Gene Networks and Natural Selection: Is There a Network Biology?

Andreas Wagner

SFI WORKING PAPER: 2007-09-036

SANTA FE INSTITUTE

Gene networks and natural selection: Is there a network biology?

Dr. Andreas Wagner
Professor, Department of Biochemistry

University of Zurich
Dept. of Biochemistry, Bldg. Y27
Winterthurerstrasse 190
CH-8057 Zurich Switzerland

Office: +41-44-635-6141
FAX: +41-44-635-6144
Email: aw@bioc.unizh.ch
Web: http://www.bioc.unizh.ch/wagner/

Can qualitative information about large molecular networks inside cells teach us fundamentally new biology? In other words, is there a network biology (distinct from a network physics or network chemistry)? The answer to this question is important, because molecular networks are bridges between individual molecules, the lowest level of biological organization, and whole organisms. Both levels of organization, molecules and organisms, are intensely studied. Nonetheless, there is still an enormous gap in our understanding of how molecules collectively produce complex organismal phenotypes. Large molecular networks have the potential to fill this gap, because they contain most of the molecules that allow an organism to survive.

To find out whether large molecular networks can teach us new biology, we first need to answer a very basic question: Does natural selection influence the structure of biological networks, and if so, how? This question is key, because natural selection is the one central feature that distinguishes biological systems from all other, non-biological systems: Only biological systems have been shaped by natural selection, a process that acts on populations of organisms, and that requires heritable fitness differences among organisms.

I will here illustrate progress in answering this question with three examples. First, I discuss recent work suggesting that the function of metabolic networks influences the rates at which its constituent enzymes evolve. Second, I show how multiple small transcriptional regulation networks may have arisen through convergent evolution. These examples demonstrates how natural selection can influence the small-scale, local structure of biological networks. Third, I discuss a number of candidate cases for the influence of natural selection on large-scale network structure, cases that illustrate the great challenges ahead.

Efforts to understand the large-scale organization of living things are generating a wealth of information about how biologically important molecules interact to sustain life. An example regards the interactions of proteins with DNA to promote transcriptional regulation. Such interactions have been elucidated for hundreds of regulators and thousands of target genes using techniques such as chromatin immunoprecipitation (Lee, Rinaldi, Robert, Odom, Bar-Joseph, Gerber, Hannett, Harbison, Thompson, Simon et al. 2002). Another example regards metabolic reactions, the interactions of enzymes with small molecules to convert food into energy and biosynthetic building blocks. A variety of information – from genome sequence to biochemical data – has been used to generate complete maps of metabolic reactions in different model organisms (Edwards and Palsson 1999; Edwards and Palsson 2000a; Forster, Famili, Fu, Palsson and Nielsen 2003). A third example regards the physical interactions of proteins in protein complexes, which have been elucidated with techniques such as the yeast two-hybrid assay and affinity purification of tagged proteins (Gavin, Bosche, Krause, Grandi, Marzioch, Bauer, Schultz, Leutwein, Bouwmeester, Kuster et al. 2002; Uetz, Giot, Cagney, Mansfield, Judson, Knight, Lockshon, Narayan, Srinivasan, Pochart et al. 2000).

Such information has undoubtedly many useful applications both in basic biology and medicine. For example, it can lead to the elucidation of new biochemical and regulatory pathways. It could speed the development of new and specific drugs targeted to key players of a regulatory system.

However, the potential of such information goes far beyond these extensions of business-as-usual in molecular biology. The reason is that we now have information about the interactions of thousands of molecules, which form large interaction networks akin to man-made information processing networks and energy distribution networks. The structure of such networks itself may contain biological information, information that would be neither evident from studying individual molecules, nor from studying whole organisms. Such networks could thus teach us fundamentally new biology. They might help build a bridge between our understanding of the molecular machinery of life and whole organisms.

The analysis of networks is not only popular in biology. It has seen an explosion of activity in fields ranging from physics to the social sciences (Albert and Barabasi 2002). This explosion has been greatly helped by graph theory, which provides students of networks both with a common language and a common set of tools. In addition, the common language of graph theory has also greatly helped researchers cross discipline boundaries. Perhaps the most prevalent question in cross-disciplinary work has been whether networks in different disciplines share common features. The answer is a clear yes, which is intriguing, because it means that the objects of disparate disciplines share common organizational laws. In contrast, there has been much less effort to study network architectures distinguishing, say, biological networks from physical networks. Put differently, little effort has been made to create a network biology that might be different from a network physics, network chemistry, or network sociology.

What distinguishes biological systems from all other systems, including physical, chemical, sociological, and man-made systems? The answer is natural selection, selective persistence of individuals with heritable features coexisting in populations. Natural selection is an absolutely essential ingredient to the creation and persistence of complex living systems. In contrast, it is required for neither the creation nor the persistence of any physical, chemical, or social system. This simple observation provides a unique perspective on the question whether there is a network biology: If we want to find out whether biological networks have any features unique to them, features that they do not share with other kinds of networks, and thus features that may teach us new biology (as opposed to new physics or chemistry), we need to ask whether natural selection influences biological networks. This is easier said than done. Biology – and not only as practiced by physicists new to the field – is full of just-so stories, where an organism’s features are postulated to be shaped by natural selection without a shred of evidence other than plausibility. Nearly everybody can come up with such postulates, but relatively few are equipped to provide supporting evidence.

In sum, information about large-scale biological networks might teach us qualitatively new biology. If this is the case, if there are laws that characterize and distinguish biological networks, if there is a network biology, we better look for it by studying natural selection and its interaction with networks. In the following, I will illustrate three small steps in this direction, with different kinds of networks and on different levels of network organization. As will become obvious, natural selection influences the small-scale, local structure of networks. In contrast, on the larger, global scale of network organization, we have no lack of hypotheses but few solid answers. A new research program in evolutionary biology lies in wait here.

Natural selection and network parts.

In this section, I discuss recent work that suggests how natural selection may influence the evolution of the smallest components of a metabolic network, namely its enzymes. Importantly, this work underlines that an understanding of network function may be essential to understand this influence.

Every living thing is sustained by a complex network of thousands of chemical reactions. In heterotrophic organisms, these reactions transform food into energy and new building blocks for growth and reproduction. Complete (or nearly so) maps of core metabolism, comprising hundreds of reactions and metabolites are now available for several model organisms (Edwards and Palsson 1999; Edwards and Palsson 2000a; Forster, Famili, Fu, Palsson and Nielsen 2003). They can be used to study the structure, function, and evolution of large-scale metabolic networks. Some early work in this area focused merely on network structure, for example by characterizing one of a variety of graph representations of a metabolic network. Graphs (Figure 1a) are mathematical objects that consist of nodes and edges which connect neighboring nodes. (An example of a metabolic graph representation is a graph whose nodes are enzymes and metabolites, and where two nodes are connected if they participate in the same chemical reactions.) Such structural analysis, however, has one key limitation: It is poor at capturing the flow of matter through a metabolic network, which is at the heart of metabolic network function.

Metabolic network function can be computationally analyzed, even though information about enzymatic reaction rates in metabolic networks is very limited. Central to any such functional analysis are approaches such as flux balance analysis (Schilling, Edwards and Palsson 1999; Varma and Palsson 1993). Flux balance analysis uses information about the stoichiometry and reversibility of chemical reactions to determine the possible rates (fluxes) at which individual chemical reactions can proceed if fundamental constraints such as that of mass conservation are taken into consideration. Within the limits of such constraints, flux balance analysis can then be used to 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* (Edwards and Palsson 2000b; Forster, Famili, Fu, Palsson and Nielsen 2003; Segre, Vitkup and Church 2002). However, such prediction may fail if an organism has not been subject to natural selection to optimize growth in a particular environment.

Because natural selection clearly affects cell growth rates (the optimized codon usage bias of highly expressed genes in microbes is ample evidence), and because the distribution of metabolic fluxes in a network bears a direct relation to cell growth rates, one can ask how natural selection may influence the rate of evolution of the enzymes responsible for these fluxes. Vitkup and collaborators (Vitkup, Kharchenko and Wagner 2006) asked this question for the yeast metabolic network (Forster, Famili, Fu, Palsson and Nielsen 2003) where fluxes had been optimized to maximize growth. Specifically, they studied the relation between flux through individual enzymatic reactions and the ratio K_a/K_s, where K_s is the fraction of synonymous (silent) substitutions per silent nucleotide site in an enzyme-coding gene, and K_a is the fraction of amino acid replacement substitutions per replacement site. The ratio K_a/K_s ratio is typically (much) smaller than one and a good indicator of the evolutionary constraint a protein is subject to: Proteins that can tolerate very few amino acid substitutions will have a smaller K_a/K_s than proteins that can tolerate more such substitutions. They also accumulate fewer amino acid substitutions per unit time and thus have a smaller K_a.

Is flux through any one enzymatic reaction associated with the evolutionary constraint K_a/K_s on the corresponding enzyme? The answer is yes: There is a significant and negative association between flux and evolutionary constraint. That is, enzymes with high associated metabolic flux under optimal growth conditions can tolerate fewer amino acid changes. This association persists if one takes into account that many enzymatic reactions are carried out by several isoenzymes. And it is observed only for carbon sources such as glucose and fructose that are likely to be abundant in yeast's wild environment, but not for less relevant carbon sources, such as acetate. A potential confounding factor in such an analysis is that many enzymes with high associated flux are expressed at high levels, and high expression is known to be associated with slow evolution. However enzymes with high flux also evolve slowly if one corrects for differences in expression (Vitkup, Kharchenko and Wagner 2006).

Why do enzymes with high associated metabolic flux evolve more slowly? An immediately obvious hypothesis is this. Most amino acid changes will lead to a reduction in the catalytic rate at which an enzymatic reaction proceeds. Enzymatic reactions with high associated metabolic flux occur preferentially in central metabolic pathways, whose products are distributed to many peripheral pathways. In contrast, enzymatic reactions in peripheral pathways often show lower (but no less essential) metabolic fluxes. Thus, a reduction in reaction rate caused by an amino acid substitution may have a greater effect in a central pathway, where it may affect the substrates available to multiple peripheral pathways. It may thus reduce the output of multiple pathways, as opposed to a mutation in a peripheral pathway, which may change only the output of the affected pathway.

Even though the distinction between central and peripheral pathways in a complex network may not be sharp, the hypothesis is clear: reducing the flux through high-flux enzymes has a greater overall effect than reducing the flux through low-flux enzymes. This hypothesis produces a testable prediction: If correct, one would expect that mutations *increasing* flux through high-flux enzymes would also have a greater overall effect (except that these effects might be beneficial as opposed to deleterious). One class of such mutations are gene duplications, which potentially double the flux through an enzymatic reaction. If the hypothesis is correct, one would expect that such duplications can increase overall biomass production by a greater extent if they affect high-flux enzymes than if they affect low-flux enzymes. Because of the beneficial effects of increased biomass production, one would thus effect such mutations to become preferentially preserved in evolution. This is exactly what one observes, as we and others have shown (Papp, Pal and Hurst 2003; Vitkup, Kharchenko and Wagner 2006) (Figure 2).

In sum, the rates at which enzymatic genes evolve varies. Some of this variation may be explicable through differences in protein structure or through differences in expression level. However, some fraction of this variation is explicable through differences in metabolic flux under conditions likely to have been important in the evolutionary history of yeast. Importantly, the distribution of metabolic fluxes optimal for cell growth cannot be explained by studying one enzyme. It cannot even be explained by studying an entire metabolic pathway. It emerges as a property of an entire complex metabolic reaction network comprising hundreds of reactions. In this sense, the network influences the evolution of its parts. Having information on the whole network thus clearly matters in explaining the evolution of its parts.

Natural selection and small-scale, local network features. In this section, I will focus not on the smallest parts of a network but on the local neighbourhood of individual nodes. The network in question is that of transcriptional regulation in the yeast *Saccharomyces cerevisiae* and in *Escherichia coli*. The nodes are transcriptional regulators and their target genes. A transcriptional regulator is directly connected to a target gene (which may itself be a transcriptional regulator) if it regulates the target gene's transcription by binding to its regulatory region.

The work I discuss builds on recent studies that have identified small and highly abundant genetic circuit motifs in transcriptional regulation networks of the yeast *Saccharomyces cerevisiae* (Ho, Gruhler, Heilbut, Bader, Moore, Adams, Millar, Taylor, Bennett, Boutilier et al. 2002) and the bacterium *Escherichia coli* (Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii and Alon 2002; Shen-Orr, Milo, Mangan and Alon 2002). These circuit motifs include regulatory chains, feed-forward circuits, and a "bi-fan" circuit (see Figure 1b). There are two extreme possibilities for the evolutionary origin (and an entire spectrum of intermediates) of these circuits. First, these circuits may have come about through the duplication – and subsequent functional diversification – of one or a few ancestral circuits (Figure 3), that is, through the duplication of each of their constituent genes in a duplication event. This scenario is plausible, given the high frequency at which chromosome segments large and small, and even whole genomes undergo duplication (Dunham, Badrane, Ferea, Adams, Brown, Rosenzweig and Botstein 2002; Lynch and Conery 2000; Wolfe and Shields 1997). 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. It is ubiquitous on both the largest and the smallest levels of biological organization. For instance, eyes of similar basic design may have evolved multiple times independently; the wings of birds and bats have similar architectures, and both fish and whales – descended from land mammals – have similarly streamlined body shapes (Futuyma 1998). On the smallest scale, lysozymes in foregut-fermenting herbivores have independently evolved digestive functions in bovids, colubine monkeys such as languars, and a foregut-fermenting bird (Kornegay, Schilling and Wilson 1994; Stewart, Schilling and Wilson 1987)]. Another case in point are antifreeze glycoproteins, which have independently evolved similar amino acid sequences (Chen, DeVries and Cheng 1997) in antarctic notothenioid fish and northern cods. Evidence for convergent evolution is exciting to evolutionary biologists, because it can reveal the power of natural selection on an organismal feature.

A combination of genome and transcriptional regulation data allowed Gavin Conant and myself (Conant and Wagner 2003) to test the hypothesis that abundant circuits in trannscriptional regulation networks have evolved through convergent evolution. Specifically, we asked how frequent duplicated circuit pairs (Figure 3) that may have shared a common ancestor are, using available whole genome sequence information for yeast and *E. coli*. Information on the abundance of such pairs allowed us to develop measures of common ancestry that quantify how many circuits in a genome may have shared a common ancestor (Figure 4).

Overall, in an analysis of 20 circuit types (18 in yeast and 2 in E. coli) preciously few circuit pairs shared any common ancestor. Specifically, for 17 circuit types not even one pair of circuits were duplicates of each other. Even for the remaining three circuit types, the vast majority of circuit pairs show independent ancestry. For example, 44 out of 48 feed-forward loops in yeast show independent ancestry, a number that is no greater than that expected by chance alone due to the large number of single gene duplications that have occurred in the yeast genome.

In sum, this analysis suggested that highly abundant regulatory circuits in the transcriptional regulation network of two different organisms have arisen through convergent evolution (Conant and Wagner 2003). But 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 (Mangan and Alon 2003; Mangan, Zaslaver and Alon 2003; Shen-Orr, Milo, Mangan and Alon 2002). For example, a feed-forward loop may activate the regulated ('downstream') genes only if the upstream-most regulator is persistently activated. It can thus act as a filter for ubiquituous intracellular gene expression noise. Conversely, this type of circuit can rapidly deactivate downstream genes when the upstream-most regulator is shut off .

Taken together, such functional information, the high abundance of some small regulatory circuit motifs, and their independent evolutionary origin indicate the importance of natural selection for the evolution of genetic networks. Importantly for the purpose of this chapter, it shows that natural selection can shape and maintain the small-scale features of a large biological network.

Natural selection and global network structure

What is global network structure? An aspect of structure that cannot be reduced to characterizing individual nodes (proteins, genes) and their neighbors (Figure 1a). Examples include a network's degree distribution, that is, the distribution of each node's number of immediate neighbors. Although each node's degree is a local property, the degree distribution is a property of the whole network. Similar examples include pairwise correlations among node degrees, or the distribution of the number of edges linking any two nodes in a network. I will now discuss several candidate examples of natural selection on global network structure, all of them with important flaws, before outlining the lessons they provide.

Natural selection and the degree distribution. The connectivity or degree distribution of biological networks is often broad-tailed and sometimes consistent with a power-law. In a power-law degree distribution, the probability $P(d)$ that a node has d immediate neighbors is proportional to $P(d) \propto d^{-\gamma}$, γ being some constant. Among networks with a broad-tailed degree distribution are metabolic networks, where nodes can be enzymes or metabolites, depending on the chosen representation, protein interaction networks (Figure 5), where two nodes (proteins) are connected if they interact physically inside the cell, and other cellular networks (Jeong, Mason, Barabasi and Oltvai 2001; Rzhetsky and Gomez 2001; Wagner 2001; Wagner and Fell 2001; Wuchty 2001; Wuchty 2002).

Because broad-tailed degree distributions are abundant among biological networks, the question arises whether they exist for some biological reason. Are networks with this degree distribution better suited for some biological function than other networks? In other words, has their degree distribution been shaped by natural selection? If so, their degree distribution may reveal profound organizational principles of biological networks.

It has indeed recently been suggested that the degree distribution of biological networks is optimal (Albert, Jeong and Barabasi 2000; Jeong, Mason, Barabasi and Oltvai 2001; Jeong 2000). At the basis of this suggestion stands the following observation. The mean number of edges connecting pairs of nodes is very small and it increases only very little upon random node removal (Albert, Jeong and Barabasi 2000) in networks with a broad-tailed degree distribution. This feature – a network's mean path length – is a measure of network compactness. In networks with narrow-tailed degree distributions (e.g., a Poisson distributions), random node removal leads to a more substantial increase in mean path length, and a ready fragmentation into disconnected components. The robust compactness of networks with broad-tailed degree distributions, so the hypothesis goes, might be advantageous for biological networks.

The problem with this hypothesis starts with the question whether any such (unknown) advantage would be biologically sensible. For example, protein interaction networks are very heterogeneous mixtures of proteins in which some interact to provide structural support to a cell, others interact to transmit signals, and yet others interact to catalyze chemical reaction. It is thus not clear that such networks have one clearly defined biological function that needs to be preserved by keeping the network compact and connected.

A second potential problem has been explored explicitly for the chemical reaction networks defined by metabolism. Is it possible that the broad-tailed degree distribution of metabolic networks is a feature of many or all large chemical reaction networks, be they formed by natural selection or not? If so, then metabolic networks would join a large group of other networks whose broad-tailed degree distribution is not due to a benefit they provide. In support of this possibility, Gleiss and collaborators (Gleiss, Stadler, Wagner and Fell 2001) have studied the chemical reaction networks of planetary atmospheres, where atmospheric photochemistry determines network structure. The available data stems from multiple planets, including Venus, Jupiter, and Earth. The respective chemical reaction networks, despite having very different structure, have a broad-tailed degree distribution (Gleiss, Stadler, Wagner and Fell 2001). Such data suggests that broad-tailed degree distributions may be general features of chemical reaction networks, whether they exist in a living cell or not.

A third problem is an extension of the second one. Broad-tailed degree distributions are generally highly abundant in systems ranging from physics to sociology networks. In contrast, networks that do not have this degree distribution are rare. This observation further weakens the argument that such distributions might be shaped by natural selection and thus presumably specific to biological systems.

Finally, there is a growing number of mathematical models, grounded in empirical data, showing how the degree distribution of biological networks changes in evolutionary time (Berg 2004; Pagel, Meade and Scott 2007; Sole, Pastor-Satorras, Smith and Kepler 2002; van Noort, Snel and Huynen 2004; Wagner 2003). Although they differ in many details, these models demonstrate how power-law degree distributions can emerge in evolution, without natural selection having shaped this distribution. In sum, several lines of evidence speak against the possibility that broad-tailed degree distributions in biological networks have been shaped by natural selection.

Global features of transcriptional regulation networks. My next example concerns the abundance of regulatory cycles in signal transduction networks. Regulatory cycles or feedback loops are important for biological networks (Fell 1997; Savageau 1976). They may endow a gene or network with robustness to environmental change or intracellular noise. Alternatively, they make multistability – the adoption of multiple stable states – in a biological network possible. (Ferrell 2002; Freeman 2000).

Their biological importance makes cyclic structures good candidate features that might be under the influence of natural selection. Specifically, natural selection may influence the number and distribution of cycles in a biological network. One can find evidence for or against this hypothesis by asking whether cycles are more abundant in biological networks than one would expect by chance alone. Jeremiah Wright and I (Wagner and Wright 2005) asked this question for a variety of signal transduction networks, such as the c-Jun N-terminal MAPK network, and the B and T lymphocyte receptor signaling network, whose structure has been curated by experts and is publicly available (http://www.stke.org/cgi/cm/). We created from each of these networks 1000 randomized networks (Wagner and Wright 2005), and compared the number of cycles in the biological networks with the distribution of the number of cycles in the randomized networks. In general, exhaustive counting of cycles in large graphs may be prohibitive computationally, requiring alternative means to estimate the abundance of cycles (Gleiss, Stadler, Wagner and Fell 2001). However, because of the moderate size of our networks, we were able to enumerate the cycles in them.

An example of the distribution of cycle numbers is shown in Figure 6 for the B-cell antigen receptor network. The number of cycles in the actual network (two) is significantly smaller than in the randomized network (P=0.002). Overall, seven out of 15 signal transduction networks show a significantly smaller (!) number of cycles with length between two and ten than randomized networks, and one out of 15 networks showed more cycles than expected by chance alone.

As stated above, cycles may cause complex dynamical behavior in a network, in particular multistability (Ferrell 2002; Thomas and D'Ari 1990). On one hand, multistability may allow a cell to adopt a stable response to extracellular information. On the other hand, complex dynamical behavior and multistability may render a network more fragile to perturbations. It is thus tempting to hypothesize that networks with an underabundance of cycles have experienced a selective purging of cycles for this reason. However, no evidence beyond the significant underabundance of cycles currently speaks to this hypothesis.

Alternative pathways in transcriptional regulation networks. In a genome-scale transcriptional regulation network, a transcriptional regulator and its target molecule need not be immediate neighbors in the network. That is, the regulator may not regulate its target directly, but it may act through one or more intermediate regulators. If so, the path connecting the regulator and its target may be either unique or not (Figure 7). In the extreme, regulator-target pairs may be connected by large 'bundles' of alternative regulatory pathways. In the yeast transcriptional regulation network, such bundles of alternative pathways are not uncommon. For example, there are more than 100 regulator-target pairs connected by 10 or more alternative pathways (Wagner and Wright 2007). For such regulator-target pairs, perturbation of one of these pathways may be compensated through alternative regulatory routes. Jeremiah Wright and I (Wagner and Wright 2007) showed that this appears to be the case. Specifically, we compared the rate at which amino acid substitutions accumulate in intermediate regulators of pathways where regulators and targets are (i) connected by a single pathway or (ii) connected by a pathway bundle. Using the ratio K_a/K_s of the fraction of amino acid substitutions at amino acid replacement sites K_a to the rate of synonymous substitutions at synonymous sites K_s (Li 1997) as a measure of evolutionary rate, we found that the rate of evolution is significantly faster in intermediate regulators that are embedded in a bundle of pathways. (The same holds if one uses only K_a as a measure of amino acid divergence, or if one subdivides all regulator-target pairs according to the shortest distance of paths connecting them.) This means that in their evolutionary history intermediate regulators that are part of a pathway bundle have tolerated more amino acid change (most of which has deleterious effects on protein function). We were able to exclude the possibility that such regulators only evolve more rapidly because they are expressed (i) at lower levels of either mRNA or protein concentration, (ii) in a more limited spectrum of environmental conditions.

Taken together, this evidence suggests that alternative pathways between regulators and their targets can cause mutational robustness of the intermediate regulators. The distribution of the number of alternative pathways in a regulatory network is a global feature of the network. The key question for my present purpose is whether it has any adaptive significance. Has it been shaped by natural selection? And this is where an otherwise intriguing story collapses like a soufflé. We tested in two ways the hypothesis that the distribution of alternative pathways in the yeast transcriptional regulation network has adaptive significance. First, if so, the number of alternative pathways should be significantly different from that in randomized networks with the same degree distribution. This is not the case**.** For example, more than 50% of randomized networks have more regulator-target pairs with greater than 10, 20, or 30 connecting paths than the actual transcriptional regulation network. Second, if the hypothesis is correct, then many alternative pathways between regulator-target genes should exist for target genes that are "important" to the cell (as indicated by their strong deletion effects or slow rate of evolutionary change). Again, this is not the case. Specifically, there is no statistical association between the number of paths terminating at a target gene and its deletion effect or its rate of evolution K_a/K_s. We thus have no evidence to support the notion that the alternative pathway structure of the yeast transcriptional regulation network is shaped by natural selection.

Putting it all together. The first of the preceding sections has shown how natural selection can influence the rate at which enzyme-coding genes, the smallest parts of large-scale metabolic networks, evolve. Incidentally, it also illustrated how information on a biological network's function may be important in influencing evolutionary rates. (In other words, it shows that networks matter.) The second section showed how natural selection can shape small, local parts of a network's larger scale structure.

These are intriguing results but they leave the biggest challenge unmet. This challenge is to demonstrate that natural selection shapes the large-scale structure of biological networks. Arguably, we can speak only then of a network biology created by new and large-scale information about molecular interactions, a field of investigation distinct from a physics or chemistry of networks. (Small-scale features of networks have been accessible for a long time through the tools of conventional molecular biology.) Notice that this influence of natural selection on large-scale network features need not exist. In other words, natural selection may well shape the small, local structure of biological networks but not global network features. The earlier example of a postulated advantage – robust compactness – of the broad-tailed distribution of biological networks provides a case in point. As I argued, there are several lines of evidence speaking against the notion that it might be important specifically for biological systems, One of them is that such distributions are ubiquitous in all natural and man-made systems. In addition, an increasing number of mathematical models that can explain this and other large-scale features of biological networks without recourse to selection on any one feature.

The above example of the degree distribution is symptomatic of many others in the literature, where the influence of natural selection is postulated without prove. Just like the examples that followed (cycles in signal transduction networks, and alternative pathways in transcriptional regulation networks) it illustrates both the seductive appeal of such hypotheses – anybody could come up with one – and the difficulty in providing proof. In the signal transduction networks, where many networks we examined contain few cycles, the hypothesis is that cycles can cause complex dynamics and multistability, which may not always be desirable. However, without further evidence this is mere speculation. We cannot examine the regulatory dynamics of such networks, because we have incomplete information about their structure and no good quantitative models. In addition, not enough is known about their biology to *prove* that multistationarity is a bad thing. And not even our efforts to ask whether such cycles are significantly less abundant than expected "by chance alone" help much. The reason is that in order to do so we need to randomize network structure, and there we make an implicit assumption about the structure of networks (the randomized networks) that are not under the influence of natural selection. If this assumption was correct, we would be done. But there are many ways of generating randomized networks, and they may yield substantially different results (Artzy-Randrup, Fleishman, Ben-Tal and Stone 2004). We have no way of knowing which one yields the correct "null" network that we would expect if selection is absent. In the final example, the hypothesis is that the alternative pathways structure of transcriptional regulation networks has been shaped by natural selection. The availability of molecular evolution data in this example allows different routes of attack, which lead to the eventual failure of the hypothesis. The evidence is thus consistent with the notion that the robustness (as indicated by evolutionary constraint) of intermediate regulators is a mere consequence, not a cause, of the structure of this network. In sum, for the examples I discussed here, there is thus either evidence against the action of natural selection on global network features, or a lack of evidence. I know of no example where there is convincing positive evidence.

What information would we need to make the case for natural selection shaping large biological networks, and thus the case for a network biology? A look at a rich history of studies in both organismal and molecular evolution (Futuyma 1998; Li 1997) provides some answers. First, one needs a reference standard of how a network would evolve without the influence of natural selection, in order to compare it to existing networks. Such a standard has been available for a long time for molecules, through the neutral theory of molecular evolution (Kimura 1983). It is absent for networks. Second, by far the most successful approach in evolutionary biology is the comparative approach, which consists of comparing organismal features that are subject different selective pressures. Comparative data is conspicuously absent for networks. Chances are that we won't get far without it. And finally, the truly compelling cases for natural selection always use a variety of evidence, including comparative data and a functional characterization of the feature of interest. On this count also, we have a long way to go for genetic networks. The generation of this kind of information encompasses a research program that could take evolutionary biology to a completely new level, and that could build the long-sought bridge between molecules and living organisms.

Acknowledgments

This work was in part supported by SNF grant 315200-116814. I thank the Santa Fe Institute for its continued support.

Figure Captions

Figure 1: **a)** A simple graph comprised of 7 nodes. Neighboring nodes are connected by edges. The number of edges emanating from a node is known as the node's degree. For example, the grey node in the figure has degree 5. Nodes with especially high degree are also known as hubs. A simple way of characterizing a graph is through its degree distribution. **b)** Three examples of small transcriptional regulation circuits that are overabundant in genome-scale transcriptional regulation networks (Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii and Alon 2002; Shen-Orr, Milo, Mangan and Alon 2002).

Figure. 2. The horizontal axis shows the number of duplicates of enzyme-coding genes in the yeast genome. The vertical axis shows the metabolic flux through the respective enzyme-coding genes in aerobic growth on a minimal medium with glucose as the sole carbon source, as predicted by flux balance analysis. There is a highly significant positive association, that is, enzymes with high associated flux under these conditions have more isoenzymes. After Figure 9.6 in (Wagner 2005).

Figure 3: Hypothetical example of a duplicated gene circuit. Dashed lines connect paralogous genes, genes that arose in the duplication event that created the two feed-forward loops shown here.

Figure 4: Quantifying common circuit ancestry as illustrated with a hypothetical example of *n* =5 feed-forward loops. Each circuit is represented as a node in a graph. Nodes are connected if they are derived from a common ancestor, that is, if all *k* pairs of corresponding genes in the two circuits are duplicate genes. An index A of common ancestry can be defined that ranges from A=0 if no circuits share a common ancestor (left panel), to A≈1 if all circuits share one common ancestor (right panel). (Conant and Wagner 2003)

Figure 5: The left panel shows a graph representing the yeast protein interaction network as elucidated through high-throughput identification of interacting protein pairs (Uetz, Giot, Cagney, Mansfield, Judson, Knight, Lockshon, Narayan, Srinivasan, Pochart et al. 2000). The right panel shows the distribution of the number of neighbors each protein has. This distribution is heavy-tailed and consistent with a power-law.

Figure 6: The distribution of the number of regulatory cycles of length greater than 2 in 1000 randomized version of the B-cell antigen receptor network. After Figure 1a of (Wagner and Wright 2005)., which also describes in detail how randomization was carried out. The number of cycles in the actual network (two) is smaller than in all but a fraction P=0.002 randomized networks. This suggests a significant underabundance of such cycles and is consistent with the possibility that natural selection shapes these cycles.

Figure 7: Regulatory molecules and their regulatory targets can either be linked directly (not shown) or indirectly through only one regulatory path involving intermediate regulators (left panel) or many such paths (right panel). Which of these is the case may have consequences on the rate at which intermediate regulators evolve.

Literature Cited

Albert, R. and A.L. Barabasi. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74:** 47-97.

Albert, R., H. Jeong, and A.L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature ;* **406:** 378-382.

Artzy-Randrup, Y., S.J. Fleishman, N. Ben-Tal, and L. Stone. 2004. Comment on "network motifs: Simple building blocks of complex networks" and "superfamilies of evolved and designed networks". *Science* **305**.

Berg, J., Lassig, M., Wagner, A. 2004. Structure and evolution of protein interaction networks: a

statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology* **4:51**.

Chen, L.B., A.L. DeVries, and C.H.C. Cheng. 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences of the United States of America* **94:** 3811-3816.

Conant, G.C. and A. Wagner. 2003. Convergent evolution in gene circuits. *Nature Genetics* **34:** 264-266.

Dunham, M.J., H. Badrane, T. Ferea, J. Adams, P.O. Brown, F. Rosenzweig, and D. Botstein. 2002. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 16144-16149.

Edwards, J.S. and B.O. Palsson. 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *Journal of Biological Chemistry* **274:** 17410-17416.

Edwards, J.S. and B.O. Palsson. 2000a. 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* **97:** 5528-5533.

Edwards, J.S. and B.O. Palsson. 2000b. Robustness analysis of the Escherichia coli metabolic network. *Biotechnology Progress* **16:** 927-939.

Fell, D. 1997. *Understanding the control of metabolism*. Portland Press, Miami.

Ferrell, J.E. 2002. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology* **14:** 140-148.

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

Freeman, M. 2000. Feedback control of intercellular signalling in development. *Nature* **408:** 313-319.

Futuyma, D.J. 1998. *Evolutionary Biology*. Sinauer, Sunderland, Massachusetts.

Gavin, A.C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, C. Leutwein, T. Bouwmeester, B. Kuster et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Faseb Journal* **16:** A523-A523.

Gleiss, P.M., P.F. Stadler, A. Wagner, and D.A. Fell. 2001. Small cycles in small worlds. *Advances in complex systems.* **4:** 207-226.

Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415:** 180-183.

Jeong, H., S.P. Mason, A.-L. Barabasi, and Z.N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature* **411:** 41-42.

Jeong, H., Tombor, B. Albert, R. Oltvai, Z.N., Barabasi, A.L. 2000. The large-scale organization of metabolic networks. *Nature.* **407:** 651-654.

Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge.

Kornegay, J.R., J.W. Schilling, and A.C. Wilson. 1994. Molecular adaptation of a leaf-eating bird - stomach lysozyme of the hoatzin. *Molecular Biology and Evolution* **11:** 921-928.

Lee, T., N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799-804.

Li, W.-H. 1997. *Molecular Evolution*. Sinauer, Massachusetts.

Lynch, M. and J.S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science ;* **290:** 1151-1155.

Mangan, S. and U. Alon. 2003. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 11980-11985.

Mangan, S., A. Zaslaver, and U. Alon. 2003. The coherent feedforward loop serves as a sign-sensitive delay Element in transcription networks *Journal of Molecular Biology* **334:** 197-204.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: Simple building blocks of complex networks. *SCIENCE* **298:** 824-827.

Pagel, M., A. Meade, and D. Scott. 2007. Assembly rules for protein networks. *BMC Evolutionary Biology (in press).*

Papp, B., C. Pal, and L.D. Hurst. 2003. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends in Genetics* **19:** 417-422.

Rzhetsky, A. and S.M. Gomez. 2001. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics ;* **17:** 988-996.

Savageau, M.A. 1976. *Biochemical systems analysis: a study of function and design in molecular biology.* Addison-Wesley, Reading, MA.

Schilling, C.H., J.S. Edwards, and B.O. Palsson. 1999. Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnology Progress* **15:** 288-295.

Segre, D., D. Vitkup, and G. Church. 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the U.S.A.* **99**.

Shen-Orr, S., R. Milo, S. Mangan, and U. Alon. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics* **31:** 64-68.

Sole, R.V., R. Pastor-Satorras, E.D. Smith, and T. Kepler. 2002. A model of large-scale proteome evolution. *Advances in Complex Systems* **5:** 43-54.

Stewart, C.B., J.W. Schilling, and A.C. Wilson. 1987. Adaptive Evolution in the Stomach Lysozymes of Foregut Fermenters. *Nature* **330:** 401-404.

Thomas, R. and R. D'Ari. 1990. *Biological feedback*. CRC Press, Boca Raton, FL.

Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403:** 623-627.

van Noort, V., B. Snel, and M.A. Huynen. 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by asimple model. *EMBO reports* **5:** 280-284.

Varma, A. and B.O. Palsson. 1993. Metabolic capabilities of Escherichia coli. synthesis of biosynthetic precursors and cofactors. *Journal of theoretical biology.* **165:** 477-502.

Vitkup, D., P. Kharchenko, and A. Wagner. 2006. Metabolic flux and molecular evolution in a genome-scale metabolic network. . *Genome Biology* **7:** R39.

Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution.* **18:** 1283-1292.

Wagner, A. 2003. How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society of London Series B.* **270:** 457-466.

Wagner, A. 2005. *Robustness and evolvability in living systems.* Princeton University Press, Princeton, NJ.

Wagner, A. and D. Fell. 2001. The small world inside large metabolic networks. *Proc. Roy. Soc. London Ser. B* **280:** 1803-1810.

Wagner, A. and J. Wright. 2005. Compactness and cycles in signal transduction and transcriptional regulation networks: a signature of natural selection? *Advances in Complex Systems* **7:** 419-432.

Wagner, A. and J. Wright. 2007. Alternative pathways and mutational robustness in molecular networks. *BioSystems (in press).*

Wolfe, K.H. and D.C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708-713.

Wuchty, S. 2001. Scale-free behavior in protein domain networks. *Molecular Biology and Evolution* **18:** 1694-1702.

Wuchty, S. 2002. Interaction and domain networks of yeast. *Proteomics* **2:** 1715-1723.

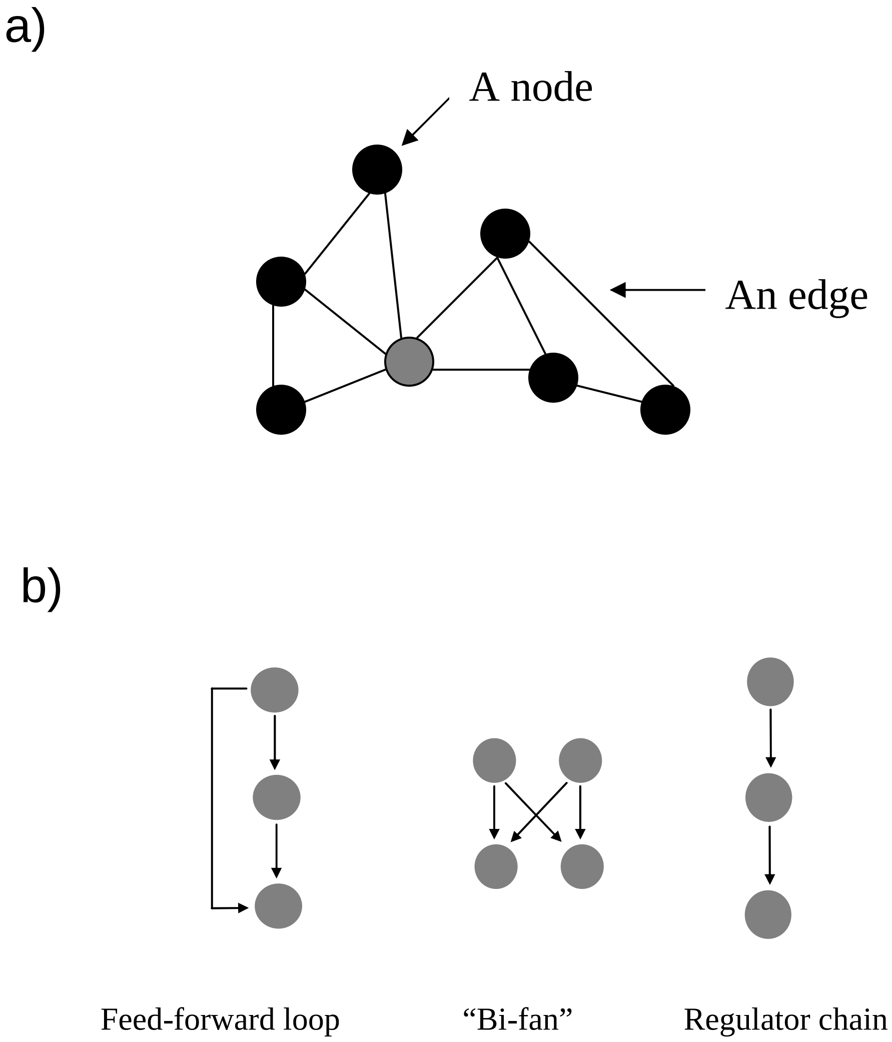


Figure 1

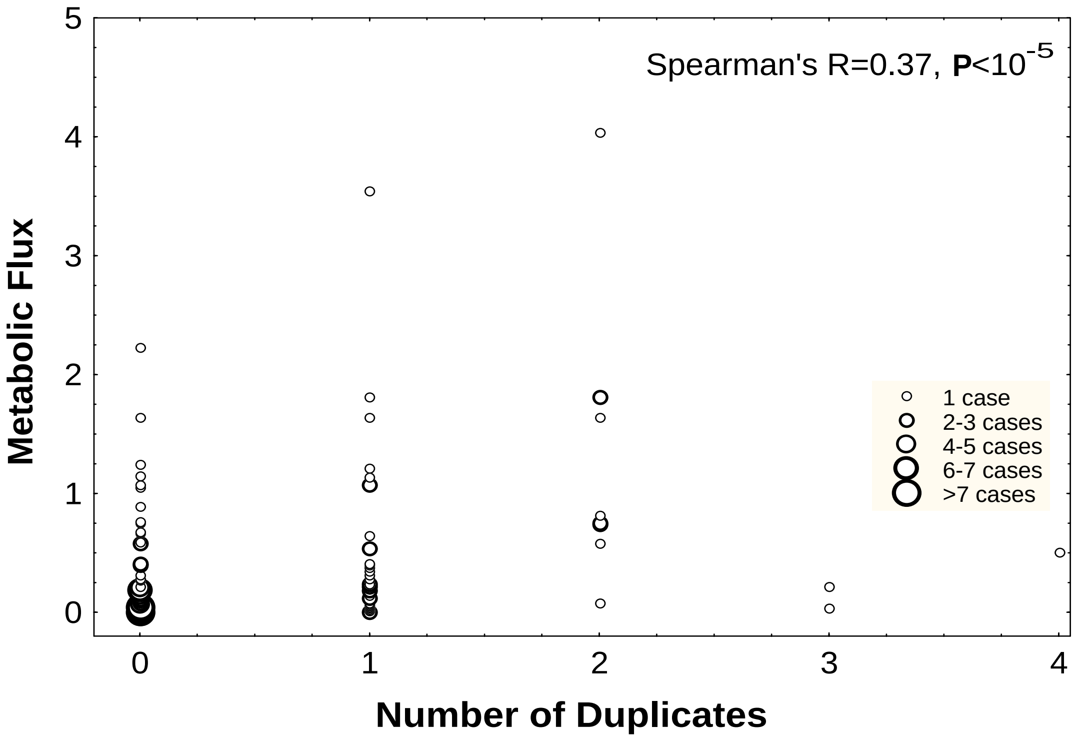
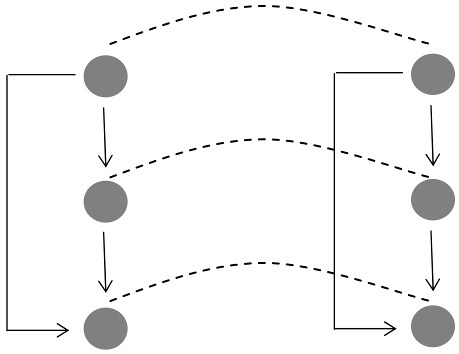

Spearman's R=0.37, **P**<10^{-5}
Metabolic Flux
Number of Duplicates
5
4
3
2
1
0
0
1
2
3
4
1 case
2-3 cases
4-5 cases
6-7 cases
>7 cases

Figure 2



Duplication of a gene circuit

Figure 3

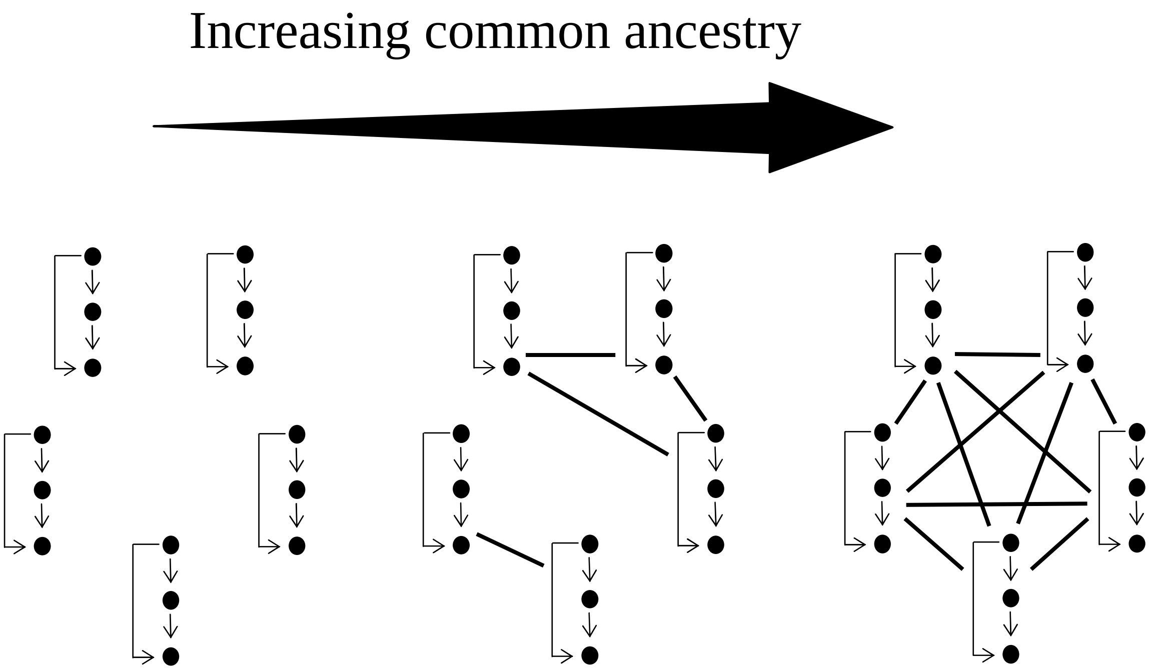


Figure 4

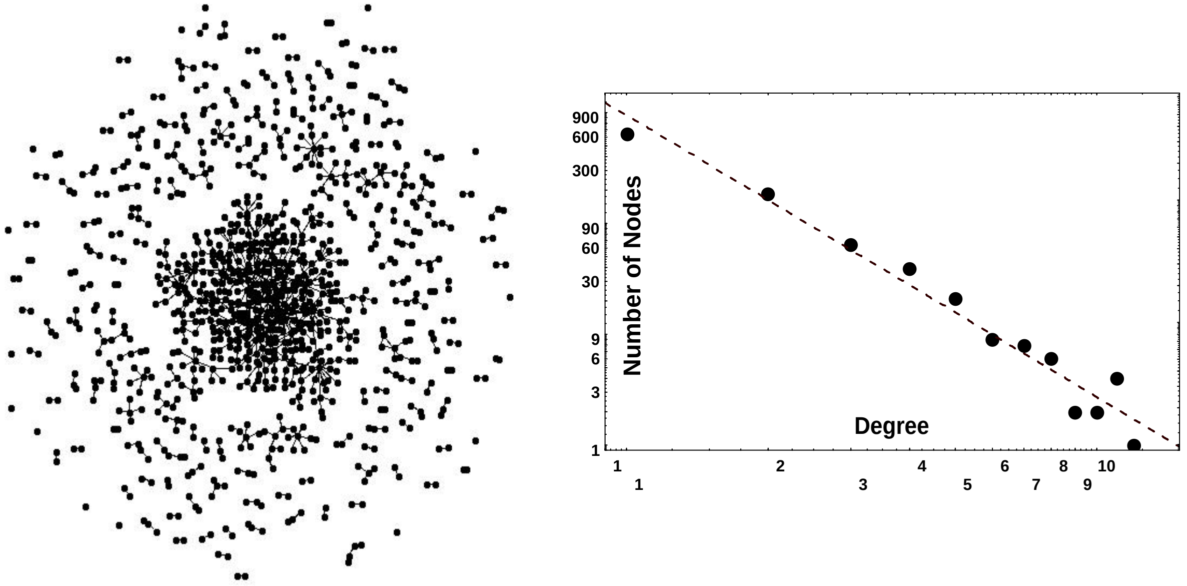

900
600
300
90
60
30
9
6
3
1
Number of Nodes
Degree
1
2
4
6
8
10
1
3
5
7
9

Figure 5

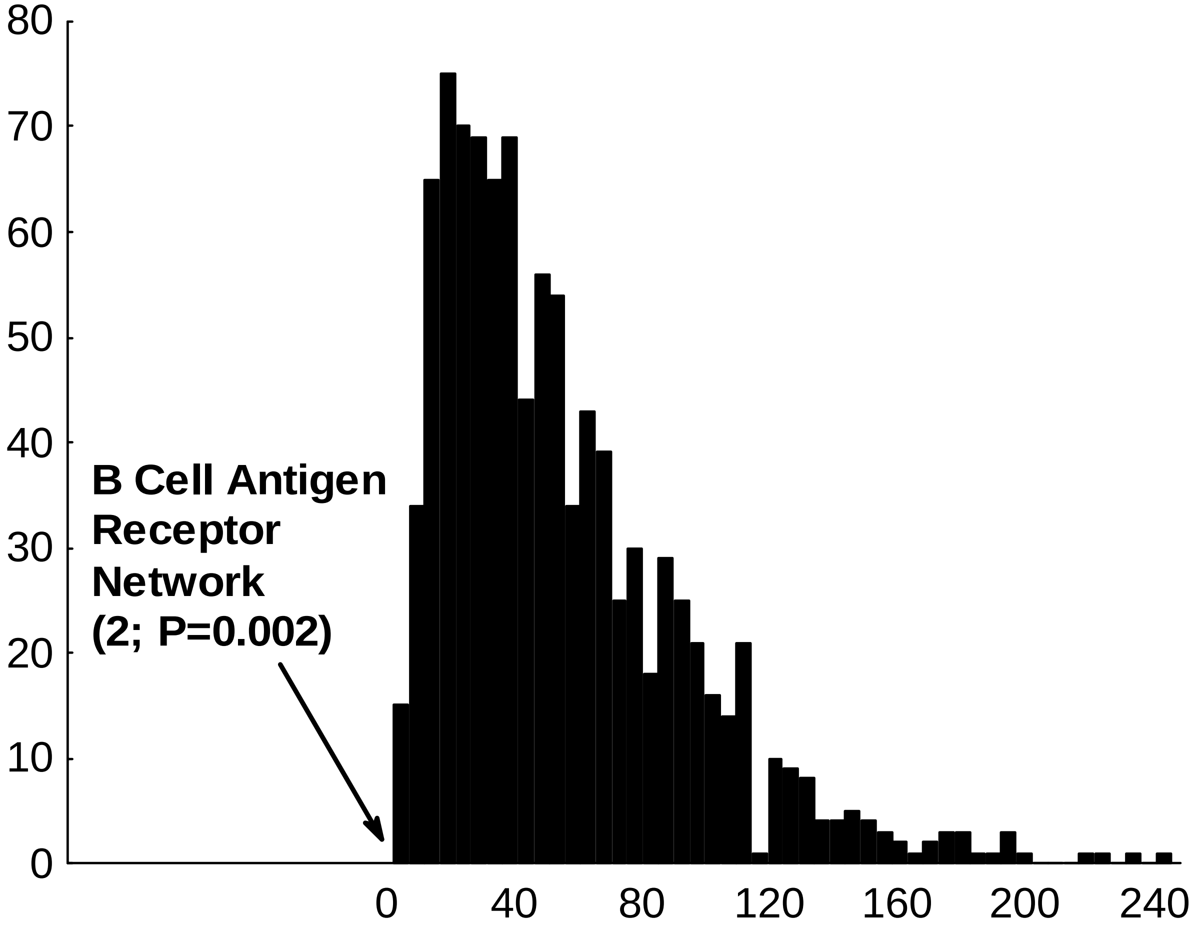
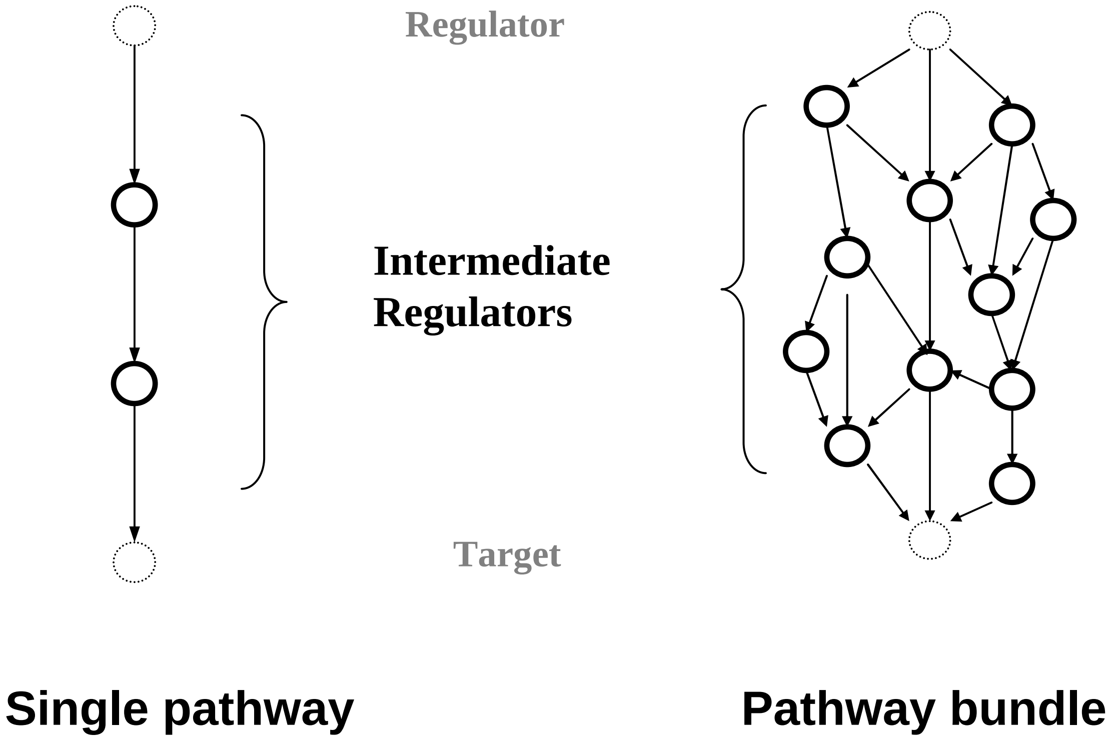

80
70
60
50
40
30
20
10
0
0
40
80
120
160
200
240
B Cell Antigen
Receptor
Network
(2; P=0.002)

Figure 6



Regulator
Intermediate Regulators
Target
Single pathway
Pathway bundle

Figure 7