Analysis of Coevolving Gene Families Using Mutually Exclusive Orthologous Modules

Xiuwei Zhang[1,2], Martin Kupiec[3], Uri Gophna[3], and Tamir Tuller[*,4,5]

[1]Laboratory for Computational Biology and Bioinformatics, School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Israel

[4]Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel

[5]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

Zhang and Tuller have contributed equally to this work.

*Corresponding author: E-mail: tamirtul@post.tau.ac.il.

Accepted: 29 March 2011

Abstract

Coevolutionary networks can encapsulate information about the dynamics of presence and absence of gene families in organisms. Analysis of such networks should reveal fundamental principles underlying the evolution of cellular systems and the functionality of sets of genes. In this study, we describe a new approach for analyzing coevolutionary networks. Our method detects Mutually Exclusive Orthologous Modules (MEOMs). A MEOM is composed of two sets of gene families, each including gene families that tend to appear in the same organisms, such that the two sets tend to mutually exclude each other (if one set appears in a certain organism the second set does not). Thus, a MEOM reflects the evolutionary replacement of one set of genes by another due to reasons such as lineage/environmental specificity, incompatibility, or functional redundancy. We use our method to analyze a coevolutionary network that is based on 383 microorganisms from the three domains of life. As we demonstrate, our method is useful for detecting meaningful evolutionary clades of organisms as well as sets of proteins that interact with each other. Among our results, we report that: 1) MEOMs tend to include gene families whose cellular functions involve transport, energy production, metabolism, and translation, suggesting that changes in the metabolic environments that require adaptation to new sources of energy are central triggers of complex/pathway replacement in evolution. 2) Many MEOMs are related to outer membrane proteins, such proteins are involved in interaction with the environment and could thus be replaced as a result of adaptation. 3) MEOMs tend to separate organisms with large phylogenetic distance but they also separate organisms that live in different ecological niches. 4) Strikingly, although many MEOMs can be identified, there are much fewer cases where the two cliques in the MEOM completely mutually exclude each other, demonstrating the flexibility of protein evolution. 5) CO dehydrogenase and thymidylate synthase and the glycine cleavage genes mutually exclude each other in archaea; this may represent an alternative route for generation of methyl donors for thymidine synthesis.

Introduction

Coevolution of genes is an important force constraining the evolution of genes, proteins, and other cellular features and is useful for predicting physical and functional interactions (Chen and Dokholyan 2006; Juan et al. 2008).

A natural way of representing coevolutionary relations is by "coevolutionary networks" (Dagan et al. 2008; Lima-Mendez et al. 2008; Halary et al. 2009; Tuller et al. 2009). In the case of coevolution of genes, each node in a coevolutionary network corresponds to a gene family (e.g., a Cluster of Orthologous Groups [COGs] of proteins[Tatusov et al. 2003]), and pairs of nodes are connected by an edge if they tend to appear in the same organisms. In some cases, the network can also encompass information about "anti-occurrence," that is, proteins or cliques of proteins that tend "not to appear" in the same organisms are connected with a different type of edge (Tuller et al. 2009).

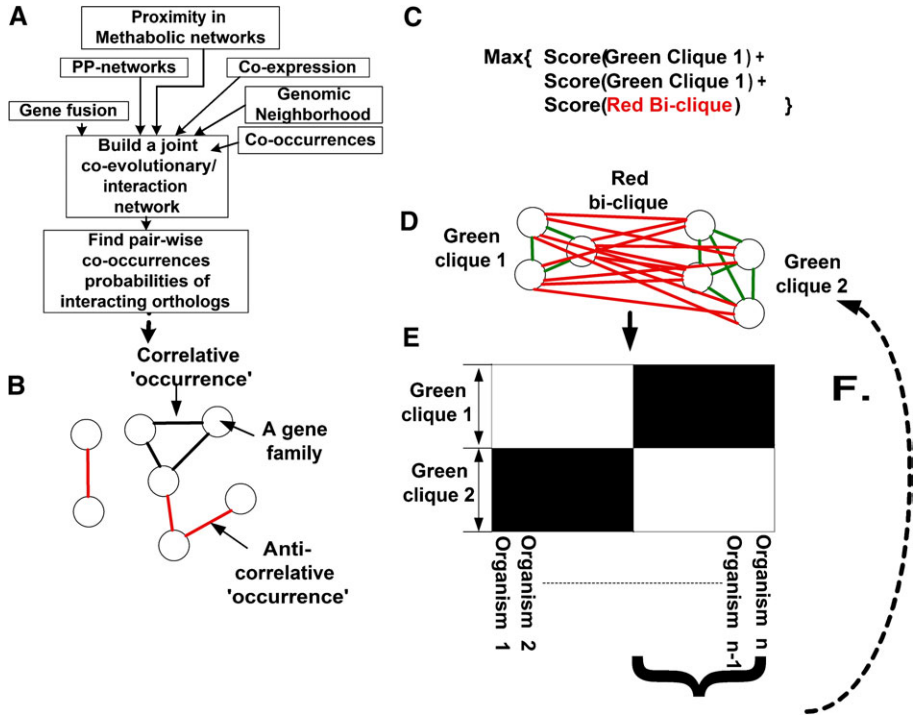


Fig. 1.—Finding MEOMs in coevolutionary networks: a flow diagram. (*A*) A coevolutionary network is reconstructed based on biological (e.g., physical interaction, coexpression, etc.) and statistical evidence. (*B*) In a coevolutionary network, each node corresponds to a gene family; green edges denote co-occurrence of the two nodes in the same organisms; red edges denote mutually exclusive occurrence (i.e., occurrence in different organisms). (*C*) The optimization score of a MEOM was based on Kelley and Ideker (2005). (*D*) Each MEOM is composed of two green (close to) cliques that are connected by a red (close to) bi-clique; in practice, the green (red) cliques (bi-cliques) may not be perfect (i.e., some of the edges may be missing; Materials and Methods). (*E*) Projection of a MEOM on a (gene family) × (organism) table; black denotes a case where there is a gene from the gene family in the organism; white denote cases where no gene from the gene family appears in the organism. (*F*) The approach can be implemented iteratively on subtaxonomic/phylogenetic groups.

Previous systems biology studies have mainly focused on other types of biological networks such as protein interaction networks (see, e.g., Kelley et al. 2003; Sharan et al. 2005), genetic interaction networks (Kelley and Ideker 2005; Ulitsky and Shamir 2007), and regulatory networks (e.g., Milo et al. 2004; Hershberg et al. 2005) to understand cellular systems. One important difference between coevolutionary networks and other biological networks is the fact that coevolutionary networks encapsulate information about the evolutionary dynamics of cellular systems, whereas the other networks provide a snapshot of a particular cellular system at a certain time point (i.e., in a specific organism). Thus, by analyzing coevolutionary networks, it should be possible to detect biological phenomena that cannot be detected by analyzing other biological networks.

In this study, we describe a new computational approach for analyzing coevolutionary networks. The aim of our method is to find Mutually Exclusive Orthologous Modules (MEOMs)—pairs of sets of gene families such that each set tends to appear in the same organisms, but the two sets usually do not co-occur (see fig. 1). These sets may describe cases of lineage specificity, where only one divergent lineage can benefit from a particular functional set. Another explanation can be incompatibility—cases where different cellular pathways cannot physically coexist in the same organisms. Alternatively, they may simply represent functional streamlining where the loss of one cellular subfunction can be either due to redundancy or an evolutionary consequence of different ecological constraints.

Materials and Methods

Information about Taxonomy

This information was downloaded from ftp://ftp.ncbi.nih.gov/pub/taxonomy/.

Obtaining the Input Network

The nodes in the analyzed networks correspond to the 4,870 COGs (Tatusov et al. 2003) in a large data set of organisms that will be discussed later.

The coevolutionary networks used in this study were based on two sources of information: 1) Evidence of co-functionality or physical interactions (e.g., the String

databases mentioned below). 2) Statistical evidences related to co-occurrence of pairs of COG in the analyzed genomes.

We put edges between pairs of COGs only if they satisfy these two conditions.

This is a generic and "modular" approach: that is, it is possible to replace the first source of information with a different one. It is also possible to use only the second source of information. Similar ideas were used in Tuller, Birin, et al. (2009) with very good results.

Specifically, the coevolutionary edges were based on information downloaded from the String database (Jensen et al. 2009). This database includes information on genomic neighborhood, coexpression, gene fusion, and more. Each coevolutionary relation in this data set was based on a composite score, which is a weighted average of these sources of information. The initial number of coevolutionary relations downloaded from this database was 962,618. For a coevolutionary edge to be included in the network, it should also satisfy additional statistical requirements:

Let pr00 denote the empirical probability that a pair of orthologs does not appear in the same organism; let pr01 denote the probability that the first ortholog does not appear while the second does; let pr10 denote the probability that the first ortholog appears, whereas the second does not appear; finally, let pr11 denote the probability that the two orthologs appear in the same organism. Let "green" denote a positive co-occurrence relation between the two orthologs and "red" denote a negative co-occurrence relation. We considered only significant edges that exhibit

(A) a strong coevolutionary relation, and
(B) the pattern of occurrence/absence of the pair of proteins is variable: In the case of green edges, there are organisms with both proteins but also organisms with neither. In the case of red edges, there are genomes encoding one of the proteins but not the other, but there are also cases in which the presence/absence pattern of these proteins is the opposite (the second protein is encoded in the genome but the first one is not).

Statistically, pairs of orthologs with stronger condition (B) are more surprising as they correspond to cases where there have been more evolutionary changes in the presence/absence pattern of each of the proteins/orthologs but the changes of the two proteins were correlated.

Specifically, we considered the following conditions ($k1$ and $k2$ are two thresholds): To satisfy condition A:

1. Candidates for green edges: calculate $g1 = (\text{pr00} + \text{pr11})/(\text{pr01} + \text{pr10})$ for each edge and choose the first $(1 - k1) \times 962{,}618$ edges which have the highest $g1$ value.
2. Candidates for red edges: calculate $r1 = (\text{pr01} + \text{pr10})/(\text{pr00} + \text{pr11})$ for each edge and choose the first $(1 - k1) \times 962{,}618$ edges which have the highest $r1$ value (these edges also have the lowest g1 value).

To satisfy condition B:

3. For each potentially green edge, calculate an additional value: $g2 = \max(\text{pr00}, \text{pr11})/\min(\text{pr00}, \text{pr11})$; choose only the edges that have $g2 < k2$ to be green.
4. For each potentially red edge, calculate an additional value: $r2 = \max(\text{pr01}, \text{pr10})/\min(\text{pr01}, \text{pr10})$; choose only the edges who have $r2 < k2$ to be red.

The Small Network. This graph corresponds to a set of 95 bacteria, archaea, and eukaryotes downloaded from Tuller, Birin, et al. (2009). By letting $k1 = 95\%$ and $k2 = 10$, we obtain a network with 24,944 edges and 2,421 different orthologs. Specifically, it includes 5,242 red edges and 19,702 green edges.

The Large Network. This graph corresponds to the gene content of 383 bacteria and archaea and eukaryotes. These data were downloaded from National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/genomes/) and based on the data set from Tuller et al. (2011); the data were analyzed hierarchically. The hierarchical partitioning of the data was based on NCBI taxonomy (ftp://ftp.ncbi.nih.gov/pub/taxonomy/).

At the first stage, we analyzed the entire data set by letting $k1 = 90\%$ and $k2 = 10$, we obtain a network with 31,451 edges and 2,324 different orthologs. Specifically, it includes 3,655 red edges and 27,796 green edges.

At the next stage, we analyzed eukaryotes (18 organsims; 68,527 green edges, 1,444 red edges, and 1,095 COGS), archaea (27 organsims; 672 green edges, 1,456 red edges, 1,047 COG), and bacteria (340 organsims; 15,836 green edges and 1,083 red edges, 1,855 COGs) separately.

At the next stage, we analyzed a few large bacterial subgroups (groups with more than 20 organisms):

Proteobacteria (199 organisms; 8,962 green edges, 490 red edges, and 1,405 COGs), Firmicutes (59 organisms; 28,021 green, 522 red edges, and 1,235 COGs), Actinobacteria (26 organisms; 6,923 green, 490 red edges, and 1,127 COGs; we used here $k1 = 0.85$).

Scoring MEOMs

We used an algorithm that is based on Ulitsky and Shamir (2007). The optimization score was based on Kelley and Ideker (2005). Let $G^g(V, E^g)$ be the network induced when considering only the positive coevolutionary relations (the green edges); let $G^r(V, E^r)$ be the network induced when considering only the reciprocal coevolutionary relations (the red edges). The nodes, V, in these networks are the set of genes families (COGs). A MEOM is a pair of disjoint sets V_1, V_2, such that (a) $/V_2/, /V_1/ \geq 2$; (b) For each V_i, there are unusually many

green edges between V_i; and (c) there are unusually many red edges between V_1 and V_2 (see fig. 1).

To quantify property (b), we derive a log-odds score reflecting the likelihood that the density of green edges between nodes in V_1 or in V_2 is unusually high. We compare two hypotheses: under the MEOM hypothesis, every pair of genes, one from V_i ($i = 1$ or $i = 2$), has a coevolutionary relation (green edges) with a high probability α, independently of all other gene pairs, and the likelihood of a model V_i is thus $\prod_{(a,b)\in(V_i\times V_i)} \alpha.I(a,b) + (1-\alpha)(1-I(a,b))$, where $I(a,b)$ equals 1 if there exists a red edge between a and b and otherwise it equals 0; in the null hypothesis, every pair (a, b) is connected with probability $r(a,b)$, representing the chance of observing this interaction at random, given the degrees of a and b in G^g. We estimate $r(a,b)$ by generating a random ensemble of networks with the same degree sequence and counting what fraction of them contain an interaction between a and b. The log-odds score is then

$$S^g(V_i) = \log \frac{\prod_{(a,b)\in(V_i\times V_i)} \alpha.I(a,b) + (1-\alpha).(1-I(a,b))}{\prod_{(a,b)\in(V_i\times V_i)} r(a,b).I(a,b) + (1-r(a,b)).(1-I(a,b))}.$$

Similarly, to quantify property (c), we derive a log-odds score reflecting the likelihood that the density of red edges between V_1 and V_2 is unusually high. We compare two hypotheses: under the MEOM hypothesis, every pair of genes, one from V_1 and the other from V_2, have reciprocal coevolutionary relation (red edges) with a high probability β, independently of all other gene pairs, and the likelihood of a model (V_1, V_2) is thus $\prod_{(a,b)\epsilon(V_i\times V_i)} \beta I(a,b) + (1-\beta)(1-I(a,b))$, where $I(a,b)$ equals 1 if there exists a red edge between a and b and otherwise it equals 0; in the null hypothesis, every pair (a, b) is connected with probability $r(a,b)$, representing the chance of observing this interaction at random, given the degrees of a and b in G^g. We estimate $r(a,b)$ by generating a random ensemble of networks with the same degree sequence and counting what fraction of them contain an interaction between a and b. The log-odds score is then.

$$S^r(V_1,V_2) = \log\left(\frac{\prod_{(a,b)\epsilon(V_i\times V_i)} \beta.I(a,b) + (1-\beta).(1-I(a,b))}{\prod_{(a,b)\epsilon(V_i\times V_i)} r(a,b).I(a,b) + (1+r(a,b)).(1+I(a,b))}\right).$$

The final score was $S^r(V_1,V_2) + S^g(V_1) + S^g(V_2)$. The aim was to find subnetworks whose score is as large as possible. We filtered subnetworks whose probability in randomized networks with similar degree distribution (see below how these networks were computed) was $>$ 0.05. To assess the statistical significance of the results, we performed the two randomization tests described below:

Two Network Randomization Tests. In this paper, we report two randomization tests.

In the first randomization test, we generated for each data set a random ensemble of networks (a total of 20 networks) with the same degree distribution of red and green edges (see, e.g., Tuller, Kupiece, and Ruppin 2009). We counted what fraction of them contained more MEOMs than the original network.

The second randomization test was similar to the one reported in Lima-Mendez et al. (2008). In this randomization test, we generated a total of 20 networks where we randomly assigned COGs to each organism maintaining the total number of COGs in the organism. In this case, we used $k1$ such that the number of edges in the random network will be similar to the number of edges in the original network.

Jackknifing. Jackknifing (see, e.g., Shao and Tu 1995) was performed as described below.

Repeat 100 times:

1. Randomly choose 90% of the edges in the network (STRING).
2. Run the algorithm to find MEOMs.

Count for each MEOM the number of times it appears (Jaccard index [Jaccard 1912] $>$ 0.5 corresponding to the nodes in the two MEOMs) in resultant random networks.

MEOMs' Separation Versus Phylogenetic/Environmental Proximity

To show that the separation of MEOMs corresponds both to phylogentical proximity and to environmental proximity, we used the following definitions:

A pair of organisms is separated by a MEOM if one of them has most of the COGs (i.e., more than 50%) of the first cluster but does not have most of the genes (COGs) of the second cluster, whereas the second organism have most of the genes of the second cluster but does not have most of the genes of the first cluster. The MEOMs' separation score for a pair of organisms is the number of times these two organisms are separated by a MEOM. Evolutionary proximity between two organisms was defined as the number of nodes separating the two taxa in NCBI taxonomy tree (see, e.g., Farris 1969). A pair of organisms was defined to share environments if it appears in the same community according to Freilich et al. (2010).

To control for the size of the genomes of the analyzed organisms, we computed for each pair of genomes the

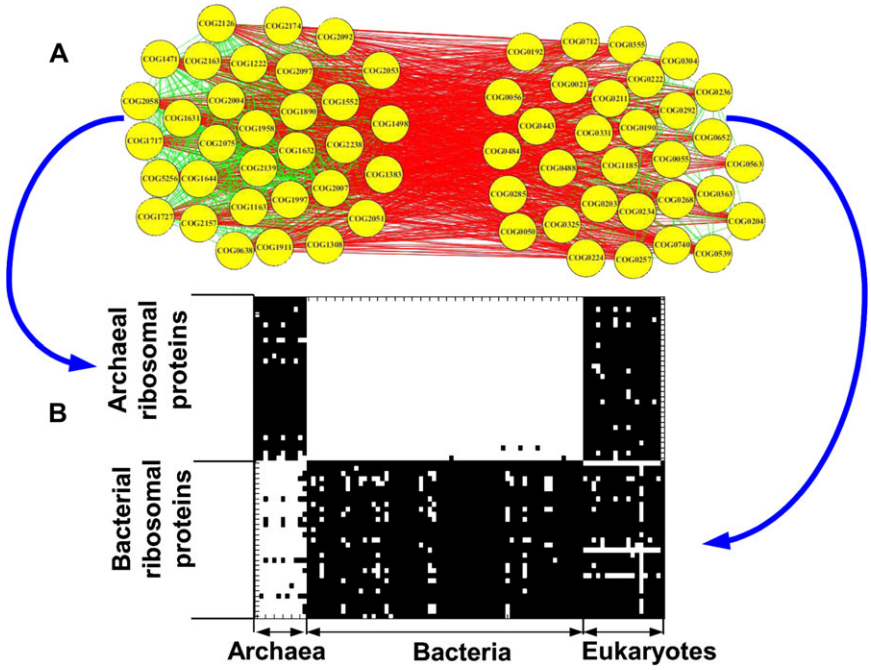


Fig. 2.—The most striking MEOM detected in data set of 95 organisms. (*A*) The coevolutionary edges between the gene families that are part of the MEOM. (*B*) The solution splits the analyzed organisms to the three domains of life (bacteria, archaea, and eukarya).

mean number of COGs in that pair. We performed nonparametric multivariate analysis (see details below) where we computed the following correlations 1) Correlation between MEOM separation and "phylogenetic" proximity when controlling for "ecological" proximity and the size of the genomes together. 2) Correlation between MEOM separation and ecological proximity when controlling for phylogenetic proximity and the size of the genomes together. The correlations were 0.53 and −0.082, respectively ($P < 10^{-16}$ in both cases).

Nonparametric Multivariate Analysis. Let X and Y denote two variables and $Z = [Z1, Z2, Z3, \ldots]$ denote a set of variables. The nonparametric multivariate analysis that is reported in this paper includes partial Spearman correlations of the form $R(X, Y|Z)$. Roughly, if such a correlation is significant it means that there is a relation between X and Y that cannot be explained by the variables in Z.

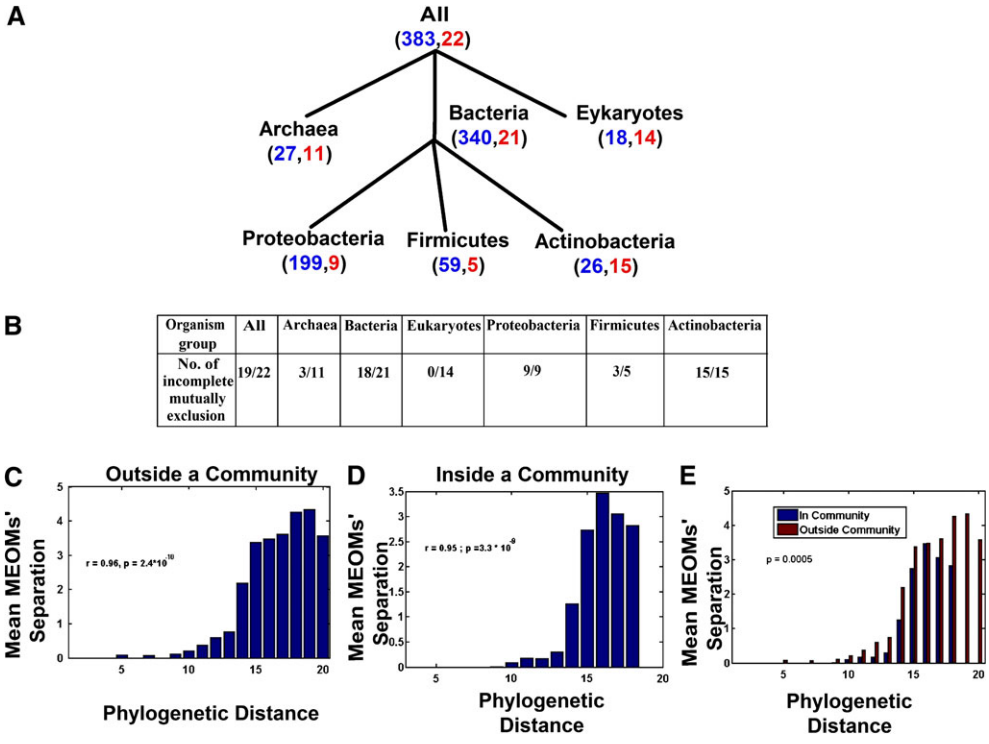
Results and Discussion

We analyze two coevolutionary networks, one that corresponds to 95 organisms and was previously used (Tuller et al. 2009) to reconstruct ancestral gene content and a new coevolutionary network extracted from 383 microorganisms (see Materials and Methods); the smaller data set is based on the commonly used COG database (Tatusov et al. 2003) and is a subset of the second data set.

We performed the following general steps to obtain MEOMs: First, based on physical and statistical evidence, we generated a coevolutionary network of 383 organisms and 4,870 COGs (fig. 1*A* and *B*; Materials and Methods). In this network, we found MEOMs (two green cliques that are connected by a red bi-clique, that is, a highly connected component of red edges; fig. 1*C*–*E*; Materials and Methods) and analyzed the COGs and organisms corresponding to each MEOM. This procedure was implemented hierarchically: in each step, we divided the set of organisms to subtaxonomical groups and applied our method to find MEOMs in each of these subsets separately (fig. 1*F*).

MEOMs Divide the Three Domains of Life According to Ribosomal and Proteasomal Proteins

To demonstrate our approach, we began with a relatively small data set (95 species of bacteria, archaea, and eukaryotes). The most striking MEOM found by our method appears in figure 2 (see supplementary table 1, Supplementary Material online, for the list of all nine MEOMs that were found for this data set). One of the green cliques in this MEOM corresponded to a set of ribosomal proteins that appear exclusively in Bacteria, whereas the second corresponded to the ribosomal proteins that only appear in Archaea (see fig. 2). Indeed, since bacterial and archaeal ribosomes differ significantly, many of the ribosomal proteins should appear either in Archaea or in Bacteria, they mutually excluding each other (Roberts et al. 2008). In addition, it was shown that the ribosomal proteins undergo less horizontal gene transfer (HGT) than other protein families (Cohen and Pupko 2009). In



Organism group	All	Archaea	Bacteria	Eukaryotes	Proteobacteria	Firmicutes	Actinobacteria
No. of incomplete mutually exclusion	19/22	3/11	18/21	0/14	9/9	3/5	15/15

Fig. 3.—General properties of MEOMS. (*A*) The subgroups that were analyzed from a data set of 383 organisms, the number of organisms in each subgroup (blue) and the number of MEOMs that were detected in each subgroup (red). (*B*) Number of incomplete mutually exclusions in each of the analyzed group of organisms (out of the total number of MEOMs detected). (*C*–*E*) MEOMs reflect both lineage specificity and environmental changes. (*C*) Mean MEOMs' separation increases with phylogenetic distance (Materials and Methods) for pair of organisms that are not in the same environment (Spearman correlation $r = 0.96$; $P = 2.4 \times 10^{-10}$). (*D*) Mean MEOM separation increases with phylogenetic distance for pair of organisms that are in the same environment (Spearman correlation $r = 0.96$; $P = 3.3 \times 10^{-9}$). (*E*) For every phylogenetic distance, pairs of organisms that are in the same community (Materials and Methods) have lower mean MEOM separation than pairs of organisms that are in different communities (Wilcoxon test: $P = 0.0005$).

contrast, the only domain of life where both sets of ribosomal proteins appear is Eukarya—the archaeal ribosome is similar to the cytosolic eukaryotic ribosome, whereas the bacterial ribosome is similar to the mitochondrial eukaryotic ribosome (see fig. 2). Reassuringly, this is exactly the result obtained in the ribosomal MEOM in Eukarya (fig. 2). Thus, although our method is unsupervised and does not assume any phylogenetic tree, it accurately divided the organisms in the data set into the three domains of life and identified the most central and evolutionarily conserved complex (the ribosome) that separates these organisms.

Another important complex that is specific to the archaeal/eukaryotic branch of the tree of life is the proteasome (the notable exception being *Mycobacterium tuberculosis* [Lin et al. 2006]). Indeed, several proteasome subunits are also domain-specific MEOMs. Curiously, two proteins of unknown function, a GTPase and a putative transcription factor, appear as part of the eukaryotes/archaea-specific MEOM set. We can therefore speculate that it is highly likely that they represent some yet unidentified component of either the translational apparatus or proteasome-mediated degradation.

Analysis of a Data set of 383 Species of Bacteria, Archaea, and Eukaryotes Reveals the Systems Biological Properties of MEOMs

To better obtain general evolutionary observations, we extended our approach to analyze a larger data set of 383 microorganisms, better representing multiple levels of prokaryotic taxonomic diversity (see fig. 3*A*). In each subgroup of this data set, dozens of MEOMs were detected (the exact number of MEOMs that was detected in each subgroup appears in fig. 3*A*; the details about all the MEOMs found in each group of organisms appear in supplementary tables 2–8, Supplementary Material online). We performed two randomization tests to show that the number of detected MEOMs is significant. In the first test, we compared the number of MEOMs to the result obtained for randomized coevolutionary networks with a similar network degree distribution (see the Materials and Methods section). In the second test, we compared our MEOM data with the results obtained when we randomly assigned COGs to genomes while maintaining the same number of COGs in each genome (Materials and Methods and Lima-Mendez et al.

2008). All the *P* values but one were significant (i.e., all *P* value < 0.0045); one *P* value was borderline significant (*P* value $= 0.083$); in most of the cases (all data sets for the first randomization, 5 of 6 data sets for the second randomization), the number of detected MEOMs in the randomized networks was zero or close to zero (see details in supplementary table 10, Supplementary Material online). In addition, we performed Jackknifing (see, details in the Materials and Methods section) to show that most of the resultant MEOMs reported in the paper are robust to changes in the input database.

We found that 78% of the MEOMs appear in at least 90% of the sampled databases (96% of the MEOMs appear in at least 70% of the sampled databases) demonstrating that the results reported in the paper are robust and that smaller data sets can also be used reliably (see details in supplementary table 10, Supplementary Material online).

Flexibility in the Coevolution of Protein Complexes

Interestingly, excluding the eukaryotes, in all the other groups analyzed, we found organisms that include both green cliques of the MEOMs (i.e., there are relatively few cases where these cliques completely mutually exclude each other, hinting to the existence of putatively incompatible gene families; fig. 3*B*). This is surprising as the aim of our algorithm is to find sets of orthologs with "as few organisms with both cliques as possible." This result demonstrates the great evolutionary flexibility of protein evolution because it implies that almost any combination of two complexes can coexist in the same cell without any deleterious effect or at least that in some organisms an evolutionary solution can be found that allows this coexistence. This inherent flexibility and adaptability of cellular life could help explain why lateral gene transfer is such a common evolutionary phenomenon because acquiring a functional complex in its entirety (e.g., by acquisition of a "selfish operon," [Lawrence and Roth 1996]) will outweigh the possible deleterious effects, at least in some lineages. More generally, this result is in accordance with previous studies emphasizing the robustness and flexibility of many biological systems (see, e.g., Barkai and Leibler 1997; Kitano 2004; Lehar et al. 2008; Li et al. 2009; Shinar and Feinberg 2010).

MEOMs Tend to Separate Organisms with Large Phylogenetic Distance but Also Separate Organisms That Live in Different Ecological Niches

In the next step, we wanted to verify if MEOMs represent both evolutionary proximity and environmental changes. Thus, we performed the following experiment: we considered the data set of all the bacteria; for each pair of bacterial species, we counted the number of times this pair of organisms appeared in different green cliques of a MEOM. We denote this number as the "MEOM separation index" of the pair of organisms.

We compared the MEOM separation of pairs of organisms with their evolutionary proximity (the topological distance in NCBI taxonomy; Materials and Methods) and their environmental proximity (according to Freilich et al. [2010], Materials and Methods). To show that the separation of MEOM decreases with phylogenetic distance while controlling for environmental proximity, we computed the correlation between the phylogenetic distance and the mean MEOM separations of organisms that do not inhabit the same environment (Spearman correlation $r = 0.96$; $P = 2.4 \times 10^{-10}$) and for organisms that do live in the same niche (Spearman correlation $r = 0.96$; $P = 3.3 \times 10^{-9}$). To show that MEOM separation decreases with environmental proximity, we computed for each possible phylogenetic distance the mean MEOM separation of pairs of organisms with this phylogenetic distance. We computed such a vector of mean MEOM separation for pairs of organisms from the same community and pairs of organisms that are in different communities. We compared these vectors to show that a component in the first vector tends to be smaller than the corresponding component in the second vector; that is, for every phylogenetic distance, pairs of organisms that are in the same community have lower mean MEOM separation indexed than pairs of organisms that are in different communities (Wilcoxon test: $P = 0.0005$). The corresponding graphs appear in figure 3C–E and show that the mean MEOM separation increases both with the mean evolutionary distance (even when controlling for environmental distance) and with environmental distance (Freilich et al. 2010) (even when controlling for evolutionary distance). In addition, we performed a multivariate regression analysis (see details in the Materials and Methods section). We showed that 1) there is correlation between MEOM separation and phylogenetic proximity even when controlling for ecological proximity and the size of the genomes together. 2) There is a correlation between MEOM separation and ecological proximity even when controlling for phylogenetic proximity and the size of the genomes together ($P < 10^{-16}$ in both cases). These results suggest that MEOMs reflect both lineage-specific and adaptation-driven changes (e.g., due to a shift in environmental conditions or lateral gene transfer).

MEOMs Are Enriched with Metabolic Genes and Outer Membrane Proteins

In order to understand if there are cellular functions that tend to appear more frequently in MEOMs, we computed for each group of organisms the number of times each cellular function appears in a MEOM. We considered three functional ontologies: the GO ontology of *S. cerevisiae* (Harris et al. 2004), the GO ontology of *Escherichia coli* (Harris et al. 2004), and the more general COG ontology (Tatusov et al. 2003). Table 1 depicts the top cellular functions that are enriched in MEOMs when considering the COG ontology. As can be seen, the main cellular functions that are enriched in MEOMs relate

Table 1
Cellular Functions (COG Ontology; Tatusov et al. 2003) That Tend to Appear in MEOMs

Group of Organisms	Cellular Functions	Number of MEOMs with the Function	Total Number of Enriched Functions	Total Number of MEOMs	Most/Least Significant *P* Value
All	Translation, ribosomal structure and biogenesis	4	16	22	$2.22 \times 10^{16}/2.27 \times 10^{-02}$
	Energy production and conversion	4	16	22	$4.46 \times 10^{-11}/6.88 \times 10^{-04}$
Archaea	Energy production and conversion	1	2	11	4.43×10^{-004}
	Amino acid transport and metabolism	1	2	11	1.95×10^{-004}
Bacteria	Energy production and conversion	5	14	21	$6.08 \times 10^{-12}/4.88 \times 10^{-04}$
	Replication, recombination, and repair	2	14	21	$1.19 \times 10^{-05}/1.06 \times 10^{-3}$
Eukaryotes	Translation, ribosomal structure, and biogenesis	1	4	14	2.21×10^{-008}
	Amino acid transport and metabolism	1	4	14	2.21×10^{-008}
Proteobacteria	Energy production and conversion	5	10	9	$2.41 \times 10^{-13}/7.58 \times 10^{-04}$
	Carbohydrate transport and metabolism	3	10	9	$6.88 \times 10^{-10}/4.45 \times 10^{-05}$
Firmicutes	Energy production and conversion	2	3	5	$4.07 \times 10^{-05}/5.16 \times 10^{-04}$
	Amino acid transport and metabolism	1	3	5	$4.44 \times 10^{-05}/3.04 \times 10^{-04}$
Actinobacteria	Cell wall/membrane/envelope biogenesis	2	8	15	$1.37 \times 10^{-06}/7.71 \times 10^{-04}$
	Intracellular trafficking, secretion, and vesicular transport	2	8	15	$5.77 \times 10^{-08}/5.42 \times 10^{-05}$

to "Transport," "Energy production," "Metabolism," and "Translation" (similar results were obtained when we used the GO ontologies of *E. coli* or *S. cerevisiae*; supplementary table 9, Supplementary Material online; all the *P* values appear in supplementary tables 1–8, Supplementary Material online) suggesting that changes in the metabolic environments that require adaptation to new sources of energy are central triggers of complex/pathway replacement in evolution.

In addition, we see many results that are related to outer membrane proteins (e.g., 5 of the 21 bacterial MEOMs were enriched with GO terms related to the outer membrane). Such proteins are involved in the interaction of Gram-negative bacteria with the environment and could thus be replaced as a result of adaptation (indeed, when considering only these MEOMs, for every phylogenetic distance, pairs of organisms that are in the same community have lower mean MEOM separation than pairs of organisms that are in different communities; Wilcoxon test: $P = 0.005$). Furthermore, different organismal lineages differ in their cellular envelope structure, imposing different constraints on the families of outer membrane proteins that can be found in them (e.g., lipid monolayers in some archaea, double membranes in Gram-negative bacterial lineages, mycolic acid in some cell walls, etc.). Thus, for example, in Bacteria (bacterial MEOM 1) the Gram-positive bacteria that do not possess the conserved outer membrane proteins common to Gram negatives, have a mutually exclusive clique of protein components of a putative ABC transporter, which includes a typical Gram-positive lipoprotein substrate-binding subunit.

Our algorithm detects pairs of cliques, each with very strong coevolutionary relations. Therefore, in many cases, the entire clique is replaced by another clique. Thus, the algorithm identifies only very rigid co-occurrence cliques, that is, protein complexes (or subcomplexes) where "subunits" cannot be substituted by members of other protein families while still maintaining the function of the complex. This often results in the identification of the most basic functional modules within a complex. Our results imply that acquisition and subsequent replacement of subunits in these complexes (or subcomplexes) by lateral gene transfer will only proceed through xenologous gene replacement rather than by introduction of nonhomologous subunits. Thus, it is possible that many of the cliques reported in this paper (e.g., the outer membrane complexes and ribosomal/proteosomal subcomplexes mentioned above) are such basic functional modules.

Mutual Exclusion of H+-ATPase and F-ATPase in Bacteria and Archaea

We described above an example of mutually exclusive complexes due to divergence of lineages, that is, the translational apparatus of Archaea and Bacteria. In the case of the ribosome, lineage divergence probably coincides with incompatibility because the same cell cannot harbor two different ribosome types, unless some compartmentalization had occurred (e.g., eukaryotic organelles). However, mutually exclusive protein complexes can also be a result of the loss of functional redundant complexes, where an activity does

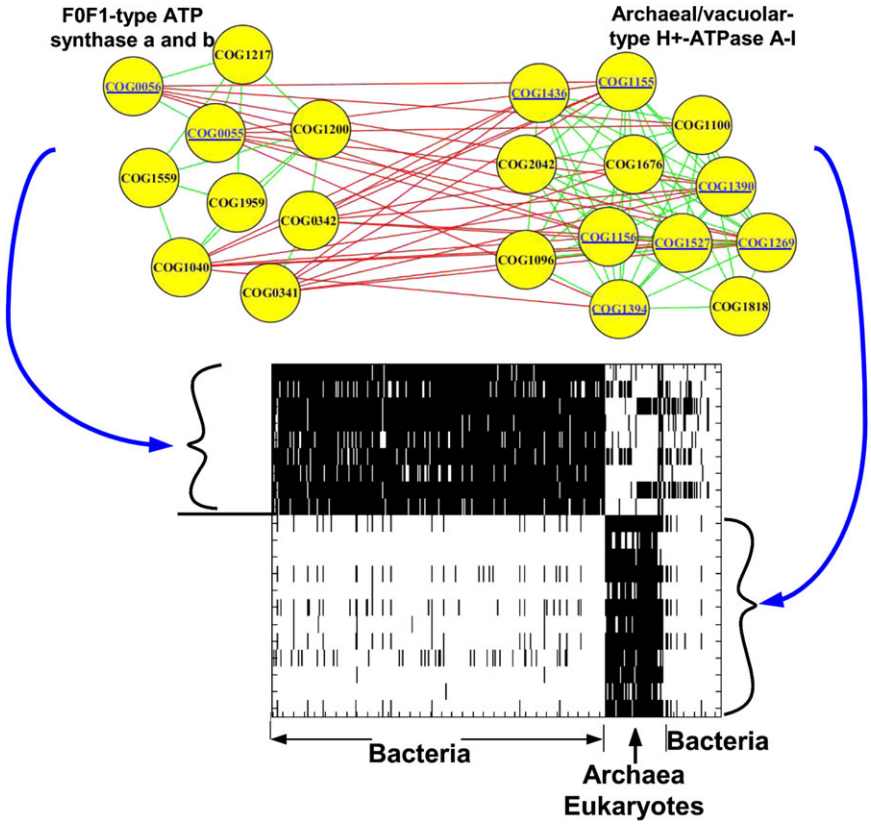


Fig. 4.—One of the MEOMs found in the entire data set (383 organisms): the "archaeal/vacuolar-type H+-ATPase" versus "F0/F1-type ATP synthase." The COGs related to these two pathways are marked.

not need to be carried out by two parallel mechanisms. This situation can arise following acquisition of a gene cluster by HGT or after a large duplication event. An interesting example is the case of the Archaeal/vacuolar-type H+-ATPase in bacteria (Bernasconi et al. 1989; Hilario and Gogarten 1993, 1998). Bacteria in general lack the Archaeal/vacuolar-type H+-ATPase complex, with a few notable exceptions attributed to lateral gene transfer, such as the spirochete *Borrelia burgdorferi* (Fraser et al. 1997; Hilario and Gogarten 1998) and *Chlamydia* species (Stephens et al. 1998), both intracellular pathogens. Curiously, these intracellular bacteria have lost the ancestral bacterial F-ATPase, hence their detection as MEOMs (when considering the entire data set; see fig. 4), and it appears likely that the laterally acquired ATPase now pumps out protons from the cytoplasm and generates a proton gradient, for energy, transport, or both, similar to the role of the ancestral bacterial system (McClarty 1999). One can speculate that the reason that the acquired ATPase replaced the ancestral one is some adaptation to life inside a eukaryotic cell.

Mutual Exclusion of CO Dehydrogenase/Thymidylate Synthase and the Glycine Cleavage Genes—Clues for Thymidine Metabolism in Archaea?

Although proteins of the putative glycine cleavage system are abundant in archaea, their roles are unclear and a *Haloferax volcanii* mutant defective for one of its components, dihydrolipoamide dehydrogenase, showed no growth inhibition on a variety of minimal media (Nakamura et al. 2004). These genes are present in all archaea in our data set, except the methanogens and *Archaeoglobus fulgidus*, an archeon that also has some methanogenesis-related genes. The latter organisms have a clique of proteins mutually exclusive with the glycine cleavage system. These proteins include CO dehydrogenase, which is found in methanogenic archaea, and can function either as an acetyl-CoA synthase or in the fermentation of acetate to methane (Ciccarelli et al. 2006). Another protein in this clique is the ThyA-type thymidylate synthase. Because methanogens do not synthesize folate and its derivatives, it is assumed that different molecules provide the required methyl group for converting dUMP to dTMP, such as the modified folate tetrahydromethanopterin. Glycine can provide a methylene group by glycine cleavage and generate either 5,10-methylenetetrahydrofolate (from tetrahydrofolate) or the analogous tetrahydromethanopterin. Thus, archaeal species that have the alternative thymidylate synthase ThyX (Jaccard 1912) may use glycine cleavage for dTMP synthesis. We therefore predict that unlike the lack of phenotype observed in the *Hfx. volcanii* mutant (Nakamura et al. 2004), the same mutation in the halophilic

archeon *Halobacterium salinarum*, which relies on ThyX rather than ThyA (Giladi et al. 2002), should yield a strain deficient in the synthesis of thymidine.

Conclusions

Coevolutionary networks represent relations between proteins that tend to coevolve. These networks can yield important evolutionary insights at the individual complex level as well as reveal important general trends in evolutionary systems biology, yet few methods to explore these networks have been developed. Here, we used a novel approach for analyzing coevolutionary networks. Our method captures both lineage-specific and environment-specific changes and can therefore be used independently of an organismal phylogeny to investigate complex evolutionary scenarios that include HGT, gene duplication, and linage-specific changes in very large number of organisms. This is important because most evolutionary algorithms rely on a given "tree of life," whereas in reality, many relationships are not well supported due to HGT (Jin et al. 2006; Doolittle and Bapteste 2007) or simply because deep relationships no longer maintain sufficient evolutionary signal (Ciccarelli et al. 2006). Another important feature of our method is that it was designed to infer both relations between organisms and between gene families. Thus, it adds an additional dimension to previous approaches in molecular evolution that mainly infer either relations between gene families (see, e.g., Nakamura et al. 2004; Tuller, Birin, et al. 2009; Tuller, Kupiece, and Ruppin 2009) or between organisms (see, e.g., Ge et al. 2005; Doolittle and Bapteste 2007; Dagan et al. 2008). Another advantage of our approach is the fact that it can be implemented in a hierarchical manner to study biologically relevant groups of organisms, at different taxonomic levels, revealing key adaptations of subsets within phylogenetic groups.

The fact that most of the cliques reported in this paper corresponded to known functional complexes confirms that our method usually detects biologically meaningful results. In addition, it suggests that it can be applied in the future for detecting functional relations between gene families. A gene family that is connected by coevolutionary relations to many gene families with a certain function probably has a similar cellular function (see also Tuller, Kupiece, and Ruppin 2009). For example, MEOM 8 of the *Actinobacteria* includes many flagellar genes but also some uncharacterized proteins. We believe that the uncharacterized proteins may also be related to motility (other examples were reported above). The MEOM principle can also be generalized to study different network types. Thus, one can analyze networks whose nodes correspond to domains within proteins or organisms in a community network (Freilich et al. 2010).

Supplementary Material

Supplementary tables S1–S10 are available at *Genome Biology and Evolution* online (http://gbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Dr Igor Ulitsky, Prof Elchanan Mossel, and Prof Yitzhak Pilpel for very helpful discussions. T.T. is a Koshland Scholar at Weizmann Institute of Science. M.K.'s research was supported by grants from the Israel Science Foundation and the McDonnell Foundation. U.G. is supported by the McDonnell Foundation and the Israeli Ministry of Health.

Literature Cited

Barkai N, Leibler S. 1997. Robustness in simple biochemical networks. Nature 387(6636):913–917.

Bernasconi P, Rausch T, Gogarten JP, Taiz L. 1989. The H+ ATPase regulatory subunit of Methanococcus thermolithotrophicus: amplification of an 800 bp fragment by polymerase chain reaction. FEBS Lett. 251(1–2):132–136.

Chen Y, Dokholyan NV. 2006. The coordinated evolution of yeast proteins is constrained by functional modularity. Trends Genet. 22(8):416–419.

Ciccarelli FD, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311(5765):1283–1287.

Cohen O, Pupko T. 2009. Inference and characterization of horizontally transferred gene families using stochastic mapping. Mol Biol Evol. 27(3):703–713.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105(29):10039–10044.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A. 104(7):2043–2049.

Farris JS. 1969. A successive approximations approach to character weighting. Syst Biol. 18(4):374–385.

Fraser CM, et al. 1997. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390(6660):580–586.

Freilich S, et al. 2010. The large-scale organization of the bacterial network of ecological co-occurrence interactions. Nucleic Acids Res. 38(12):3857–3868.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3(10):e316.

Giladi M, Bitan-Banin G, Mevarech M, Ortenberg R. 2002. Genetic evidence for a novel thymidylate synthase in the halophilic archaeon Halobacterium salinarum and in Campylobacter jejuni. FEMS Microbiol Lett. 216(1):105–109.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2009. Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci U S A. 107(1):127–132.

Harris MA, et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32(Database issue): D258–D261.

Hershberg R, Yeger-Lotem E, Margalit H. 2005. Chromosomal organization is shaped by the transcription regulatory network. Trends Genet. 21(3):138–142.

Hilario E, Gogarten JP. 1993. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. Biosystems 31(2–3):111–119.

Hilario E, Gogarten JP. 1998. The prokaryote-to-eukaryote transition reflected in the evolution of the V/F/A-ATPase catalytic and proteolipid subunits. J Mol Evol. 46(6):703–715.
Jaccard P. 1912. The distribution of flora in the alpine zone. New Phytologist. 11:37–50.
Jensen LJ, et al. 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 37(Database issue):D412–D416.
Jin G, Nakhleh L, Snir S, Tuller T. 2006. Maximum likelihood of phylogenetic networks. Bioinformatics 22(21):2604–2611.
Juan D, Pazos F, Valencia A. 2008. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci U S A. 105(3):934–939.
Kelley R, Ideker T. 2005. Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol. 23(5):561–566.
Kelley BP, et al. 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A. 100(20):11394–11399.
Kitano H. 2004. Biological robustness. Nat Rev Genet. 5(11):826–837.
Lawrence JG, Roth JR. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143(4): 1843–1860.
Lehar J, Krueger A, Zimmermann G, Borisy A. 2008. High-order combination effects and biological robustness. Mol Syst Biol. 4(215):215.
Li X, Cassidy JJ, Reinke CA, Fischboeck S, Carthew RW. 2009. A microRNA imparts robustness against environmental fluctuation during development. Cell 137(2):273–282.
Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol. 25(4):762–777.
Lin G, et al. 2006. Mycobacterium tuberculosis prcBA genes encode a gated proteasome with broad oligopeptide specificity. Mol Microbiol. 59(5):1405–1416.
McClarty G. 1999. Chlamydial metabolism as inferred from genome sequence. In: Stephens RS, editor. Chlamydia: intracellular biology, pathogenesis, and immunity. ASM Press. pp 69–100.
Milo R, et al. 2004. Superfamilies of evolved and designed networks. Science 303(5663):1538–1542.
Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 36(7):760–766.
Roberts E, Sethi A, Montoya J, Woese CR, Luthey-Schulten Z. 2008. Molecular signatures of ribosomal evolution. Proc Natl Acad Sci U S A. 105(37):13953–13958.
Shao J, Tu D. 1995. The Jackknife and bootstrap. New York: Springer-Verlag, Inc.
Sharan R, et al. 2005. Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A. 102(6):1974–1979.
Shinar G, Feinberg M. Structural sources of robustness in biochemical reaction networks. Science 327(5971):1389–1391.
Stephens RS, et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. Science 282(5389):754–759.
Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4(41):41.
Tuller T, Birin H, Gophna U, Kupiec M, Ruppin E. 2009. Reconstructing ancestral gene content by coevolution. Genome Res. 20(1):122–132.
Tuller T, Kupiec M, Ruppin E. 2009. Co-evolutionary networks of genes and cellular processes across fungal species. Genome Biol. 10(5):R48.
Tuller T, et al. 2011. Associations between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acids Res. 2011 Feb 22. [Epub ahead of print].
Ulitsky I, Shamir R. 2007. Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. Mol Syst Biol. 3(104):104.

Associate editor: Bill Martin