Molecular Evolution, Networks in

From: Meyers, R.A. (ed.) Encyclopedia of Complexity and System Science. Springer, Heidelberg (2009)

Prof. Dr. Andreas Wagner
University of Zurich
Department of Biochemistry
Computational Biology and Bioinformatics
Building Y27
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

Email: aw@bioc.unizh.ch

Article Outline

Glossary

Genome: The totality of genetic material in an organism, including all its genes.

Metabolite: A small molecule that is either produced or consumed by a chemical reaction that takes place inside an organism.

Graph: A mathematical object that consists of nodes (vertices). Pairs of nodes may be connected by edges.

Gene frequency or allele frequency: The proportion of individuals in a population that carries a specific allele.

Paralogous genes: Genes in the same genome that originated in a gene duplication event.

Orthologous genes: Genes in the genomes of two different species that shared a common ancestor in the ancestral species.

Nonsynonymous nucleotide substitution/amino acid replacement substitution: A nucleotide substitution in a gene that changes the amino acid sequence of the encoded gene.

Synonymous nucleotide substitution/amino acid replacement substitution: A nucleotide substitution in a gene that does not change the amino acid sequence of the encoded gene.

Protein domain: a protein region with a characteristic function and structure that often also folds autonomously.

I. Definition

Molecular evolution is concerned with evolutionary change of nucleic acids and proteins. It attempts to identify the evolutionary forces that cause these molecules to change their structure over millions of years. Molecular evolution as a research field emerged in the second half of the 20^{th} century, when information on DNA and protein sequences first became available. Although studies in the field initially focused on the evolution of genes and the proteins they encode, they increasingly concentrate on the evolution of whole genomes. This was made possible by the availability of whole genome sequences in the mid-1990s.

Recently, technological developments have made it possible to study molecular networks inside cells. These networks encompass hundreds or thousands of proteins that interact with each other, with DNA, and with small metabolites. A molecule's position in such a molecular network, as well as its interaction partners may influence the tempo and mode of the molecule's evolution. In addition, change in individual molecules of a network can influence the network's structure on an evolutionary time scale. These two topics form the core of this contribution.

II. Introduction.

II.1. Molecular networks. A molecular network is a highly heterogeneous assemblage of different molecules, including small metabolites, RNA, DNA, proteins, and protein complexes. Molecules in this assemblage interact with each other in a variety of ways to carry out important cellular functions. In any one cell, the structure of this network changes as a function of the cell's physiological state, and as a function of the proteins and RNA molecules that are expressed at any one time. No experimental technique is currently available that could reveal the full complexity of a molecular network, much less its temporal dynamics. However, much information is available on (sub)networks that are characterized by one kind of molecular interactions. Specifically, three kinds of such networks have been characterized extensively in different organisms. The first kind is a *protein interaction network*. It can be represented as a graph whose nodes are proteins, and where two proteins are connected by an edge if they physically interact

inside a cell. The second kind of network is a *transcriptional regulation network*. Here, the nodes of the network are genes. A directed edge connects a gene A to a gene B in such a network, if A encodes a transcriptional regulator, a protein that binds to regulatory DNA near gene B, and if A activates or represses the transcription of B. The third and final class of well-characterized networks comprises *metabolic reaction networks*. They are networks of chemical reactions that sustains life by producing energy and biochemical building blocks for cell growth. Metabolic networks consist of two kinds of key parts, metabolites and metabolic enzymes. Metabolic enzymes catalyze chemical reactions that convert metabolites into other metabolites. These enzymes are encoded by genes.

Because these three kinds of networks, protein interaction networks, transcriptional regulation networks, and metabolic networks, are by far the best studied kinds of biological networks, this contribution will focus on them.

II. 2. Molecular evolution. Genes and the proteins they encode are key components of the three networks introduced above. Genomic DNA in general, and genes in particular can undergo three principal kinds of evolutionary genetic change (mutations). The first kind is a *deletion*, whereby a gene or a part of it becomes eliminated from the genome. The second kind is a *duplication*, whereby a stretch of genomic DNA becomes duplicated, such that two copies of the DNA sequence come to exist in the genome. The duplicate DNA can reside immediately adjacent or far away from the original, depending on the mechanism of duplication. If such a duplication encompasses one or more genes, one speaks of a gene duplication. Gene duplications have received considerable attention since whole genome sequences have become available, because they lead to an increase of the number of genes in a genome, and because they may allow the evolution of genes with new functions. The third kind of change is a *point mutation*. Here, a single nucleotide changes. If the change occurs inside of a gene, then the amino acid sequence of the encoded protein may also change, leading to a potential change of function in the protein. More complicated kinds of evolutionary change also occur, such as rearrangements of parts or domains within a protein. Their impact on network evolution is less intensely studied and has been reviewed elsewhere [1].

This characterization of evolutionary genetic change distinguishes different kinds of molecular events. In addition, one can also distinguish genetic change through its effects on fitness. Here again, there are three possible classes of change. The first class consists of *neutral mutations*. Such mutations are causing a change in genomic DNA that leaves an organism's fitness unchanged. A second class comprises *beneficial mutations*, mutations that increase an organism's fitness. Natural selection increases the frequency of genes carrying such mutations in a population. A third class consists of *deleterious mutations* which decrease the fitness of an organism, and are thus eliminated from populations. For this reason, deleterious mutations do not contribute to observed molecular variation, even though they may be the most frequent mutations. Despite 40 years of research, it is still a matter of debate whether most mutations that give rise to observed variation in a population of organisms are neutral or beneficial.

Molecular evolution as a research field emerged in the second half of the 20^{th} century, with the availability of the first DNA and protein sequences. A key theoretical development in the field was Kimura's neutral theory of molecular evolution [2]. This

theory makes specific predictions about the fate of neutral mutations. Specifically, the rate at which neutral mutations arise that will eventually go to fixation, that is, attain a frequency of one, equals the rate of neutral mutations itself, and is constant and independent of population size. The time neutral mutations take to go to fixation is proportional to the size of a population. These simple predictions do not hold for beneficial mutations whose fate also depends on the amount of fitness benefits they confer. These predictions of the neutral theory are well corroborated, but Kimura and others made additional claims that were more controversial. Specifically, they maintained that neutral mutations comprised the vast majority of mutations that give rise to genetic variation in a population, a claim that gave rise to the neutralist-selectionist controversy [3, 4]. Although this debate has not been fully resolved, recent analyses based on whole-genome data suggest that many mutations that occur in a genome have beneficial effects [5, 6].

Multiple sequence characteristics can be used to determine whether the DNA sequence of a gene has been subject to mostly negative selection that eliminates deleterious mutations, to positive selection that has increased beneficial mutations in frequency, or to no selection (neutral evolution) [7]. One such characteristic, the ratio K_a/K_s of non-synonymous to synonymous nucleotide substitution is simple and widely used. In order to determine this ratio, one compares two genes and the mutations that have accumulated since their common ancestry (either since a gene duplication event for paralogous genes, or since a speciation event for orthologous genes). Specifically, one estimates the number of non-synonymous mutations, mutations that did change the amino acid sequence of the encoded protein, and the number of synonymous mutations, mutations that did not change the protein. Such mutations are possible, because the genetic code is redundant, that is, multiple codons may encode the same amino acid. More specifically still, one estimates K_s, the fraction of synonymous substitutions per synonymous nucleotide site in a gene, and K_a, the fraction of amino acid replacement substitutions per replacement site. These measures of divergence account for the fact that different genes have different length. From these estimates, one then calculates K_a/K_s. If this ratio is smaller than one, than the genes in question have tolerated fewer amino acid replacement substitutions in their evolutionary history than synonymous substitutions. This means that the genes are under negative or purifying selection, because some amino acid substitutions have been eliminated from the evolutionary record. If the ratio is equal to one ($K_a=K_s$), then an equal number of silent and replacement substitutions have been preserved. Such genes evolve neutrally. This pattern of evolution is typical of pseudogenes, genes that have lost their function through mutations. Finally, if the ratio is greater than one, then more amino acid changing mutations have been preserved than synonymous mutations, meaning that the genes have been subject to net positive selection. For the vast majority of genes, the ratio K_a/K_s is much smaller than one, meaning that these genes are under net negative or purifying selection. For these genes, K_a/K_s is a good indicator of the evolutionary constraint a gene is subject to: Genes with small K_a/K_s are said to be more highly constrained than genes with a large K_a/K_s.

II. 3. Molecular networks and molecular evolution.

Two principal kinds of genetic change can be distinguished in the molecular evolution of molecular networks. First, there is change that affects the number of network parts itself, either by adding network parts through duplication, or by eliminating network parts through deletion. Second, there is change that leaves the network size unaffected, but that changes existing network parts and their interactions through point mutations. A comprehensive analysis of network evolution would study both categories of change, and it would analyze how such change affects the structure of a network. Such an analysis would also study how natural selection on network function would influence the kinds of genetic change that can be tolerated on evolutionary time scales. Partly because of a lack of necessary data, no such comprehensive analysis exists for all of the molecular networks discussed here. One kind of change and its impact on a network may have been studied for one kind of network, but hardly at all for another network. The next sections highlight insights available from studies focusing on one or the other kind of change and its effects on the evolution of protein interaction networks, transcriptional regulation networks, and metabolic networks.

III. Protein interaction networks

III.1. Characterizing protein interaction networks.

Two prominent experimental approaches exist to characterize protein interaction networks exist (reviewed in [8]). These approaches illustrate the kinds of data available for evolutionary studies on such networks. The first approach is the yeast two-hybrid assay [9], a technique to identify interactions between two specific proteins A and B (not necessarily from yeast). This assay first uses recombinant DNA techniques to generate two hybrid proteins. In one of these hybrids, protein A is fused to the transcriptional activation domain of a yeast transcription factor. In the other hybrid, the transcriptional activation domain of the same transcription factor is fused to protein B. If protein A and B interact in vivo, then their interaction physically links the transcriptional activation and the DNA binding domain of the transcription factor, thus allowing transcriptional activation of a suitably chosen "reporter" gene, which can be easily detected. The two hybrid approach has been applied to detect interactions of most protein pairs A-B in a genome [10-18]. Even for a small genome like that of the yeast *Saccharomyces cerevisiae*, this requires screening millions of pairwise interactions.

The first genome-wide protein interaction screens that used the two-hybrid assay were carried out in the yeast proteome itself. They yielded maps of protein interactions involving some 1000 proteins [14, 15]. Variations of the approach have been applied successfully to analyze protein interactions in other microbes, such as the bacterium *Helicobacter pylori* [16], and protein interactions between viral and cellular proteins [11, 12]. The yeast two-hybrid approach has several commonly recognized shortcomings. One of them is the use of fusion proteins, which can lead to protein misfolding. Another problem is that the assay forces coexpression of proteins in the same compartment of a cell or an organism, although the proteins may not co-localize *in vivo*. These shortcomings lead to potentially high false positive and false negative error rates, i.e., to the detection of spurious interactions, and to the failure to detect actual interactions.

These error rates may well exceed 50% [19, 20]. This complication means that it is currently difficult to evaluate which of the (vast) differences in network composition and interactions observed among distantly related organisms is due to evolutionary divergence, and which part is due to experimental error.

Another class of techniques to characterize protein interaction networks identifies the proteins that are part of a multiprotein complex [21-23]. Here, the departure point of a typical experiment is some protein A of interest, and the experiment asks which protein complexes –groups of interacting proteins – this protein A is a part of. In the experiment, protein A is reversibly attached to a solid support via a chemical tag. This solid support is exposed to a protein extract from cells. As a result, proteins that can interact with A, become attached to the support via protein A. Protein A and all proteins attached to it are then released from the support, at which point the proteins can be isolated and characterized, for example through mass spectrometry. The whole approach is a variation of affinity chromatography, a chemical separation technique that takes advantage of specific binding of one molecule to another. The largest-scale approaches so far have identified more than 400 protein complexes in the yeast *Saccharomyces cerevisiae* [21-24].

The yeast two-hybrid assay and affinity chromatography based methods lead to different and complementary kinds of information. The yeast two-hybrid assay yields information about pairwise protein interactions. In contrast, affinity chromatography-based methods lead to information about the proteins that occur in a protein complex, where not all of the proteins in a complex may interact directly with each other.

III.2. Characterizing network structure.

Perhaps the most basic and general question that one can ask about protein interaction networks (or any other molecular network) is why a network has its observed structure. To answer this question ultimately requires an evolutionary perspective, because any network's structure needs to be explained from its evolutionary history and the evolutionary forces shaping it. To answer this question, however, one has to first know what a network's structure *is*. Because molecular networks have thousands of parts, visual inspection is of little use in identifying a network's structure, and it is not always clear what features of the structure to focus on. Most existing work focuses on the simplest structural network characteristics, three of which are given below. Others are also in use, but many biologically sensible such characteristics may still await discovery.

Perhaps the simplest structural characteristic one can study is the distribution of the number d of interactions per protein, the so-called degree distribution of a network. A second characteristic are degree correlations among proteins, that is, one can ask whether highly (lowly) connected proteins preferentially connect to highly (lowly) connected other proteins. A third basic characteristic of a molecular network is the clustering coefficient C [25]. To define the clustering coefficient $C(v)$ of a node (protein) v in a graph, consider all k_v nodes adjacent to a node v, and count the number m of edges that exist among these k_v nodes(not including edges connecting them to v). The maximally possible m is $k_v(k_v-1)/2$, in which case all m nodes are connected to each other. Let $C(v):=m/(k_v(k_v-1)/2)$. $C(v)$ measures the "cliquishness" of the neighborhood of v, i.e.,

what fraction of the nodes adjacent to v are also adjacent to each other. The clustering coefficient *C* of the whole network is defined as the average of *C(v)* over all *v*.

The degree distribution of protein interaction networks resembles a power law, $P(d) \sim d^{-\gamma}$, where γ is a constant characteristic of the network.[26, 27], protein degrees are anticorrelated, that is, highly connected proteins preferentially interact with lowly connected proteins [28], and the clustering coefficient of protein interaction networks is much higher than that of random networks with the same number of interactions.

III.3. Protein network structure and molecular evolution

A variety of evolutionary models have attempted to ask why networks have their observed structure with respect to the above and some other simple structural features [29-37]. These models rely on two main ingredients, addition and deletion of network proteins (caused by gene duplications and deletions) which can change the size of a network, and "rewiring" of network interactions driven by point mutations in the genes encoding network proteins. Both processes undoubtedly play a role in network evolution. Network rewiring must occur, because individual mutations can change protein-protein interface necessary for interactions. Gene duplication and gene deletion must also play a role, because genomes vary in size by orders of magnitude, and so do the number of genes, encoded proteins, and protein interaction network size. In addition, some families of interacting proteins such as heterodimerizing transcription factors have arisen largely through gene duplication [38, 39]. Furthermore, gene duplication play a role in the evolution of new protein complexes in yeast [40].

Beyond these generalities, the available models differ widely in their assumptions, and about the importance they ascribe to rewiring and duplication/deletion. They include differences in assumptions about (i) rates of duplication, deletion, and rewiring, (ii) whether these processes are random with respect to network structure, or whether their rate depends on a protein's position in the network, and (iii) whether duplication/deletion and rewiring occur independently from one another or whether they are in some way coupled. Most existing models constitute mathematical proofs of principle, that is, they attempt to show that a particular network feature, such as the degree distribution *could* be explained by a particular evolutionary process, whereas a few models attempt to stay close to available molecular evolution data. However, this data is currently very limited, because no information is available about the structure of protein interaction networks in closely related organisms. That is, the available data is either derived from comparisons of protein content and/or network structure of very distantly related organisms, or from within one genome, such as from gene duplicates (whose age can estimated) and their common interaction partners [26, 41]. Although such data is insufficient to validate or refute any one of the models to the exclusion of all others, a limited amount of evidence favors a preferential attachment mechanism of network evolution. In this mechanism, proteins that have arisen early during network evolution tend to be highly connected proteins, and such highly connected proteins may acquire more interactions subsequently [36, 42, 43, 44, but see also Kunin, 2004 #2133].

Despite all their differences, existing network evolution models have an important unifying feature: None of them require that natural selection molds any global feature of network structure, such as the degree distribution. This observation is significant, because

early work on molecular networks assumed that features of protein interaction networks, such as the power-law degree distribution reflect evolutionary optimization of some aspect of network function. For example, in protein interaction networks and other networks with power-law degree distributions, the mean distance between network nodes that can be reached from each other (via a path of edges) is very small and it increases only very little upon random removal of nodes [45]. This distance can be thought of as a measure of how compact a network is. In graphs with other degree distributions, this mean distance can increase substantially upon node removal. From this observation emerged the proposition that robustly compact networks confer some (unknown) advantages on a cell, and that the power law degree distribution reflects the action of natural selection on the degree distribution itself. The observation alone that power-law degree distributions are ubiquituous in biological and non-biological systems argues against this proposition. The models mentioned above, none of which require natural selection on the degree distribution, further speak against it. In addition, an even simpler hypothetical explanation of observed network structure has been proposed. This hypothesis explains the degree distribution and other network features by a random model of desolvation energies among interacting protein pairs [46].

One might be tempted to call network evolution in the absence of natural selection optimizing a global network feature *neutral evolution*. Doing so, however, would neglect that natural selection almost certainly influences which duplication/deletion/rewiring events are preserved in the evolutionary record. In other words, even though natural selection may not influence global network structure, it may affect the local events that change network structure in evolutionary time. Multiple lines of evidence hint at this influence of natural selection. The first comes from a study on protein complexes. In the yeast *Saccharomyces cerevisiae*, over- or underexpression of members of a protein complex may have adverse effects on fitness. The likely reason is that such expression changes affect the stoichiometric balance of the proteins in a cell which is necessary for forming complexes with the correct protein composition [47, 48]. Gene duplications of proteins interacting in a complex may be harmful, because such duplications effectively change gene expression, which distorts this balance. In agreement with this observation, proteins encoded by members of large gene families, genes that have often undergone duplication, are underrepresented in protein complexes [48]. A second, similar indication of the influence of natural selection on network evolution is that the number of proteins in a complex encoded by single copy genes rises with complex size [49]. Thirdly, gene duplications seem to have been preferentially preserved in the sparsely connected parts of the yeast protein interaction network, parts that are characterized by low degree [50] or low clustering coefficients [51]. This suggests that gene duplications in densely connected network parts may have deleterious effects.

Rather than focusing on global networks structure, a limited amount of work has focused on small subgraphs of a protein interaction network. Such subgraphs comprise only few (3-5) proteins, are characterized by specific patterns of interactions, and are also known as network motifs. Proteins that occur in larger and more densely connected motifs have a greater likelihood to be preserved across distantly related species [52, 53].

All work discussed thus far has focused on the evolution of the network itself. Another line of inquiry asks how a protein's position within a network constrains the

protein's evolution. For example, as already discussed above, gene duplications tend to be observed preferentially for genes in sparsely connected parts of a network [50, 54]. Also, proteins that have a more central role in the protein interaction network evolve more slowly [55]. In addition, early work suggested that proteins with more interaction partners are evolutionarily more constrained [56, 57]. This association has become controversial, because it may be caused by bias in protein interaction data sets, and because it may be explained by differences in gene expression level among proteins with different numbers of interaction partners [58-64]. Specifically, highly expressed proteins evolve more slowly, and much of the observed variation in evolutionary rates among proteins may be due to variation in expression level [65, 66], leaving only a minor role for the influence of protein-protein interactions.

Rather than just considering the numbers of interactions of a protein in a protein network when trying to explain evolutionary rate differences, it may be necessary to distinguish between different kinds of interactions. One important distinction here is that between transient and permanent interactions. Proteins that enter permanent interactions, thus forming stable complexes with other proteins, evolve at lower rates than proteins that undergo transient interactions, or proteins that are not known to interact with other proteins [54, 67]. A closely related distinction is that between proteins that have multiple protein interaction interfaces, and that can thus interact with multiple proteins at the same time, and between proteins that have a single interaction interface, and that interact with multiple partners successively and transiently. Multi-interface proteins evolve more slowly, which may be readily explained by the larger fraction of their surface that is constrained [68]. Yet other distinctions among interactions may also affect evolutionary rates. For example, interactions between proteins of different cellular functions may constrain evolutionary rates particularly strongly [69].

In sum, models of protein on network evolution agree that natural selection on global network structure is not necessary to produce protein interaction networks with the global features that have been studied so far. Nonetheless, these models differ in the relative importance they ascribe to deletion and duplication on one hand, and interaction rewiring on the other hand, in network evolution. Studies focusing on the evolution of network parts within an existing network suggest that a protein's position in a network influences these constraints. Nonetheless, this influence may be minor compared to other factors, especially protein expression level.

IV. **Transcriptional regulation networks**

IV.1. Characterizing transcriptional regulation networks.

In a transcriptional regulation network, transcription factors bind to regulatory DNA near network genes, and activate or repress the expression of these genes. Transcription factors are proteins that are themselves encoded by genes in the network. Transcriptional regulation networks thus comprise two main kinds of genes, genes encoding transcriptional regulators, and their regulatory target genes. However, the two classes of genes overlap, because genes encoding transcriptional regulators may themselves be transcriptionally regulated. Even small genomes such as that of the yeast *Saccharomyces cerevisiae* contain hundreds of genes encoding transcriptional regulators.

Two principal approaches have been pursued to characterize transcriptional regulation networks. One of them is manual curation, whereby data from existing experimental literature about the targets of individual transcription factors, is assembled into a network [70, 71]. The second approach is high-throughput experimental analysis of DNA binding by transcriptional regulators. This approach permits the genome-scale identification of regulatory DNA regions bound by transcription factors. It thus provides hints which genes may be regulated by which transcription factors, although transcription factor binding is only a necessary, but not a sufficient criterion for transcriptional regulation. A prominent technique used in this area is chromatin immunoprecipitation. In this technique, a transcriptional regulator is labeled with an epitope tag, a molecule that can be recognized by a specific antibody. Genomic DNA, some of which is bound by the regulator, is then isolated. This isolate is then exposed to the antibody, in order to precipitate the DNA bound by the regulator, hence the name immunoprecipitation. The precipitated DNA is then hybridized to a DNA microarray, allowing its identification and localization in the genome. In one prominent study using this technique putative candidate target genes of 106 yeast transcriptional regulators were identified in this way [72].

These two approaches to characterize transcriptional regulation networks are complementary: Manual curation may reveal high quality information about individual transcriptional regulators, but it may capture only a limited number of regulatory interactions. The high-throughput approach, on the other hand, provides more comprehensive information at the price of greater uncertainty about the biological relevance of the observed interactions.

It is noteworthy that transcriptional regulation networks, as opposed to protein interaction networks, are directed networks. This means that interactions occur from a regulator to its target gene, but not necessarily vice versa. In a graph representation of such a network, genes are thus connected by directed edges.

IV.2. Transcriptional regulation networks and molecular evolution

The molecular evolution of transcriptional regulation networks has received less attention than that of protein interaction networks. In existing work, some parallels to evolutionary patterns in protein interaction networks are evident. First, a gene's connectedness within the network may have only a weak or no impact on its rate of evolutionary. Specifically, the number of target genes of a transcriptional regulator does not affect the regulator's evolutionary rate, as indicated by the ratio K_a/K_s. Similarly, the number of transcriptional regulators that regulate a given target gene does not strongly influence the evolutionary

rate of the target gene. Second, as in protein interaction networks, gene duplications are also very important in the formation of transcriptional regulation networks [73-76]. For example, a large proportion of transcriptional regulators themselves are products of duplicate genes. The exact proportion depends on how duplicates are identified. For example, approaches that identify duplicate genes through their domain architecture may reveal that a majority of transcriptional regulators are the results of gene duplication, whereas approaches based on significant sequence similarity among transcriptional regulator genes may ascribe a lesser role to duplication. Thirdly, rewiring of transcriptional regulation interactions has also played a prominent role in transcriptional regulation networks [77-80]. Such rewiring can be accomplished in two ways. First, a mutation may change the DNA binding domain of a transcription factor, such that the factor recognizes a different spectrum of regulatory DNA motifs. However, because any one transcriptional regulator may regulate hundreds of genes, many such changes are likely to be deleterious and may not be preserved in the evolutionary record. Second and perhaps more importantly, changes in a gene's regulatory region may affect which transcription factors can bind to and regulate a gene of interest. Because the regulatory DNA sequence motifs at which a transcription factor binds are often very short, binding sites can be easily created or destroyed through mutations. For example, in a study comparing gene expression patterns between the yeasts *Candida albicans* and *Saccharomyces cerevisiae*, a strong expression correlation was found between cytoplasmatic and ribosomal proteins in *C. albicans* but not in *S. cerevisiae*. This difference was associated with a change in multiple short regulatory DNA elements that drive the expression of these genes [79].

How fast and to what extent gene functions diverge after gene duplication is a subject of considerable interest to molecular evolutionists. Gene regulation and gene expression are an important aspect of gene function. Duplicate genes in a transcriptional regulation network are thus ideal study subjects to help answer this question. For example, one can ask to what extent duplicate genes of different sequence similarity (and thus different age) share transcription factors that bind at their regulatory regions. The answer is that duplicate genes rapidly diverge in the number of shared transcription factors they share [76, 81]. For example, duplicate genes in yeast may lose 3% of common transcription factors for every 1% of sequence divergence [81]. The process of divergence, however, does not only involve loss of common transcription factor binding sites. A gain of new sites unique to each member of a duplicate gene pair may be equally important [76, 82].

One area that has received perhaps more attention than in protein interaction networks is the analysis of small and highly abundant genetic circuit motifs in transcriptional regulation networks [22, 83, 84]. An example for such a regulatory motif is a transcriptional feed-forward loop, where a transcriptional regulator A regulates the expression of a regulator B, which regulates the expression of some target gene C, which is also regulated by A. Multiple other classes of network motifs are known. A wide spectrum of possibilities exist for the evolutionary origin of these circuits. At the two extremes of this spectrum stand two scenarios. First, these circuits may have arisen through the duplication – and subsequent functional diversification – of one or a few

ancestral circuits, that is, through the duplication of each of their constituent genes in a series of duplication events. Alternatively, most of these circuits may have arisen independently by recruitment of unrelated genes. In this case, abundant circuits would have arisen through *convergent evolution*. Convergent evolution – the independent origin of similar organismal features– is a strong indicator of optimal "design" of a feature.

Because the complete genome sequence is available for the yeast *Saccharomyces cerevisiae*, one can ask which of these scenarios better reflects the evolutionary history of transcriptional regulation motifs. The answer is that the vast majority of highly abundant transcriptional regulation motifs have not originated through gene duplication, but independently and convergently [78]. What are the favourable functional properties of such networks, the properties that would drive such convergent evolution? Answers are beginning to emerge from a mix of computational and experimental work [84-86]. For example, a feed-forward loop may activate the regulated ('downstream') genes only if the upstream-most regulator is persistently activated. It can thus filter intracellular gene expression noise, which is known to be ubiqituous.

V. Metabolic networks

V.1. The analysis of genome-scale metabolic networks.

Complex chemical reaction networks comprising hundreds to thousands of reactions sustain all of life. In free-living, heterotrophic organisms, these reaction networks transform food into energy and biosynthetic building blocks for growth and reproduction. Complete (or nearly so) maps of core metabolism, comprising hundreds of reactions and metabolites are available for several model organisms [87-89]. These maps have been assembled through painstaking analysis of decades worth of biochemical literature, aided by genome sequence analysis, which may help determine whether a genome contains a gene catalyzing a particular chemical reaction.

Some work in this area focuses on network structure, by characterizing one of a variety of graph representations of a metabolic network. For example, metabolic networks can be represented as graphs whose nodes are enzymes and metabolites. Two nodes are connected if they participate in the same chemical reactions. Structural network analysis, however, has one key limitation: It does not capture the flow of matter through a metabolic network, which is at the heart of metabolic network function. This function can be computationally analyzed, even though information about enzymatic reaction rates in metabolic networks is very limited. Central to any such computational analysis are approaches such as flux balance analysis that use only information about the stoichiometry and reversibility of chemical reactions. [90, 91]. Flux balance analysis determines the rates (fluxes) at which individual chemical reactions can proceed if fundamental constraints such as that of mass conservation have to be fulfilled. Within the limits of such constraints, flux balance analysis can determine the distribution of metabolic fluxes that will maximize some metabolic property of interest. The rate of biomass production is one of these properties. It is a proxy for cell growth-rate, itself an important component of fitness in single-cell organism. Flux balance analysis makes predictions that are often in good agreement with experimental evidence in *E. coli* and

the yeast *S. cerevisiae* [89, 92, 93]. However, such predictions may fail if an organism has not been subject to natural selection to optimize growth in a particular environment.

V.2. Metabolic networks and molecular evolution.

Much as for the other two classes of networks, considerable attention has focused on the question how the structure of metabolic networks evolved [94-102]. With some exceptions [92, 103, 104], most work has not focused on metabolites, but on enzymes and their role in network evolution. The reason is that the evolution of metabolic enzymes can be better reconstructed through gene sequence and protein structure comparisons.

In the evolution of metabolic network structure, rewiring clearly has much lower importance than in the previous networks discussed here. In contrast to protein interaction networks and regulatory networks, where interactions could form between a wide variety of different proteins, in metabolic networks interactions are largely dictated by chemistry. That is, only two enzymes that share substrates or products of their reactions can be neighbors in the network.

In contrast to rewiring, gene duplications and gene deletions play an important role in network evolution. Long before information on genome-scale metabolic reaction networks became available, gene duplication already played an important role in two major hypotheses about the evolution of metabolic pathways. The first such hypothesis is that of "retrograde" evolution. According to this hypothesis, metabolic pathways evolved backwards from their (essential) end-products, through the addition of new enzymes produced by gene duplication, and in response to the depletion of substrates that are necessary for production of these end-products [105]. The second, "patchwork evolution" hypothesis postulates that enzymes originally had broad substrate specificities, and that they subsequently evolved more specialized functions through gene duplication [106]. Recent genome-scale analyses of metabolic network evolution suggest that both processes may occur, but that retrograde evolution by gene duplication is relatively rare [95, 99, 107]. For example, only a small fraction of adjacent enzymes in the same pathway arose from gene duplication. In contrast, recruitment of duplicate enzymes into new pathways is very frequent [99]. Gene duplications can also produce isoenzymes, enzymes with the same catalytic function, and thus the same position in a pathway, but possibly differential regulation. The locations in a network where such duplications are most likely preserved are not random. For example, isoenzymes are most often observed for enzymes with high metabolic flux, enzymes through which a lot of matter flows per unit time [108, 109].

The phenomenon of horizontal gene transfer, which can add new genes from a different organism to a network has also received some attention in the analysis of metabolic network evolution [110, 111]. Genes encoding metabolic enzymes are frequently transferred horizontally among bacterial genomes. However, such transfer does not affect all classes of enzymes equally. Specifically, peripheral network reactions, which are often reactions that are involved in an organism's response to specific environmental demands, are more frequently added to a network by horizontal transfer than central reactions [111]. This does not mean, however, that central parts of metabolism are completely invariant. Even a metabolic cycle as central as the tricarboxylic acid cycle can undergo substantial evolutionary change [112].

Gene duplication and horizontal gene transfer are both mechanisms by which metabolic networks can increase in size. Gene deletions, in contrast, reduce network size. They are especially important in the evolution of organelles or organisms with reduced genome size, such as chloroplasts [113] and endosymbiotic cells [96]. In such cells, the host cell provides most metabolites necessary for survival, and metabolic networks are thus often drastically reduced in size. Deletions of enzyme-coding genes have also been studied in the evolution of individual pathways. Examples include the vitamin B6 synthesis pathway, which has been lost multiple times independently through gene deletions during animal evolution [114].

Like in the other two networks discussed above, some work has also focused on the evolution of enzyme-coding genes *within* a network and within metabolic pathways. An example regards enzymes involved in the biosynthesis of anthocyan, a plant pigment. In this pathway, upstream enzymes are subject to greater evolutionary constraints than downstream enzymes, as indicated by their lower rate at which non-synonymous substitutions accumulate [115]. Here, the upstream enzymes are located above metabolic branch points that lead to other metabolic pathways. It is thus possible that mutations in them are more likely to be deleterious, because they affect more than one pathway. This evolutionary pattern of higher conservation in upstream genes is, however, not universal [116].

More recent work has asked how the amount of metabolic flux through an enzyme might affect its evolutionary rate. Here, a negative association exists between the flux through individual enzymatic reactions in yeast, as predicted by flux balance analysis, and the ratio K_a/K_s [117]. That is, enzymes with high associated metabolic flux can tolerate fewer amino acid changes. One likely explanation is that the products of high-flux enzymes play a role in multiple metabolic pathways. Thus, mutations in such enzymes, most of which reduce their metabolic output, are more likely to have deleterious effects.

VI. Summary and Outlook

Molecular evolution studies in molecular networks are still in their infancy, partly because genome-scale data on such networks has only become available recently. A few common patterns emerge from existing work. With the possible exception of metabolic networks, a gene's position in a network has a limited influence on its rate of evolution. In the evolution of network structure, both gene duplications and gene deletions play an important role in all three networks, and rewiring of existing interactions is important in protein interaction networks and transcriptional regulation networks. Natural selection may have only a minor role in shaping those feature of global network structure that have been studied, but many other such features remain poorly investigated. In contrast, natural selection undoubtedly influences what kinds of mutations can be tolerated during network evolution. A major future challenge is to explain the structure of biological networks in evolutionary terms through a quantitative framework that accounts for all the rates of evolutionary events that influence network structure.

VII. Bibliography

1. Bornberg-Bauer, E., et al., *The evolution of domain arrangements in proteins and interaction networks.* . Cellular and Molecular Life Sciences, 2005. **62**: p. 435-445.
2. Kimura, M., *The neutral theory of molecular evolution*. 1983, Cambridge: Cambridge University Press.
3. Kreitman, M. and H. Akashi, *Molecular evidence for natural selection.* Annual Review of Ecology and Systematics, 1995. **26**: p. 403-422.
4. Ohta, T., *The nearly neutral theory of molecular evolution.* . Annual Reviews of Ecology and Systematics, 1992. **23**: p. 263-286.
5. Smith, N.G.C. and A. Eyre-Walker, *Adaptive protein evolution in Drosophila.* Nature, 2002. **415**(6875): p. 1022-1024.
6. Fay, J., G. Wyckoff, and C.-I. Wu, *Testing the neutral theory of molecular evolution with genomic data from Drosophila.* Nature, 2002. **415**: p. 1024-1026.
7. Kreitman, M., *Methods to detect selection in populations with applications to the human.* Annual Review of Genomics and Human Genetics, 2000. **1**: p. 539-559.
8. Pandey, A. and M.P. Mann, *Proteomics to study genes and genomes.* Nature, 2001. **405**: p. 837-846.
9. Fields, S. and O.K. Song, *A novel genetic system to detect protein protein interactions.* Nature, 1989. **340**(6230): p. 245-246.
10. Fromont-Racine, M., J.C. Rain, and P. Legrain, *Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens.* Nature Genetics, 1997. **16**(3): p. 277-282.
11. Bartel, P.L., et al., *A protein linkage map of Escherichia coli bacteriophage T7.* Nature Genetics, 1996. **12**(1): p. 72-77.
12. Flajolet, M., et al., *A genomic approach of the hepatitis C virus generates a protein interaction map.* Gene, 2000. **242**(1-2): p. 369-379.
13. Ito, T., et al., *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(3): p. 1143-1147.
14. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(8): p. 4569-4574.
15. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-627.
16. Rain, J.C., et al., *The protein-protein interaction map of Helicobacter pylori.* Nature ;, 2001. **409**(6817): p. 211-215.
17. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173.
18. Li, S.M., et al., *A map of the interactome network of the metazoan C-elegans.* Science, 2004. **303**(5657): p. 540.

19. Edwards, A.M., et al., *Bridging structural biology and genomics: assessing protein interaction data with known complexes.* Trends in Genetics, 2002. **18**(10): p. 529-536.
20. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
21. Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Faseb Journal, 2002. **16**(4/pt.1): p. A523-A523.
22. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180-183.
23. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637.
24. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631.
25. Watts, D.J. and S.H. Strogatz, *Collective dynamics of small-world networks.* Nature, 1998. **393**(#6684): p. 440-442.
26. Wagner, A., *The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes.* Molecular Biology and Evolution., 2001. **18**: p. 1283-1292.
27. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**: p. 41-42.
28. Maslov, S. and K. Sneppen, *Specificity and stability in topology of protein networks.* Science, 2002. **296**: p. 910-913.
29. Wagner, A., *How the global structure of protein interaction networks evolves.* Proceedings of the Royal Society of London Series B., 2003. **270**: p. 457-466.
30. Sole, R.V., et al., *A model of large-scale proteome evolution.* Advances in Complex Systems, 2002. **5**: p. 43-54.
31. Pastor-Satorras, R., E. Smith, and R.V. Sole, *Evolving protein interaction networks through gene duplication.* Journal of Theoretical Biology, 2003. **222**: p. 199-210.
32. Vazquez, A., et al., *Modelling of protein interaction networks.* Complexus, 2001. **1**: p. 38-44.
33. Berg, J., Lassig, M., Wagner, A., *Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications.* BMC Evolutionary Biology, 2004. **4:51**.
34. Przulj, N. and D.J. Higham, *Modelling protein-protein interaction networks via a stickiness index.* Journal of the Royal Society Interface, 2006. **3**: p. 711-716.
35. Ispolatov, I., P.L. Krapivsky, and A. Yuryev, *Duplication-divergence model of protein interaction network.* Physical Review E, 2005. **71**(6): p. 061911.
36. Middendorf, M., E. Ziv, and C.H. Wiggins, *Inferring network mechanisms: The Drosophila melanogaster protein interaction network.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(9): p. 3192.
37. Goh, K.I., B. Kahng, and D. Kim, *Evolution of the protein interaction network of budding yeast: Role of the protein family compatibility constraint.* Journal of the Korean Physical Society, 2005. **46**(2): p. 551.

38. Amoutzias, G.D., D.L. Robertson, and E. Bornberg-Bauer, *The evolution of protein interaction networks in regulatory proteins.* Comparative and Functional Genomics, 2004. **5**(1): p. 79.
39. Amoutzias, G.D., J. Weiner, and E. Bornberg-Bauer, *Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs.* Gene, 2005. **347**(2): p. 247.
40. Pereira-Leal, J. and S. Teichmann, *Novel specificities emerge by stepwise duplication of functional modules.* Genome research, 2005. **4**: p. 552-559.
41. Ispolatov, I., et al., *Binding properties and evolution of homodimers in protein-protein interaction networks.* Nucleic Acids Research, 2005. **33**(11): p. 3629.
42. Pereira-Leal, J.B., et al., *An exponential core in the heart of the yeast protein interaction network.* Molecular Biology and Evolution, 2005. **22**(3): p. 421.
43. Wagner, A., *How the global structure of protein interaction networks evolves.* Proceedings of the Royal Society of London Series B-Biological Sciences, 2003. **270**(1514): p. 457.
44. Eisenberg, E. and E. Levanon, *Preferential attachment in the protein network evolution.* . Physical Review Letters, 2003. **91**: p. 138701-138704.
45. Albert, R., H. Jeong, and A.L. Barabasi, *Error and attack tolerance of complex networks.* Nature ;, 2000. **406**(6794): p. 378-382.
46. Deeds, E.J., O. Ashenberg, and E.I. Shakhnovich, *A simple physical model for scaling in protein-protein interaction networks.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(2): p. 311.
47. Lemos, B., C.D. Meiklejohn, and D.L. Hartl, *Regulatory evolution across the protein interaction network.* Nature Genetics, 2004. **36**(10): p. 1059.
48. Papp, B., C. Pal, and L.D. Hurst, *Dosage sensitivity and the evolution of gene families in yeast.* Nature, 2003. **424**(6945): p. 194-197.
49. Yang, J., R. Lusk, and W.H. Li, *Organismal complexity, protein complexity, and gene duplicability.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(26): p. 15661.
50. Prachumwat, A. and W.H. Li, *Protein function, connectivity, and duplicability in yeast.* Molecular Biology and Evolution, 2006. **23**(1): p. 30.
51. Li, L., et al., *Preferential duplication in the sparse part of yeast protein interaction network.* Molecular Biology and Evolution, 2006. **23**(12): p. 2467.
52. Wuchty, S., *Evolution and topology in the yeast protein interaction network.* Genome Research, 2004. **14**(7): p. 1310.
53. Wuchty, S., A.L. Barabasi, and M.T. Ferdig, *Stable evolutionary signal in a Yeast protein interaction network.* BMC Evolutionary Biology, 2006. **6**: p. 8.
54. Mintseris, J. and Z.P. Weng, *Structure, function, and evolution of transient and obligate protein-protein interactions.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(31): p. 10930.
55. Hahn, M.W. and A.D. Kern, *Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks.* Molecular Biology and Evolution, 2005. **22**(4): p. 803.
56. Fraser, H., et al., *Evolutionary rate in the protein interaction network.* SCIENCE, 2002. **296**(5568): p. 750-752.

57. Fraser, H.B., D.P. Wall, and A.E. Hirsh, *A simple dependence between protein evolution rate and the number of protein-protein interactions.* BMC Evolutionary Biology, 2003. **3**: p. 11.
58. Batada, N., L. Hurst, and M. Tyers, *Evolutionary and physiological importance of hub proteins.* . PloS Computational Biology, 2006. **2(7)**: p. e88.
59. Hahn, M., G.C. Conant, and A. Wagner, *Molecular evolution in large genetic networks: does connectivity equal importance?* Journal of Molecular Evolution, 2004. **58**: p. 203-211.
60. Jordan, I.K., Y.I. Wolf, and E.V. Koonin, *No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly.* BMC Evolutionary Biology, 2003. **3**: p. 1.
61. Jordan, I.K., Y.I. Wolf, and E.V. Koonin, *Correction: no simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors evolve slowly.* BMC Evolutionary Biology, 2003. **3**: p. 5.
62. Bloom, J.D. and C. Adami, *Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets.* BMC Evolutionary Biology, 2003. **3**: p. 21.
63. Bloom, J.D. and C. Adami, *Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response.* BMC Evolutionary Biology, 2004. **4**: p. 14.
64. Agrafioti, I., et al., *Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks.* BMC Evolutionary Biology, 2005. **5**: p. 23.
65. Pal, C., B. Papp, and L.D. Hurst, *Highly expressed genes in yeast evolve slowly.* Genetics, 2001. **158**: p. 927-931.
66. Drummond, D.A., et al., *Why highly expressed proteins evolve slowly.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(40): p. 14338.
67. Teichmann, S.A., *The constraints protein-protein interactions place on sequence divergence.* Journal of Molecular Biology, 2002. **324**(3): p. 399.
68. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938.
69. Makino, T. and T. Gojobori, *The evolutionary rate of a protein is influenced by features of the interacting partners.* Molecular Biology and Evolution, 2006. **23**(4): p. 784.
70. Salgado, H., et al., *RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12.* NUCLEIC ACIDS RESEARCH, 2004. **32**: p. D303-D306.
71. Guelzim, N., et al., *Topological and causal structure of the yeast transcriptional regulatory network.* NATURE GENETICS, 2002. **31**(1): p. 60-63.
72. Lee, T., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.

73. Babu, M.M. and S.A. Teichmann, *Evolution of transcription factors and the gene regulatory network in Escherichia coli.* Nucleic Acids Research, 2003. **31**(4): p. 1234.
74. Teichmann, S.A. and M.M. Babu, *Gene regulatory network growth by duplication.* Nature Genetics, 2004. **36**(5): p. 492.
75. Babu, M.M., et al., *Structure and evolution of transcriptional regulatory networks.* Current Opinion in Structural Biology, 2004. **14**(3): p. 283.
76. Evangelisti, A. and A. Wagner, *Molecular evolution in the transcriptional regulation network of yeast.* Journal of Experimental Zoology/Molecular Development and Evolution, 2004. **302B**: p. 392-411.
77. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-254.
78. Conant, G.C. and A. Wagner, *Convergent evolution in gene circuits.* Nature Genetics, 2003. **34**: p. 264-266.
79. Ihmels, J., et al., *Rewiring of the yeast transcriptional network through the evolution of motif usage.* Science, 2005. **309**(5736): p. 938.
80. Babu, M.M., S.A. Teichmann, and L. Aravind, *Evolutionary dynamics of prokaryotic transcriptional regulatory networks.* Journal of Molecular Biology, 2006. **358**(2): p. 614.
81. Maslov, S., et al., *Upstream plasticity and downstream robustness in evolution of molecular networks.* BMC Evolutionary Biology, 2004. **4**: p. 9.
82. Papp, B., C. Pal, and L.D. Hurst, *Evolution of cis-regulatory elements in duplicated genes of yeast.* Trends in Genetics, 2003. **19**(8): p. 417-422.
83. Milo, R., et al., *Network motifs: Simple building blocks of complex networks.* SCIENCE, 2002. **298**(5594): p. 824-827.
84. Shen-Orr, S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nature Genetics, 2002. **31**(1): p. 64-68.
85. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(21): p. 11980-11985.
86. Mangan, S., A. Zaslaver, and U. Alon, *The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks* Journal of Molecular Biology, 2003. **334**: p. 197-204.
87. Edwards, J.S. and B.O. Palsson, *Systems properties of the Haemophilus influenzae Rd metabolic genotype.* Journal of Biological Chemistry, 1999. **274**(#25): p. 17410-17416.
88. Edwards, J.S. and B.O. Palsson, *The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(10): p. 5528-5533.
89. Forster, J., et al., *Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.* Genome Research, 2003. **13**: p. 244-253.
90. Varma, A. and B.O. Palsson, *Metabolic capabilities of Escherichia coli. synthesis of biosynthetic precursors and cofactors.* Journal of theoretical biology., 1993. **165**: p. 477-502.

91. Schilling, C.H., J.S. Edwards, and B.O. Palsson, *Toward metabolic phenomics: Analysis of genomic data using flux balances.* Biotechnology Progress, 1999. **15**(#3): p. 288-295.
92. Segre, D., D. Vitkup, and G. Church, *Analysis of optimality in natural and perturbed metabolic networks.* Proceedings of the National Academy of Sciences of the U.S.A., 2002. **99**(15112-15117).
93. Edwards, J.S. and B.O. Palsson, *Robustness analysis of the Escherichia coli metabolic network.* Biotechnology Progress, 2000. **16**(6): p. 927-939.
94. Light, S., P. Kraulis, and A. Elofsson, *Preferential attachment in the evolution of metabolic networks.* BMC Genomics, 2005. **6**: p. 159.
95. Light, S. and P. Kraulis, *Network analysis of metabolic enzyme evolution in Escherichia coli.* BMC Bioinformatics, 2004. **5**: p. 15.
96. Pal, C., et al., *Chance and necessity in the evolution of minimal metabolic networks.* Nature, 2006. **440**(7084): p. 667.
97. Sakharkar, M.K., et al., *Insights to metabolic network evolution by fusion proteins.* Frontiers in Bioscience, 2005. **10**: p. 1070.
98. Ebenhoh, O., T. Handorf, and D. Kahn, *Evolutionary changes of metabolic networks and their biosynthetic capacities.* IEE Proceedings Systems Biology, 2006. **153**(5): p. 354.
99. Teichmann, S.A., et al., *The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli.* Journal of Molecular Biology, 2001. **311**(4): p. 693.
100. Teichmann, S.A., et al., *Small-molecule metabolism: an enzyme mosaic.* Trends in Biotechnology, 2001. **19**(12): p. 482.
101. Spirin, V., et al., *A metabolic network in the evolutionary context: Multiscale structure and modularity.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(23): p. 8774.
102. Tanaka, T., K. Ikeo, and T. Gojobori, *Evolution of metabolic networks by gain and loss of enzymatic reaction in eukaryotes.* Gene, 2006. **365**: p. 88.
103. Handorf, T., O. Ebenhoh, and R. Heinrich, *Expanding metabolic networks: Scopes of compounds, robustness, and evolution.* Journal of Molecular Evolution, 2005. **61**(4): p. 498.
104. Pfeiffer, T., O.S. Soyer, and S. Bonhoeffer, *The evolution of connectivity in metabolic networks.* Plos Biology, 2005. **3**(7): p. 1269.
105. Horowitz, N., *The evolution of biochemical syntheses - retrospect and prospect.*, in *Evolving genes and proteins.*, H. Bryson and H. Vogel, Editors. 1965, Academic Press: New York. p. 15-23.
106. Jensen, R., *Enzyme recruitment in evolution of new functions.* . Annual Reviews of Microbiology, 1976. **30**: p. 409-425.
107. Alves, R., R.A.G. Chaleil, and M.J.E. Sternberg, *Evolution of enzymes in metabolism: A network perspective.* Journal of Molecular Biology, 2002. **320**(4): p. 751.
108. Vitkup, D., P. Kharchenko, and A. Wagner, *Influence of metabolic network structure and function on enzyme evolution.* Genome Biology, 2006. **7**(5): p. R39.
109. Papp, B., C. Pal, and L.D. Hurst, *Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast.* Nature, 2004. **429**(6992): p. 661-664.

110. Ma, H.W. and A.P. Zeng, *Phylogenetic comparison of metabolic capacities of organisms at genome level.* Molecular Phylogenetics and Evolution, 2004. **31**(1): p. 204.
111. Pal, C., B. Papp, and M.J. Lercher, *Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.* Nature Genetics, 2005. **37**(12): p. 1372.
112. Huynen, M.A., T. Dandekar, and P. Bork, *Variation and evolution of the citric acid cycle: a genomic perspective.* Trends in Microbiology, 1999. **7**(#7): p. 281-291.
113. Wang, Z., et al., *Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria.* BMC Genomics, 2006. **7**: p. 100.
114. Tanaka, T., Y. Tateno, and T. Gojobori, *Evolution of vitamin B-6 (Pyridoxine) metabolism by gain and loss of genes.* Molecular Biology and Evolution, 2005. **22**(2): p. 243.
115. Rausher, M.D., R.E. Miller, and P. Tiffin, *Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway.* Molecular Biology and Evolution, 1999. **16**(2): p. 266.
116. Cork, J.M. and M.D. Purugganan, *The evolution of molecular genetic pathways and networks.* Bioessays, 2004. **26**(5): p. 479.
117. Vitkup, D., P. Kharchenko, and A. Wagner, *Metabolic flux and molecular evolution in a genome-scale metabolic network.* . Genome Biology, 2006. **7**(5): p. R39.