNA nucleotides, RNA nucleotides, lipids, and several enzyme cofactors. To fulfill the dual purposes of metabolism, the metabolic network of a free-living organism requires hundreds or more reactions, depending on the complexity of the environment they operate in [1, 2]. Most of these reactions are catalyzed by enzymes, which are encoded by genes. Together, they carry out the complex chemical transformations necessary to sustain life.

The structure, function, and evolution of metabolic networks have attracted a great amount of research interest for many decades [3–12]. Older work primarily focuses on small networks, comprising a handful of reactions, or on linear sequences of reactions. Experimental analysis of such small-scale systems involves classical biochemistry, including measurements of enzyme concentrations, enzyme activities, reaction rate constants, or metabolic fluxes—the rates at which enzymes convert substrates into products. Quantitative models of such small systems typically are kinetic models that use ordinary differential equations to study the changes in the concentrations of individual metabolites over time. The parameters of these equations include biochemically measurable quantities such as those I just mentioned [12].

With the rise to prominence of systems biology in the mid-1990s increasing attention started to focus on genome-scale metabolic systems. Such systems comprise not just few but hundreds or even thousands of reactions. That is, they comprise most or all reactions that take place in an organism's metabolism. Two technological and methodological advances made the analysis of such large metabolic networks feasible [2]. The first was that complete genome sequences were beginning to become available, first for the small genomes of prokaryotes, and subsequently for the much larger genomes of eukaryotes. Comprehensive information about the genes that an organism's genome harbors can provide unprecedented insights into the metabolic enzymes a genome encodes, and into the chemical reactions that an organism's metabolic network can catalyze. The second, closely related development was the ability to identify the complete or nearly complete set of chemical reactions that proceed in an organism's metabolism. This second development was facilitated by complete genome sequences, but it also required in-depth analyses of many years of accumulated biochemical literature in well-studied organisms, such as the bacterium *Escherichia coli* or the yeast *Saccharomyces cerevisiae*.

A quantitative understanding of genome-scale metabolic networks is difficult to achieve with as much detail as is possible for smaller networks. For example, it would be very difficult to estimate kinetic rate constants for hundreds of enzymes. It would also be very difficult to measure all metabolic fluxes in a large metabolic network: Methods using isotopic tracers and other tools [12–15] can measure the metabolic flux through many but not all reactions. They need to infer the fluxes through the remaining reactions from assumptions about the structure of a metabolic network. These technical difficulties put detailed kinetic models with measured parameters for all or even most reactions of a genome-scale metabolic network beyond our reach. Therefore, many approaches to understand the function of genome-scale metabolic networks focus on coarser-grained representations of

such networks. An especially prominent and fruitful approach in this area is called flux balance analysis (FBA), which requires only stoichiometric information about individual reactions, and which can predict the biosynthetic abilities of a network under some general assumptions (Box 1).

Box 1: Constraint-based modeling and flux balance analysis (FBA)

An important goal of systems biology is to predict a metabolic *phenotype*, the identity of the molecules that a metabolic network can synthesize, as well as their rate of synthesis, from a metabolic *genotype*, the set of enzymes encoded by a genome and their regulation. Experimental techniques have made great strides in this area [13–15], but they cannot (yet) determine phenotypes of genome-scale metabolic networks. Thus, computational approaches are indispensable for this purpose. One such approach is FBA, which is based on constraint-based modeling [16–18]. FBA has two objectives. First, it uses constraints given by reaction stoichiometry, reversibility, and maximal nutrient uptake rates of an organism to predict the metabolic fluxes that are *allowed* in a metabolic steady state, for all network reactions. Such a steady state would be attained by a cell population that is exposed to the same environment over extended periods of times, such as in a chemostat. Second, FBA then uses linear programming [19] to identify those allowed metabolic fluxes that maximize certain desired phenotypic properties, such as ATP or NADPH production, or the rate at which biomass with a known chemical composition is produced [10, 16, 17, 20–22, 24]. This latter rate is particularly important, because it is a proxy for the maximal rate at which cells can grow and divide. FBA is only one among several constraint-based techniques. Other examples include minimization of metabolic adjustment (MOMA), which aims to predict how metabolic networks react to loss of individual chemical reactions [25]. Extreme pathway analysis, elementary mode analysis, and the minimal metabolic behavior (MMB) approach decompose allowable fluxes into minimal sets analogous to basis vectors [26–32].

Aside from the steady-state assumption, the main limitation of most constraint-based methods is that they do not account for the regulation of enzymes, such as through transcriptional regulation. Efforts to incorporate regulation [33–36] are still hampered by limited empirical data. Nonetheless, constraint-based metabolic phenotype predictions are often in good agreement with experimental data [21, 25, 37]. Where they are not, microbial laboratory evolution experiments have shown that within a few hundred generations, a microbial strains' growth phenotype in a given environment can approach the FBA-predicted phenotype [38]. This means that regulatory constraints can be overcome on short evolutionary time scales.

(continued)

(continued)

To use constraint-based modeling for any one organism, the reactions in its metabolic network have to be known, as do its biomass composition, and nutrient uptake constraints. It is important to realize that the quality of phenotypic predictions obtained through constraint-based modeling depends critically on the accuracy and completeness of this information. Through a combination of manual curation and integration of genome-scale sequence data and functional genomics data, metabolic networks have been reconstructed for more than 40 organisms [39]. Such reconstructions are time-consuming and challenged by several factors, such as incorrect gene annotations, missing information on enzymes, elemental reaction imbalances, and incomplete information on reaction directionality, specificity, and thermodynamics. Increasingly, methods are being developed to overcome these and other obstacles [39–41].

I will focus here on genome-scale metabolic networks for two reasons. First, we have learned a substantial amount about their structure and their evolution in recent years. Second, they are the first systems that allow a comprehensive understanding of the relationship between a metabolic *genotype* (the DNA that encodes all metabolic enzymes an organism harbors) and a metabolic *phenotype*, the biosynthetic and energetic abilities of a metabolic network in a given environment. In other words, genome-scale metabolic networks are the first class of systems for which we can build a bridge between genotype and phenotype on the scale of entire organisms. Together, these two features make metabolic networks ideal study objects for the study of evolving biological systems, that is, for an Evolutionary Systems Biology.

A metabolic network is a *whole* comprised of many enzyme *parts.* To understand its structure and function, an evolutionary perspective is useful. The whole network constrains how its parts change over time. That is, natural selection on the function of the whole imposes constraints on the parts. Conversely, the parts and their changes influence the function of the whole. I will here discuss the evolution of metabolic networks from these two complementary perspectives. First, I will discuss different aspects of the evolution of network parts, and how the whole network constrains this evolution. Second, I will discuss changes in these parts that can change the function of the whole. This latter aspect is especially important, because it can teach us about how evolutionary change in metabolic networks can lead to new biosynthetic abilities. That is, it can teach us how metabolic innovations arise in evolution. Although it is useful to distinguish these two classes of influence—whole on parts, parts on whole—I note that they are not strictly separable. For example, when an altered part changes what the whole network is doing, the new network function may create new constraints on changes in its parts.

2 A Whole Constraining Its Parts

2.1 Constrained Evolution of Network Enzymes

There are three principal processes that are relevant to the evolution of a metabolic network's parts, that is, to the enzymes that catalyze its reactions. The first is the accumulation of changes—point mutations—in the DNA sequence of the genes encoding these enzymes. The second is the duplication of enzyme-coding genes. The third includes changes in the regulation of enzyme activities, for example through changes in the regulatory DNA sequences that help regulate the transcription of enzyme-coding genes. I will discuss the three processes in this order.

Not every point mutation that occurs in an enzyme-coding gene will survive and be passed on to subsequent generations. Mutations that destroy an essential enzyme's function and eliminate the metabolic flux through an essential reaction, for example, will be lethal to their carrier. The incidence of surviving point mutations in an enzyme-coding gene can be estimated by comparing the gene's DNA sequence to that of an orthologous gene—a gene with which it shared an ancestor in the past. Since the time of their common ancestor, two classes of point mutations may have occurred in either gene. The first are called *synonymous* or silent mutations. These are mutations that changed the DNA sequence of the gene, but due to the redundancy of the genetic code did not affect the amino acid sequence of the encoded protein. The second class of mutations is called *non-synonymous* or amino acid replacement mutations. These mutations did change the amino acid sequence of the encoded proteins, and may therefore also have changed the protein's function. The relative incidence of these two kinds of mutations, and the extent to which they have been preserved in evolution is commonly estimated through the fraction of synonymous changes that occurred at synonymous sites, often denoted as K_s, and through the fraction of non-synonymous changes per non-synonymous site K_a [42]. These measures take into consideration that different nucleotide sites in a gene have a different likelihood to undergo synonymous or non-synonymous change.

Silent mutations are subject to weaker selection than non-synonymous mutations, at least for most proteins and for most nucleotide sites in a gene [42, 43]. (Some silent mutations may cause changes in gene expression that are subject to selection.) For most enzyme-coding genes, one would therefore expect that K_a is smaller than K_s. In other words, the ratio K_a/K_s will be less than one, because fewer non-synonymous than silent changes are preserved in extant genes. The smaller this ratio is, the fewer amino acid replacement changes have been tolerated in the evolutionary history of a gene. In other words, a gene with a very small ratio K_a/K_s has experienced stronger selection in its history than a gene with a large ratio K_a/K_s.

Evolutionary constraints can depend on an enzyme's location in a genome-scale metabolic network, and on the metabolic flux through the enzyme. To render this assertion more precise I need to define what I mean by the location of an enzyme. One can represent a metabolic network as a graph, a mathematical object that consists of nodes (enzymes), and where any two nodes can be connected. In a

metabolic network, two enzymes are connected, if they share at least one metabolite as a substrate or as a product [44]. In the language of graph theory, two enzymes that are connected are also called neighbors. The number of enzymes that any one enzyme is connected to is called the degree or, more colloquially, the connectivity of the enzyme. Some enzymes are highly connected (they have high degree), whereas others are not highly connected. Many enzymes in central metabolic processes, such as central energy metabolism, are highly connected, whereas enzymes involved in peripheral pathways are often lowly connected. An enzyme's connectivity can be viewed as a measure of its position in the network, and of how central a role it might play in the network. (Other notions of position and centrality are also used in graph theory [45].)

The connectivity of an enzyme can influence its rate of evolution. For instance, in the metabolic network of the yeast *S. cerevisiae*, more highly connected enzymes evolve more slowly. That is, their ratio K_a/K_s is lower than for less connected enzymes [46]. Similar observations have been made in the fruit fly *Drosophila melanogaster* [47]. The likely reason comes from the effects of perturbations—for example caused by mutations—on the rate at which a highly connected enzyme catalyzes formation of its reaction product. Products of highly connected enzymes may be substrates for many other reactions. Perturbations in forming such products are thus more likely to be detrimental than perturbations in less highly connected enzymes. The association between enzyme connectivity and constraint, however, is not strong and may even be absent in some groups of organisms, such as mammals [48] and *E. coli* [49].

Analogous observations hold for enzymes with high metabolic flux. These are enzymes that turn over many molecules of substrate per unit time, and they are often involved in central metabolic processes. Specifically, enzymes with high flux tend to evolve more slowly [46]. They can tolerate fewer amino acid changes than enzymes with low flux. The reason becomes clear if one considers that most amino acid substitutions will reduce rather than increase an enzyme's activity, and thus reduce the metabolic flux that the enzyme can support. The observation that fewer amino acid changes can be tolerated in enzymes with high flux means that reduced flux in such enzymes is more likely to have adverse consequences for the organism, and that such enzymes are thus likely to be eliminated via natural selection. In other words, the biological function of a metabolic network constrains the evolution of its parts by point mutations. More precisely, it constrains the evolution of different parts to different extent. Parts with high flux and high connectivity are more constrained, and from this perspective, more important to the network's function, than parts with low flux.

In addition to the relationship between enzyme connectivity, flux, and constraints on enzyme evolution, several other observations have been made about the constrained evolution of metabolic genes. For instance, metabolic genes can be more constrained in their evolution than non-metabolic genes, at least in mammals and in *Drosophila* [47, 48]. In addition, different classes of enzymes are constrained to a different degree. For example, in *Drosophila*, enzymes that are involved in metabolizing xenobiotic substances are less constrained in their evolution than other

enzymes [47]. In mammals, enzymes expressed in the nucleus are more highly constrained than enzymes expressed in the cytoplasm [48].

In a minority of genes, the incidence of amino acid changing substitutions may actually exceed that of silent substitutions. In these genes, the ratio K_a/K_s may exceed 1. Patterns like this indicate the action of positive selection, that is, one or more amino acid changes were favored by selection, and have swept through an evolving population, which can explain the elevated rate of amino acid change. A ratio of K_a/K_s that exceeds 1 indicates beneficial functional changes in a protein. Unfortunately, without detailed and laborious biochemical analyses it can be difficult to understand why a change is beneficial.

In general, only a minority of genes is subject to positive selection at any one time. In the genus *Drosophila*, for example, fewer than 10% of enzyme-coding genes appear to be under positive selection [47]. In many of these genes, the reason for their functional change has not been characterized, but exceptions exist. For example, the gene encoding the enzyme glutathione-S-transferase is under positive selection. The likely reason is that the changes in glutathione-S-transferase help improve the enzyme's ability to detoxify pesticides such as DDT, and thus help flies survive these pesticides [50].

2.2 Gene Duplication

The second major process that can affect metabolic network parts is the duplication of enzyme-coding genes. Gene duplication is a ubiquitous process in the evolution of most genomes. For example, as many as half of the genes in the human genome have a duplicate [51]. Gene duplications arise as by-products of DNA recombination and DNA repair processes that sometimes duplicate stretches of an organism's DNA. The duplicated stretches can be very short, comprising only a few nucleotides, or they can be very long, comprising large segments of chromosome, entire chromosomes, or even the entire genome. If any duplicated stretch of DNA includes at least one gene, a gene duplication has occurred. Most duplicate genes are eliminated from a genome shortly after the duplication [52]. However, a small fraction of duplicates is usually preserved, indicating that their duplication either did no harm or was favored by selection. Over time duplicates may preserve a similar function, they may acquire specialized functions, or they may evolve completely new functions [53, 54].

If the functional demands on a metabolic network were irrelevant for duplications in its enzyme-coding genes, then the incidence of preserved duplications should be the same for all metabolic genes. This, however, is not the case, indicating that network structure and function influences gene duplication patterns. For example, in mammalian metabolic networks [55], duplications are preferentially preserved in genes whose products transport metabolites into cells. In cattle, genes encoding metabolic enzymes that are involved in milk production are more likely to have duplicates, indicating that natural selection may have influenced duplication patterns

in these genes [55]. Even adaptive genetic changes in laboratory evolution experiments, that is, changes that occur on short evolutionary time-scales, can be mediated by gene duplications. For instance, in populations of yeast cells cultivated under conditions where glucose limits the rate of cell growth, duplications in high affinity hexose transporter genes accumulate [56]. Such duplications allow yeast cells to scavenge scarce glucose from the environment.

The metabolic significance of gene duplications is that they can increase the level of an enzyme's expression. Enzymes that are products of duplicated genes may occur in higher concentrations in the cell, and they may therefore support greater metabolic flux through them. One might therefore predict that enzymes with high metabolic flux should often be the product of duplicate genes. This prediction is borne out by existing observations. For example, high-flux enzymes in the metabolism of the yeast *S. cerevisiae* are more often encoded by duplicate genes than low-flux enzymes [46]. Thus, here again a function of the whole network constrains the evolution of its parts, in this case through gene duplication. Specifically, the preservation of gene duplications is favored in enzyme-coding genes whose protein products catalyze high-flux reactions. Many such genes occur in central metabolism.

An extreme form of duplication is the duplication of an entire genome. After such a genome duplication, most duplicated genes typically get lost over time, and only a small fraction of them remain. The remaining fraction may not comprise a random subset of metabolic genes. For example, it has been shown that the enzyme-coding genes preserved in duplicate after an ancient genome duplication in *S. cerevisiae* preferentially encode glycolytic enzymes. This preferential preservation allows a higher flux through glycolysis relative to other parts of yeast's metabolism, because it increases the total amount of glycolytic enzymes relative to other enzymes. It allows yeast cells to ferment glucose more effectively, and it may have helped yeast cells survive in a glucose-rich environment [57].

Taken together, these observations suggest that the constraint that a whole metabolic network imposes on the duplication of its parts arises through the increased enzyme expression that such duplications cause. If increased expression of an enzyme is advantageous, for example because it allows greater flux through a metabolic reaction, duplications in the gene encoding the enzyme may be preserved preferentially.

2.3 Gene Regulation

The third and final major process that can affect metabolic network parts is the evolution of their regulation. It is the most difficult process to study, because regulation can have many facets. Enzymes can be regulated on the level of their RNA expression, their protein expression, their biochemical activity, for example through phosphorylation, and in many other ways. Studies of how individual enzymes are regulated have a long history [12]. However, information about such

small-scale regulatory changes has not yet given rise to a principled understanding of how the regulation of *all* enzymes in a metabolic network evolves. Only this much is certain: Regulation is extremely malleable and can change on short evolutionary time-scales for many enzymes. For example, laboratory evolution experiments in which *E. coli* cells adapt evolutionarily to new nutrients show that such change can occur in a few hundred generations, can alter the transcription of many genes, and can occur differently in parallel experiments [58]. Regulatory changes like those observed in laboratory evolution experiments reflect changes in the demands that a whole metabolic network operating in a new environment places on the function of its parts.

In closing this section, it is worth mentioning that all three processes—gene sequence evolution, gene duplication, and regulatory evolution—usually occur simultaneously. For example, several enzyme-coding genes in the yeast tricarboxylic acid (TCA) cycle have undergone duplication, and have subsequently diverged in their sequence and expression, which reflects their adaptation to operate in different cell compartments [59].

3 Parts Transforming the Whole

I will next discuss changes that affect the number and identity of the chemical reactions in a metabolic network. These are *qualitative* changes that can alter a network's biosynthetic abilities profoundly. As opposed to the *quantitative* changes that I discussed so far, which typically just reduce or increase the rate at which a network can synthesize biomass in a given environment, such qualitative changes are changes in parts that can transform the whole network. They may eliminate the network's ability to sustain life in a given environment, or they may allow the network to sustain life in new chemical environments. The latter kind of change is an especially worthy subject of study, because it speaks to the fundamental evolutionary question of how new traits arise in evolution.

The reaction complements of metabolic networks can vary greatly among organisms. For example, metabolic annotations available for more than 200 completely sequenced bacterial genomes suggest that metabolic networks can differ in more than 50% of their reactions [60]. Even different strains of the same organism, such as *E. coli*, may differ in more than 100 metabolic reactions [61].

It is often useful to think of a metabolism as being partitioned into two major parts, a core and a periphery. Core metabolism comprises processes central to life, such as glycolysis, the TCA cycle, or the pentose phosphate shunt. The periphery includes reactions that are needed to metabolize specific sources of chemical elements. It converts these elements into compounds that the core metabolism can process further. The periphery also includes secondary metabolism, which synthesizes molecules such as alkaloids or pigments that are not absolutely essential for life, but that serve other important functions, such as protection against a hostile environment.

Core metabolism is held to be highly optimized in different ways [62, 63]. For example, it has been suggested that among a number of alternative "designs" of the TCA cycle, the structure of the cycle realized in nature uses the smallest number of chemical transformations, and produces the highest yield in ATP [63]. However, even such central parts of metabolism can vary among different organisms. For example, analysis of completely sequenced bacterial genomes suggests that the TCA cycle may be incomplete in multiple species [64]. Although changes in core metabolism do occur, variation in the reaction complement of a metabolic network tends to be more frequent in the periphery of metabolism.

3.1 Reaction Deletions

The first of two major kinds of qualitative changes in a metabolic network is the elimination of reactions. Such elimination can occur through loss of function mutations in enzyme-coding genes. It is often observed for organisms living in environments that undergo little change, such as endoparasitic or endosymbiotic single-celled organisms, which live inside other organisms. Examples include *Buchnera aphidicola*, an endosymbiotic relative of *E. coli*, which lives inside the cells of aphids [65, 66]. *Buchnera* provides its host with essential amino acids in an association that has persisted for many million years [66]. During this time the genome of *Buchnera* has lost many genes, and its metabolic network has lost many chemical reactions [67]. For example, while the metabolic network of *E. coli* has more than 900 reactions [68], that of *Buchnera* has merely 263 metabolic reactions [67]. *E. coli* is a metabolic generalist whose metabolic network can sustain life on dozens of different carbon sources in otherwise minimal chemical environments. The metabolic network of *Buchnera* has lost this versatility, because it is no longer needed. Similar reductions in genome sizes and metabolic networks have been observed in other organisms, such as the human pathogen *Mycoplasma pneumonia*, whose metabolic network comprises only 189 reactions [69]. More generally, a reduction in network size and versatility to live in multiple environments would be expected under prolonged exposure to the same environment [70, 71].

Flux balance analysis (FBA, Box 1) can predict the spectrum of molecules that can be synthesized by a given metabolic network from a set of nutrients in the environment. FBA is also useful to reconstruct the evolutionary trajectory that can transform a complex metabolic network like that of *E. coli* into the much simpler network of its relative *Buchnera* through a sequence of mutations that eliminate enzyme-coding genes and reactions from a metabolic network [72,73]. For example, one can predict the reaction complement of *B. aphidicola* with about 80% accuracy from knowledge about the *E. coli* metabolic network, and about the environment in which *Buchnera* lives [73].

3.2 Reaction Additions

The second major class of qualitative changes to a metabolic network is the addition of chemical reactions. There are several mechanisms by which reactions can get added to a network. For example, after a duplication of an enzyme-coding gene, one of the duplicates may preserve its enzymatic function, whereas the other may evolve a new catalytic function. Mechanisms like this require the origin of new catalytic functions in enzymes. Other mechanisms do not. Consider horizontal gene transfer. Through this mechanism, new enzyme-coding genes can be imported into a genome from the genomes of other organisms. Through horizontal gene transfer reactions can get added to a metabolic network without the need to evolve new enzymatic activities from scratch. It is thus an especially powerful way of evolving new metabolic traits. I will briefly discuss its incidence before returning to metabolic network evolution.

Horizontal gene transfer occurs both in prokaryotes and eukaryotes, but it is much more prevalent in prokaryotes. It can change genome organization on short evolutionary time-scales [74–82]. For example, DNA is transferred into the *E. coli* genome at a rate of 64 kilobase pairs per million years [83]. With an average gene length of approximately 1 kilobase pairs [84], this rate amounts to the transfer of 64 genes per million years. Even closely related *E. coli* strains can differ by more than one megabase pair of DNA [77], or more than 20% of their genome, and they may have experienced of the order of 100 gene additions through horizontal transfer relative to other strains [74]. Because some 30% of *E. coli* genes have metabolic functions [1, 84], the effect of such horizontal gene transfer on metabolism is surely profound. The addition of new DNA can be compensated by the deletion of other DNA, and many newly added genes reside in the genome only for short amounts of time [75, 83]. Gene turnover in microbial genomes can thus be very high.

A recent study used FBA (Box 1), as well as information about horizontal gene transfer into the *E. coli* genome to examine evolutionary changes in *E. coli* metabolism [75]. It concluded that metabolic genes that are preserved after horizontal transfer are often responsible for metabolic reactions that transport and metabolize nutrients. Such genes may be preserved, because they allow the organism to survive in specific nutrient environments. The relevant reactions are located at the periphery of metabolism and not at its core. The study also showed that gene duplication played a relatively small role in the evolution of *E. coli* metabolism, at least in the last hundred million years [75]. This observation underscores the importance of horizontal gene transfer in metabolic evolution. Horizontal gene transfer may be one of the reasons why prokaryotes are masters of metabolic innovation. They have evolved the ability to survive on an immensely broad spectrum of nutrients, including sources of carbon such as crude oil, hydrogen, methane, toxic xenobiotics, and antibiotics [85–91].

4 A Systematic Analysis of Metabolic Innovation

New phenotypes that provide a qualitative advantage to an organism's ability to survive or reproduce are also known as evolutionary innovations. The ability to sustain life on a new nutrient can be considered an evolutionary innovation in metabolism. We know many evolutionary innovations (metabolic and others). They are fascinating and well-studied examples of natural history [92]. But beyond the well-worn idea that innovations require a combination of mutation and natural selection, we know little about the principles underlying their origins. To identify such principles requires that one can study the relationship between genotype and phenotype systematically, not just for one genotype and one phenotype, but for many genotypes and many phenotypes. To determine phenotypes of many organisms is still difficult, time consuming, and an area of active methods development [93]. Thus, systems where one can *predict* phenotype from genotype are currently the best starting points for understanding principles of innovation. Metabolism is one such system, because tools such as FBA (Box 1) can help us understand its genotype–phenotype relationship. In the next section, I will summarize recent work that has advanced our understanding of metabolic innovations.

To appreciate the key difficulties in understanding the origins of metabolic innovations, I first need to make the notion of metabolic genotype and phenotype more precise (Fig. 2.1). An organism's metabolic *genotype* is the part of the organism's genome that encodes metabolic enzymes. However, it is often more expedient to represent this genotype more compactly, such as through the presence or absence of specific enzyme-catalyzed reactions in the network [95]. The current known "universe" of metabolic reactions comprises more than 5,000 such reactions, each of which can be present or absent in the metabolic network of any one organism. This means that there are more than 2^{5000} possible metabolic networks [95,96], distinguished from one another through the presence or absence of different reactions (enzyme-coding genes). Together, they form a vast collection, a *space* of metabolic genotypes. This space is much larger than the number of metabolic networks that could have existed on earth since life's origin.

In this space, one can define a *distance* between metabolic genotypes as the fraction of metabolic reactions in which these genotypes differ. Two genotypes (metabolic networks) would differ maximally if they did not share a single reaction. Two genotypes are *neighbors* in this space if they differ minimally, that is, in only one metabolic reaction. The *neighborhood* of a genotype G comprises all of its neighbors, more than 5,000 metabolic networks, each of which differing from G in one reaction. Metabolic genotype space is a high dimensional space with many counterintuitive properties, whose structure is akin to that of hypercubes—cubes in multidimensional spaces [97,98].

To classify metabolic *phenotypes*, it is expedient to focus on metabolism's central task, the ability to sustain life—to synthesize all biomass molecules—in different chemical environments [95]. For example, if one focuses on carbon metabolism, one can ask which molecules can serve as sole carbon and energy sources for a

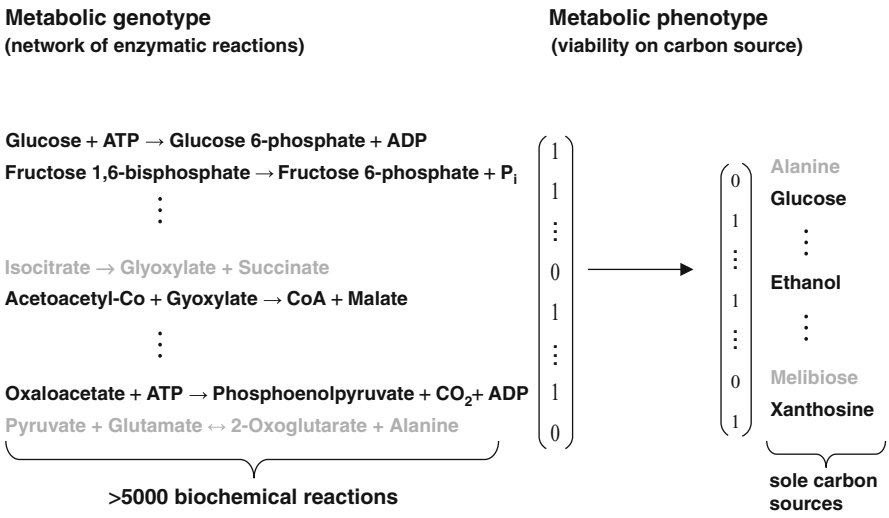

Fig. 2.1 *Metabolic genotypes and phenotypes*. The metabolic genotype of a genome-scale metabolic network can be represented in discrete form as a binary string, each of whose entries corresponds to one biochemical reaction in a "universe" of known reactions. Individual entries indicate the presence ("1," *black type* in stoichiometric equation) and absence ("0," *gray type*) of an enzyme-coding gene whose product catalyzes the respective reaction. Metabolic phenotypes can be represented by a binary string whose entries correspond to individual carbon sources. The string contains a "1" for every carbon source (*black type*), for which a metabolic network can synthesize all major biomass molecules, if this source is the only available carbon source. Flux balance analysis can be used to predict metabolic phenotypes from metabolic genotypes. Figure and caption adapted from [94]. Used by permission from Oxford University Press

metabolic network. To represent such phenotypes systematically, one can use some number of common carbon sources, say 100 different molecules, and write these as a list (Fig. 2.1, right panel). A metabolic phenotype can then be represented as a binary string, where one writes a one next to a carbon source in the list, if the network can sustain life on it, and a zero if it cannot. Note that for 100 carbon sources, there is already an astronomical number of 2^{100} possible metabolic phenotypes, each of them encapsulating viability in a different spectrum of chemical environments. Analogous classifications are possible for sources of other elements [71]. FBA and constraint-based modeling (Box 1) allow us to compute metabolic phenotypes from metabolic genotypes.

All evolution occurs in populations of organisms. We can envision such a population, each of whose members may have a different metabolic genotype, as a collection of points in metabolic genotype space. Such a population explores metabolic genotype space through mutation (changes in enzyme-coding genes that add or delete reactions from a network) and natural selection that preserves well-adapted phenotypes. Suppose that individuals in this population have a metabolic phenotype that is well adapted to a population's current environment. When that environment changes, a new phenotype may become superior to the old phenotype.

For example, individuals with the old phenotype may not have been able to thrive on some carbon source, say ethanol. In the new environment ethanol may be an abundant carbon source. It would be advantageous if organisms in the population could "find" genotypes with this phenotype, and thus begin to use ethanol as a sole carbon source.

The following considerations illustrate two major difficulties with finding such novel and superior metabolic phenotypes through a blind evolutionary search conducted by a population in the vast metabolic genotype space. First, imagine that only one or a few metabolic genotypes in this space have the superior phenotype. Because this space is so large, it would be difficult or impossible to find these genotypes in realistic amounts of time. Second, during this search, individuals in a population have to preserve their old phenotype, which allows them to survive on existing nutrients. If any mutation abolished this ability, its carrier would perish. In other words, while the population explores the vast genotype space for new and potentially useful phenotypes, it needs to preserve its old phenotype. It needs to conserve the old while exploring the new.

These problems may seem difficult to overcome. However, systematic analyses of metabolic genotype space, conducted by sampling thousands of metabolic networks from this space and by computing their phenotypes, reveal two major features of this space that help overcome them [71, 94, 96].

The first feature is that there are not few but hyperastronomically many genotypes with a given metabolic phenotype. For example, there are more than 10^{800} metabolic networks with 2,000 reactions that can synthesize all the small biomass molecules of the bacterium *E. coli* using glucose as the sole carbon source. What is more, these metabolic genotypes are connected in metabolic genotype space in the following sense [71, 99]. One can step from one metabolic genotype to its neighbor, to the neighbor's neighbor, and so forth, without changing the metabolic phenotype, until one has traversed a large fraction of the space. Specifically, metabolic networks with the same phenotype may share as little as 30% of their reactions [71]. The reactions they do share form part of core metabolism. Most other reactions can vary.

Figure 2.2 illustrates schematically how one can envision the organization of metabolic genotypes with any one particular phenotype. The left-hand panel shows a large rectangle which stands for genotype space. Inscribed in this rectangle is a single open circle, intended to illustrate that a metabolic genotype (a metabolic network) is a single point in this space. The right-hand side shows an identical rectangle, but with many open inscribed circles. Each of them corresponds to a single metabolic genotype with the same phenotype *P*. Two genotypes (circles) are connected by a straight line if they are neighbors. The panel illustrates that metabolic networks with the same phenotype form a vast network of networks—a *genotype network*—that reaches far through genotype space. I note that a two-dimensional image like this just provides a crude visual crutch. It allows us merely to get a modicum of visual intuition about the organization of a space that is vast and that has many dimensions.

Large genotype networks that extend far through metabolic genotype space are not a peculiarity of specific metabolic phenotypes. They exist for a broad range of

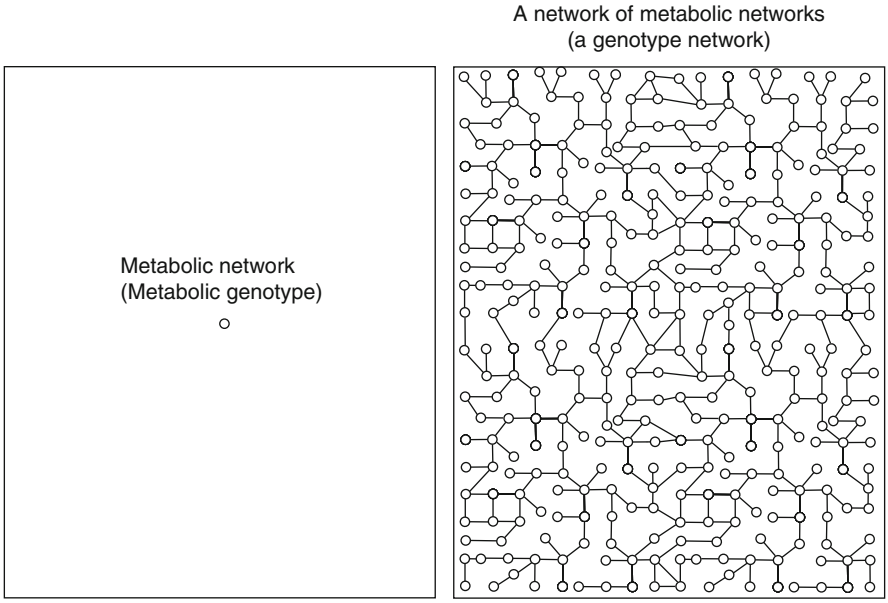

Fig. 2.2 *Genotype networks*. The *large rectangle* in each panel stands for genotype space. The left panel shows a single open circle inscribed in this space, which stands for a hypothetical metabolic genotype, that is, a metabolic network with a specific set of enzyme-catalyzed reactions and some phenotype P. The *right panel* shows a large collection of *circles*, each corresponding to a metabolic genotype with the same phenotype P. Two circles are linked by a straight line if they are neighbors, that is, if the metabolic networks that they represent differ in a single chemical reaction. The linked circles form a large network of metabolic genotypes—a genotype network. See text for details

phenotypes able to sustain life on many different sole carbon sources, on multiple carbon sources, as well as on sources of other chemical elements [71, 94, 96]. That is, each such phenotype has an associated genotype network that is typically large and reaches far through genotype space. Genotype networks are generic features of metabolic genotype space. Their existence is a consequence of their robustness to genetic change, which in turn is linked with life in changing environments [70, 100–105].

A second important feature regards the neighborhoods of different genotypes with the same phenotype. Consider two genotypes G_1 and G_2 that have identical phenotypes P, and all genotypes in the two neighborhoods of these two genotypes. Using tools such as FBA, one can examine the genotypes in these neighborhoods one by one, and establish a list $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ of all phenotypes different from P in the neighborhoods of G_1 and G_2, respectively. One can then ask whether the new phenotypes in $\boldsymbol{P}_1$ are mostly the same as the new phenotypes in $\boldsymbol{P}_2$, or if they are very different. Here is the answer: Even if G_1 and G_2 differ only modestly in the reactions that they contain, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ typically contain mostly different new phenotypes. In other words, the spectrum of new phenotypes in the neighborhood

of one metabolic genotype is typically not identical to that in the neighborhood of another genotype. In other words, different neighborhoods of metabolic networks—even networks with the same phenotype—contain different novel phenotypes. The extent of this diversity is not very sensitive to specific phenotypes *P* [71,94,96]. It is another generic feature of metabolic genotype space.

Figure 2.3 illustrates these observations. Like the right panel of Fig. 2.2, this figure also shows a hypothetical genotype network (open circles) whose members have some phenotype *P*. In addition, it shows multiple colored circles, each of which stands for a genotype with a phenotype different from *P*. Each color corresponds to a different phenotype. Each of these genotypes are neighbors of a genotype on the genotype network. The figure also shows two dashed circles that circumscribe the neighborhood of two different genotypes in the circles' center. The two circles contain different new phenotypes (colors), illustrating the principle I just mentioned. Note again that this figure is a highly simplified sketch of a high-dimensional genotype space. For example, metabolic genotypes have thousands of neighbors, not just the few neighbors shown here. In addition, the genotypes with new phenotypes (colors) generally also form large genotype networks, which are not shown here.

In sum, two generic properties characterize metabolic genotype space. The first is that genotypes with the same phenotype form large and far-reaching genotype networks. The second is that the neighborhoods of different genotypes on the same genotype network typically contain different metabolic phenotypes. Together, these features facilitate the evolutionary search of novel phenotypes through mutation and natural selection in genotype space. First, the fact that there are astronomically many and not few genotypes with the same phenotype facilitates the encounter of any one genotype with this phenotype. Second, genotype networks with their diverse neighborhoods facilitate the exploration of many novel phenotypes while preserving existing phenotypes. The reason is that genotype networks allow metabolic genotypes to be changed through addition and elimination of reactions, while preserving their phenotype. During such change, individuals in a population of evolving organisms can explore ever-changing neighborhoods of genotype space, which allows them to access a broad spectrum of novel phenotypes, many more than if genotype networks did not exist.

I note that the features of metabolic genotype space that I described here may depend on the particular class of phenotype one studies. However, they are probably widespread, because they also exist in multiple other classes of systems, including regulatory circuits, proteins, and RNA [106–110]. In general, they occur in systems whose genotype–phenotype relationship is such that more genotypes than phenotypes exist, and where phenotypes are to some extent robust to changes in genotype [94].

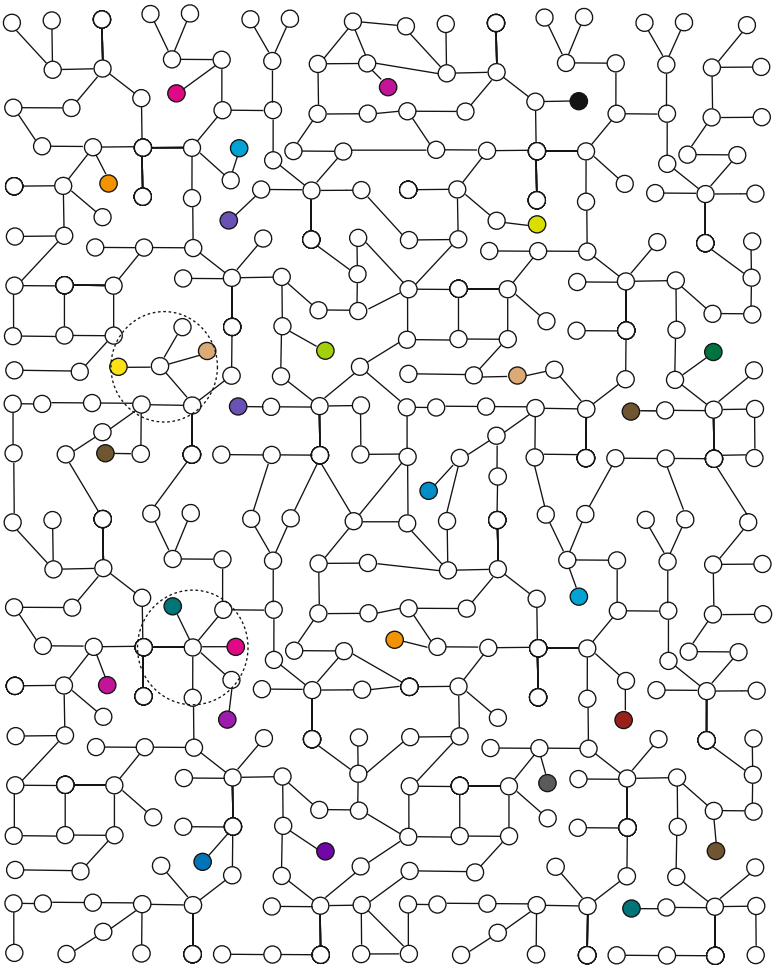

Fig. 2.3 *Diverse genotypic neighborhoods in genotype space.* As in Fig. 2.2, the large collection of *open circles* stands for a hypothetical genotype network, that is, a large connected set of metabolic genotypes with the same phenotype. Circles in different colors correspond to genotypes that are neighbors of a genotype on this genotype network, but that have different phenotypes. Each color stands for a different phenotype. Each of the two large *dotted circles* stands for the neighborhood of a genotype, which is at the center of the circle. The two neighborhoods each contain two genotypes with new phenotypes (*colored circles*). However, the identity of these phenotypes differ between the two neighborhoods, as indicated by their different colors (*yellow* and *beige* in one neighborhood, *blue* and *red* in the other). See text for details. Adapted from [94]. Used by permission from Oxford University Press

5 Conclusions and Future Challenges

Theodosius Dobzhansky's old adage that "nothing in biology makes sense except in the light of evolution" [92] also applies to metabolism. We will understand the structure of genome-scale metabolic networks to the extent that we will understand their evolution. Our efforts in this area are just beginning. In recent years, our ability to reconstruct evolutionary processes in the laboratory has made great strides, as have efforts to determining different aspects of metabolic phenotypes. Many of the studies I discussed here are based on comparative analyses of metabolic networks, aided by computational predictions of metabolic phenotypes. In the foreseeable future, it may become possible to integrate the observations I discussed here with experimental observations. Doing so may lead to a more comprehensive understanding of how a whole metabolic network influences the evolution of its parts, and how these parts influenced the whole.

The ability to predict metabolic phenotype from metabolic genotype has opened completely new avenues for a systematic understanding of metabolic innovation. It allows us to study metabolic innovations not one by one, as case studies in natural history, but systematically, as part of a metabolic genotype space that encapsulates all *possible* metabolisms. Such a systematic approach allows us to ask whether fundamental principles exist that facilitate metabolic innovations. Here also, we are at a beginning. Genotype networks and their diverse neighborhood are two features of genotype space that facilitate innovation, but this space may harbor many other secrets. The tools of Evolutionary Systems Biology will allow us to uncover these secrets.

References

1. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3. doi:121.10.1038/msb4100155
2. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2):129–143. doi:10.1038/nrmicro1949
3. Holms WH (1986) The central metabolic pathways of Escherischia coli: relationship between flux and control at a branch point, efficiency of conversion to biomass and excretion of acetate. Current Topics Cell Regul 28:69–105
4. Dykhuizen DE, Dean AM, Hartl DL (1987) Metabolic flux and fitness. Genetics 115(#1):25–31
5. Keightley PD, Kacser H (1987) Dominance, pleiotropy and metabolic structure. Genetics 117(#2):319–329
6. Joshi A, Palsson BO (1989) Metabolic dynamics in the human red-cell.1. A comprehensive kinetic model. J Theor Biol 141(4):515–528
7. Hofmeyr J-HS (1991) Control pattern analysis of metabolic pathways: flux and concentration control in linear pathways. Eur J Biochem 275:253–258

8. Varma A, Palsson BO (1993) Metabolic capabilities of Escherichia coli. Synthesis of biosynthetic precursors and cofactors. J Theor Biol 165:477–502
9. Veech RL, Fell DA (1996) Distribution control of metabolic flux. Cell Biochem Funct 14(#4):229–236
10. Bonarius HPJ, Schmid G, Tramper J (1997) Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. Trends Biotechnol 15(8):308–314
11. Thomas S, Fell DA (1998) A control analysis exploration of the role of ATP utilisation in glycolytic-flux control and glycolytic-metabolite-concentration regulation. Eur J Biochem 258(#3):956–967
12. Fell D (1997) Understanding the control of metabolism. Portland Press, Miami
13. Fischer E, Sauer U (2005) Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. Nat Genet 37(6):636–640
14. Blank LM, Lehmbeck F, Sauer U (2005) Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. Fems Yeast Res 5(6–7):545–558
15. Blank LM, Kuepfer L, Sauer U (2005) Large-scale C-13-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol 6(6):R49
16. Price N, Reed J, Palsson B (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886–897
17. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2(3):727–738. doi:10.1038/nprot.2007.99
18. Heinrich R, Schuster S (1996) The regulation of cellular systems. Chapman and Hall, New York
19. Cormen TH, Leiserson CE, Rivest RL, Stein C (2005) Introduction to algorithms. 2nd edn. MIT Press, Cambridge, MA
20. Forster J, Famili I, Fu P, Palsson B, Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res 13:244–253
21. Edwards JS, Palsson BO (2000) The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proc Natal Acad Sci USA 97(10):5528–5533
22. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol 3. doi:119.10.1038/msb4100162
23. Savinell JM, Palsson BO (1992) Network analysis of intermediary metabolism using linear optimization.1. development of mathematical formalism. J Theor Biol 154(4):421–454
24. Fell DA, Small JR (1986) Fat synthesis in adipose-tissue - an examination of stoichiometric constraints. Biochem J 238(3):781–786
25. Segre D, Vitkup D, Church G (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 99:15112–15117
26. Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004) Comparison of network-based pathway analysis methods. Trends in Biotechnology 22(8):400–405. doi:10.1016/j.tibtech.2004.06.010
27. Palsson BO, Price ND, Papin JA (2003) Development of network-based pathway definitions: the need to analyze real metabolic networks. Trends Biotechnol 21 (5):195–198. doi:10.1016/s0167–7799(03)00080–5
28. Papin JA, Price ND, Palsson BO (2002) Extreme pathway lengths and reaction participation in genome-scale metabolic networks. Genome Res 12(12):1889–1900. doi:10.1101/gr.327702
29. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. Nature 420(6912):190–193
30. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18(3):326–332
31. Klamt S, Stelling J (2003) Two approaches for metabolic pathway analysis? Trends Biotechnol 21(2):64–69

32. Larhlimi A, Bockmayr A (2006) A new constraint-based description of the steady-state flux cone of metabolic networks. In: Workshop on Networks in Computational Biology, Ankara, TURKEY, Sep 10–12 2006. pp. 2257–2266. doi:10.1016/j.dam.2008.06.039
33. Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. Plos Comput Biol 4(5). doi:e1000082.10.1371/journal.pcbi.1000082
34. Herrgard MJ, Fong SS, Palsson BO (2006) Identification of genome-scale metabolic network models using experimentally measured flux profiles. Plos Comput Biol 2(7):676–686. doi:e72q.10.1371/journal.pcbi.0020072
35. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429(6987):92–96
36. Herrgard MJ, Lee BS, Portnoy V, Palsson BO (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Genome Res 16(5):627–635. doi:10.1101/gr.4083206
37. Forster J, Famili I, Palsson BO, Nielsen J (2003) Large-scale evaluation of in-silico gene deletions in Saccharomyces cerevisiae. Omics 7:193–202
38. Fong SS, Palsson BO (2004) Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet 36(10):1056–1058
39. Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26(6):659–667. doi:10.1038/nbt1401
40. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophys J 92(5):1792–1805. doi:10.1529/biophysj.106.093138
41. Mavrovouniotis ML (1991) Estimation of standard Gibbs energy changes of biotransformations. J Biol Chem 266(22):14440–14445
42. Li W-H (1997) Molecular evolution. Sinauer, Massachusetts
43. Parmley JL, Hurst LD (2007) How do synonymous mutations affect fitness? Bioessays 29(6):515–519. doi:10.1002/bies.20592
44. Wagner A, Fell D (2001) The small world inside large metabolic networks. Proc Roy Soc London Ser B 280:1803–1810
45. Newman MEJ (2003) The structure and function of complex networks. Siam Review 45(2):167–256
46. Vitkup D, Kharchenko P, Wagner A (2006) Influence of metabolic network structure and function on enzyme evolution. Genome Biol 7(5). doi:R3910.1186/gb-2006–7–5-r39
47. Greenberg AJ, Stockwell SR, Clark AG (2008) Evolutionary constraint and adaptation in the metabolic network of Drosophila. Mol Biol Evol 25(12):2537–2546. doi:10.1093/molbev/msn205
48. Hudson CM, Conant GC (2011) Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. BMC Evolutionary Biol 11. doi:89.10.1186/1471–2148–11–89
49. Hahn M, Conant GC, Wagner A (2004) Molecular evolution in large genetic networks: does connectivity equal importance? J Mol Evol 58:203–211
50. Low WY, Ng HL, Morton CJ, Parker MW, Batterham P, Robin C (2007) Molecular evolution of glutathione S-transferases in the genus drosophila. Genetics 177(3):1363–1375. doi:10.1534/genetics.107.075838
51. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei YD, Li ZY, Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK,

Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun JT, Wang ZY, Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH, Yao A, Ye J, Zhan M, Zhang WQ, Zhang HY, Zhao Q, Zheng LS, Zhong F, Zhong WY, Zhu SPC, Zhao SY, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An HJ, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi HY, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu XJ, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen MY, Wu D, Wu M, Xia A, Zandieh A, Zhu XH (2001) The sequence of the human genome. Science 291(5507):1304–1351

52. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155
53. Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Ann Rev Genet 38:615–643
54. Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9(12):938–950. doi:10.1038/nrg2482
55. Bekaert M, Conant GC (2011) Copy number alterations among mammalian enzymes cluster in the metabolic network. Molecular Biology and Evolution 28(2):1111–1121. doi:10.1093/molbev/msq296
56. Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D (2002) Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc Natal Acad Sci USA 99(25):16144–16149
57. van Hoek MJA, Hogeweg P (2009) Metabolic adaptation after whole genome duplication. Mol Biol Evol 26(11):2441–2453. doi:10.1093/molbev/msp160
58. Fong SS, Joyce AR, Palsson BO (2005) Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res 15(10):1365–1372. doi:10.1101/gr.3832305
59. McAlister-Henn L, Small W (1997) Molecular genetics of yeast TCA cycle isozymes. Prog Nucleic Acid Res Mol Biol 57:317–339
60. Wagner A (2009) Evolutionary constraints permeate large metabolic networks. BMC Evolutionary Biol 9. doi:231.10.1186/1471–2148–9–231
61. Vieira G, Sabarly V, Bourguignon PY, Durot M, Le Fevre F, Mornico D, Vallenet D, Bouvet O, Denamur E, Schachter V, Medigue C (2011) Core and panmetabolism in Escherichia coli. J Bacteriol 193(6):1461–1472. doi:10.1128/jb.01192–10
62. Noor E, Eden E, Milo R, Alon U (2010) Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. Mol Cell 39(5):809–820. doi:10.1016/j.molcel.2010.08.031
63. Melendez-Hevia E, Waddell TG, Cascante M (1996) The puzzle of the Krebs citric-acid cycle: assembling the pieces of chemically feasible reactions; and opportunism in the design of metabolic pathways during evolution. J Mol Evol 43(#3):293–303

64. Huynen MA, Dandekar T, Bork P (1999) Variation and evolution of the citric acid cycle: a genomic perspective. Trends Microbiol 7(7):281–291
65. Moran NA, Wernegreen JJ (2000) Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol Evol 15(8):321–326
66. Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. Ann Rev Genet 42:165–190. doi:10.1146/annurev.genet.41.110306.130119
67. Thomas GH, Zucker J, MacDonald SJ, Sorokin A, Goryanin I, Douglas AE (2009) A fragile metabolic network adapted for cooperation in the symbiotic bacterium *Buchnera aphidicola*. BMC Sys Biol 3:24. doi:10.1186/1752–0509–3–24
68. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol 4(9):R54
69. Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen WH, Wodke JAH, Guell M, Martinez S, Bourgeois R, Kuhner S, Raineri E, Letunic I, Kalinina OV, Rode M, Herrmann R, Gutierrez-Gallego R, Russell RB, Gavin AC, Bork P, Serrano L (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326(5957):1263–1268. doi:10.1126/science.1177263
70. Soyer OS, Pfeiffer T (2010) Evolution under fluctuating environments explains observed robustness in metabolic networks. PLoS Comp Biol 6(8). doi:e1000907.10.1371/journal.pcbi.1000907
71. Rodrigues JF, Wagner A (2011) Genotype networks in sulfur metabolism. BMC Sys Biol 5:39. doi:10.1186/1752–0509–5–39
72. Yizhak K, Tuller T, Papp B, Ruppin E (2011) Metabolic modeling of endosymbiont genome reduction on a temporal scale. Mol Syst Biol 7. doi:479.10.1038/msb.2011.11
73. Pal C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440(7084):667–670
74. Pal C, Papp B, Lercher MJ (2005) Horizontal gene transfer depends on gene content of the host. In: Joint meeting of the 4th european conference on computational biology/6th meeting of the spanish-bioinformatics-network, Madrid, SPAIN, Sep 28-Oct 01 2005. pp 222–223. doi:10.1093/bioinformatics/bti1136
75. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37(12):1372–1375. doi:10.1038/ng1686
76. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson LD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature 399(6734):323–329
77. Ochman H, Lawrence J, Groisman E (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304
78. Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3(5):e130
79. Ochman H, Lerat E, Daubin V (2005) Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci USA 102:6595–6599
80. Choi IG, Kim SH (2007) Global extent of horizontal gene transfer. Proc Natl Acad Sci USA 104(11):4489–4494
81. Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Ann Rev Microbiol 55:709–742
82. Daubin V, Ochman H (2004) Quartet mapping and the extent of lateral transfer in bacterial genomes. Mol Biol Evol 21(1):86–89
83. Lawrence JG, Ochman H (1998) Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA 95(16):9413–9417

84. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Colladovides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of Escherichia-Coli K-12. Science 277(#5331):1453–1462
85. Postgate JR (1994) The outer reaches of life. Cambridge University Press, Cambridge, UK
86. Dantas G, Sommer MOA, Oluwasegun RD, Church GM (2008) Bacteria subsisting on antibiotics. Science 320(5872):100–103. doi:10.1126/science.1155157
87. Rehmann L, Daugulis AJ (2008) Enhancement of PCB degradation by *Burkholderia xenovorans* LB400 in biphasic systems by manipulating culture conditions. Biotechnol Bioeng 99(3):521–528. doi:10.1002/bit.21610
88. van der Meer JR, Werlen C, Nishino SF, Spain JC (1998) Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. Appl Environ Microbiol 64(11):4185–4193
89. van der Meer JR Evolution of novel metabolic pathways for the degradation of chloroaromatic compounds. In: Beijerinck centennial symposium on microbial physiology and gene regulation - emerging principles and applications, The Hague, Netherlands, Dec 1995. pp 159–178
90. Copley SD (2000) Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. Trends Biochem Sci 25(6):261–265
91. Cline RE, Hill RH, Phillips DL, Needham LL (1989) Pentachlorophenol measurements in body-fluids of people in log homes and workplaces. Arch Environ Contam Toxicol 18(4):475–481
92. Dobzhansky T (1964) Biology, molecular and organismic. Am Zool 4:443–452
93. Benfey PN, Mitchell-Olds T (2008) Perspective - From genotype to phenotype: Systems biology meets natural variation. Science 320(5875):495–497. doi:10.1126/science.1153716
94. Wagner A (2011) The origins of evolutionary innovations. A theory of transformative change in living systems. Oxford University Press, Oxford, UK
95. Rodrigues JF, Wagner A (2009) Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comp Biol 5(12):e1000613
96. Samal A, Rodrigues JFM, Jost J, Martin OC, Wagner A (2010) Genotype networks in metabolic reaction spaces. BMC Sys Biol 4:30
97. Gavrilets S, Gravner J (1997) Percolation on the fitness hypercube and the evolution of reproductive isolation. J Theor Biol 184(#1):51–64
98. Reidys CM, Stadler PF (2002) Combinatorial landscapes. SIAM Rev 44:3–54
99. Ndifon W, Plotkin JB, Dushoff J (2009) On the accessibility of adaptive phenotypes of a bacterial metabolic network. Plos Comput Biol 5(8). doi:e1000472.10.1371/journal.pcbi.1000472
100. Meiklejohn C, Hartl D (2002) A single mode of canalization. Trends Ecol Evol 17(10):468–473
101. Wagner A (2005) Robustness and evolvability in living systems. Princeton University Press, Princeton, NJ
102. Wagner GP, Booth G, Bagherichaichian H (1997) A population genetic theory of canalization. Evolution 51(#2):329–347
103. Papp B, Teusink B, Notebaart RA (2009) A critical view of metabolic network adaptations. HFSP J 3(1):24–35. doi:10.2976/1.3020599
104. Wang Z, Zhang J (2009) Abundant indispensable redundancies in cellular metabolic networks. Genome Biol Evol 1:23–33
105. Freilich S, Kreimer A, Borenstein E, Gophna U, Sharan R, Ruppin E (2010) Decoupling environment-dependent and independent genetic robustness across bacterial species. PLoS Comp Biol 6(2). doi:e1000690.10.1371/journal.pcbi.1000690
106. Ciliberti S, Martin OC, Wagner A (2007) Innovation and robustness in complex regulatory gene networks. Proc Natal Acad Sci USA 104:13591–13596
107. Ferrada E, Wagner A (2008) Protein robustness promotes evolutionary innovations on large evolutionary time scales. Proc Roy Soc Lond B Biol Sci 275:1595–1602

108. Schuster P, Fontana W, Stadler P, Hofacker I (1994) From sequences to shapes and back - a case-study in RNA secondary structures. Proc Roy Soc Lond B 255(1344):279–284
109. Lipman D, Wilbur W (1991) Modeling neutral and selective evolution of protein folding. Proc Roy Soc Lond B 245(1312):7–11
110. Raman K, Wagner A (2011) Evolvability and robustness in a complex signaling circuit. Mol BioSyst 7:1081–1092