Research

The society of genes: networks of functional links between genes from comparative genomics

Itai Yanai*† and Charles DeLisi*

Address: *Bioinformatics Graduate Program and Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA. †Current address: Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel.

Correspondence: Itai Yanai. E-mail: itai.yanai@weizmann.ac.il

Published: 25 October 2002

Genome ***Biology*** 2002, **3(11)**:research0064.1–0064.12

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/11/research/0064

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 12 March 2002
Revised: 2 August 2002
Accepted: 11 September 2002

Abstract

Background: Comparative genomics provides at least three methods beyond traditional sequence similarity for identifying functional links between genes: the examination of common phylogenetic distributions, the analysis of conserved proximity along the chromosomes of multiple genomes, and observations of fusions of genes into a multidomain gene in another organism. We have previously generated the links according to each of these methods individually for 43 known microbial genomes. Here we combine these results to construct networks of functional associations.

Results: We show that the functional networks obtained by applying these methods have different topologies and that the information they provide is largely additive. In particular, the combined networks of functional links contain an average of 57% of an organism's complete genetic complement, uncover substantial portions of known pathways, and suggest the function of previously unannotated genes. In addition, the combined networks are qualitatively different from the networks obtained using individual methods. They have a dominant cluster that contains approximately 80%-90% of the genes, independent of genome size, and the dominant clusters show the small world behavior expected of a biological system, with global connectivity that is nearly random, and local properties that are highly ordered.

Conclusions: When the information on functional linkage provided by three emerging computational methods is combined, the integrated network uncovers large numbers of conserved pathways and identifies clusters of functionally related genes. It therefore shows considerable utility and promise as a tool for understanding genomic structure, and for guiding high throughput experimental investigations.

Background

Complex systems, ubiquitous in science, owe their complexity to the interrelatedness of their elements [1]. Investigations of the local and global structures of network representations of these systems have advanced our understanding of the systems themselves as well as some of their emergent, system-level properties [2,3]. A network of interactions also relates genes of a cell as each gene product carries out its function in the context of other gene products. Thanks to the availability of complete genomic sequences, the elements of

cellular networks - the gene products - have been identified. Methods for identifying functional relationships between genes have also been introduced. We now have the opportunity to get a glimpse of the structure of the networks that lie at the core of life at the cellular level.

Novel high-throughput experimental methods for identifying protein-protein interactions, such as yeast two-hybrid [4,5] and mass spectrometry [6,7], are now complemented by computational analyses of sequenced genomes to detect functional links between genes. The three comparative genomics methods applied here (Figure 1) utilize correlations in the properties and occurrence of genes across known genomes [8-11]. The first method, phylogenetic profiling, infers functional linkage between genes whose orthologs have identical phylogenetic patterns of occurrence across genomes [12-14]. More generally, this method links genes when the similarity between their phyletic distributions is unlikely to have occurred by chance. Conserved chromosomal proximity of genes in multiple genomes, whether enforced by operons or co-horizontal transfer [15], forms the basis for the second method for detecting functional correlations [16-19]. Finally, the domain fusion method [20-23] is based on the observation that distinct non-homologous genes are functionally related if their orthologs are fused in another organism.

Previously, we reported on a database of functional links generated by the comparative genomics methods [24]. Here we combine the total sets of links generated by these three methods for each of 43 microbial genomes with the following findings.

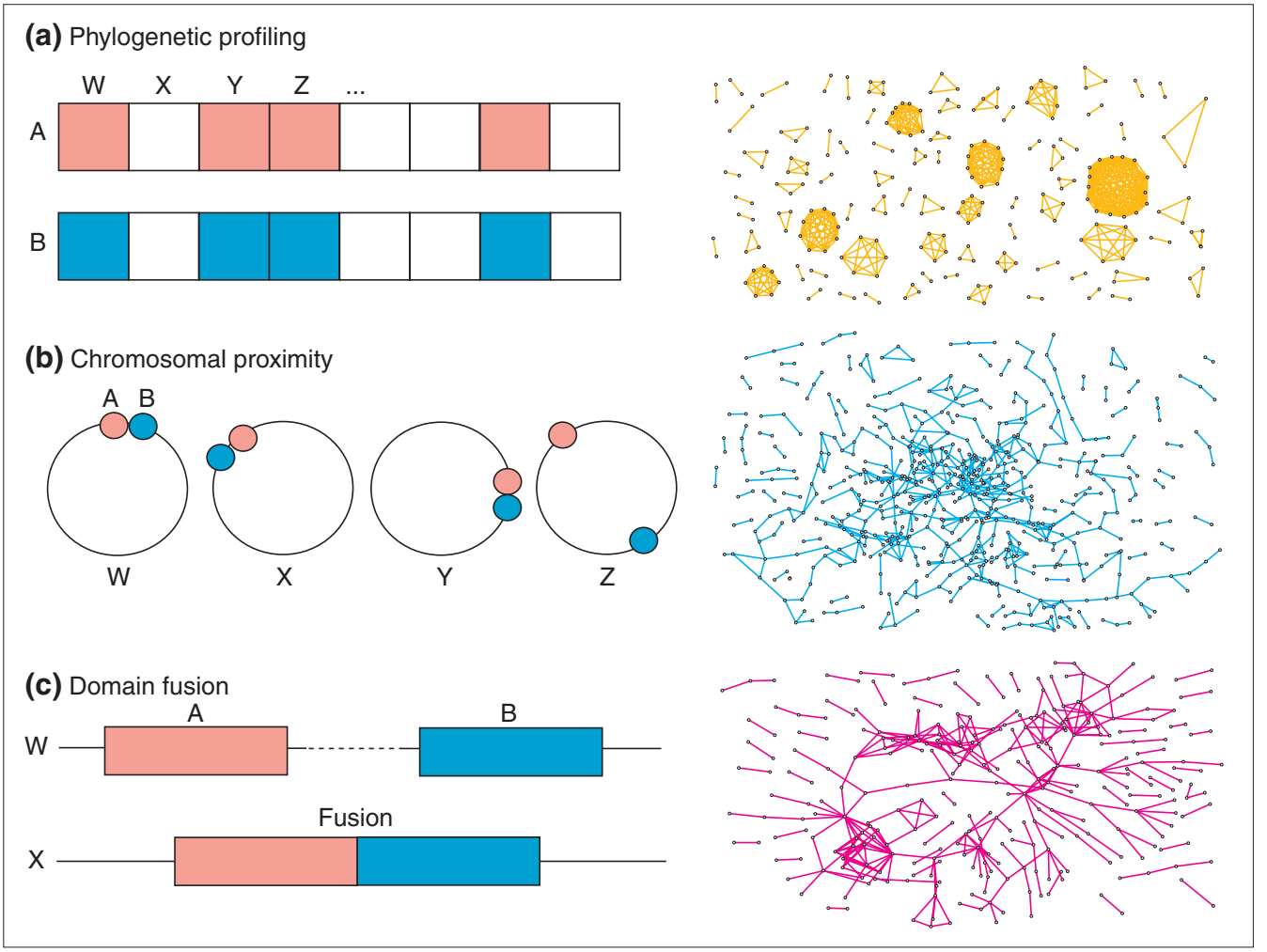

Figure 1
Three comparative genomics methods for identifying functional links between genes and the networks they produce. The schematics on the left show links between gene A (pink) and gene B (blue) based on **(a)** the same phyletic distribution across the known genomes, here arbitrarily labeled W, X, Y, Z, etc.; **(b)** their proximity on chromosomes from different genomes; and **(c)** fusion of A and B into one multidomain gene in another organism. On the right are networks found in *H. pylori* using each of the three methods. All network figures were made using the Pajek program [42].

Sets of functional links provided by the three methods are largely additive (that is, the sets largely do not overlap). The reliability of the links is approximately 70%, well above the background noise, which is estimated to be in the 10-15% range. Networks obtained by individual methods coalesce into a combined network covering, on average, 57% of an organisms' genes. The local structures of the combined networks correspond to genes participating in the same pathway. In the *Escherichia coli* network, 26 clusters are identified such that each is composed of seven or more genes and corresponds to different functional pathways. Functional predictions can be made based upon a gene's location in the combined networks. For example, *Methanococcus jannaschii* genes MJ1313 and MJ1407 are of completely unknown function, but are unambiguously predicted to be related, and therefore should be studied together. A giant cluster covering between 80 to 90% of the genes of the total network and demonstrating random global properties and highly cliquish local properties characterizes the structure of the combined networks.

Results and discussion

Networks from individual comparative genomic methods

The relationships uncovered by each method form networks whose structures are method-dependent (Figure 1a). Phylogenetic profile links, which we will refer to as 'phylo links', are transitive; that is, two genes linked to a common gene are also linked to each other. An example is the fully connected clique formed by flagella motor proteins (also shown in Pelligrini *et al.* [14]); that is, the proteins are generally either all present or all absent in a given genome and therefore have identical phylogenetic patterns. Phylo networks are characterized by the number and size distribution of their cliques. The occurrence of fully connected cliques follows from the definition that requires perfectly correlated patterns of occurrence to assign a link. A more permissive phylogenetic profiling method would result in reduced transitivity.

Gene networks uncovered by conserved chromosomal proximity links, which we will refer to as 'chromo links', typically exhibit a chain structure. For example, one of the long chains in the network of *Helicobacter pylori* genes (Figure 1b) corresponds to the chromosomal region containing highly conserved ribosomal genes. This network formation results from the fact that the conservation of gene order tends to involve more than two genes. As a particular conserved group tends to be arranged in an order conserved among organisms, linear concatenation dominates the network.

The domain fusion links, which we will refer to as 'fusion links', have complicated network relations, including the appearance of a major cluster. Even for small networks, as in Figure 1c, we find a few nodes that have a large number of links, and a large number that have few links [25-27]. We discuss this power-law behavior below.

In summary, the three methods for constructing networks all produce distinctive, non-random structures: branched but highly cliquish for phylo networks; generally branched structures for fusion networks; and linearly concatenated structures for chromo networks. By combining these three representations on the same grid of genes, we generate a largely non-overlapping map of functional linkages that provides more information than any of the three maps alone. The combined networks have a number of important functional and structural properties.

Functional properties of the combined comparative genomics networks

When the three networks are superimposed, four new types of links are formed combinatorially (Figure 2). The combined network of fusion, chromo and phylo links for a given genome (henceforth, combined network) captures between 30 and 80% of the genes in a genome and 57% on average, whereas the chromo, fusion and phylo networks independently capture 48%, 29% and 19% of the genes on average. As these numbers suggest, the graphs overlap significantly in terms of nodes (Table 1). We found, as did Huynen *et al.* [28] in *Mycoplasma genitalium*, that of the three methods the chromo networks have the greatest coverage. The fusion and phylo networks share 72% and 75% of their nodes with chromo networks; that is, 72% and 75% of the nodes found in fusion and phylo networks are also found in chromo networks. Functional links, on the other hand, tend to be complementary. Only 20% and 14% of the links in fusion and phylo networks, respectively, are found in chromo networks. The least overlapping of all are the phylo and fusion graphs, with only 41% of the nodes in a phylo net found in a fusion net and 6% of the links in a phylo net found in a fusion net. These results indicate that although the three methods capture overlapping sets of genes, and conserved chromosomal proximity captures more than the other two methods combined, the links generated by the different methods individually show much less overlap than the nodes.

A fundamental question is the quality of the links in terms of a direct functional correlation between the linked members. This quality can be estimated by reference to databases that classify genes into broad functional categories (clusters of orthologous groups (COGs) [12]) or biological pathways (Kyoto Encyclopedia of Genes and Genomes (KEGG) [18]). In particular, 72%, 68% and 64% of the fusion, chromo and phylo links are in the same COGs functional category, respectively [19,22,24].

We find, as did Marcotte *et al.* [29], that the links corroborated by multiple methods are of exceedingly good quality. However, as discussed above, this intersecting set corresponds to a small fraction of the total links (approximately 10%). Thus, in our study of combined networks, we construct the union, instead of the intersection, of the links generated by each method.

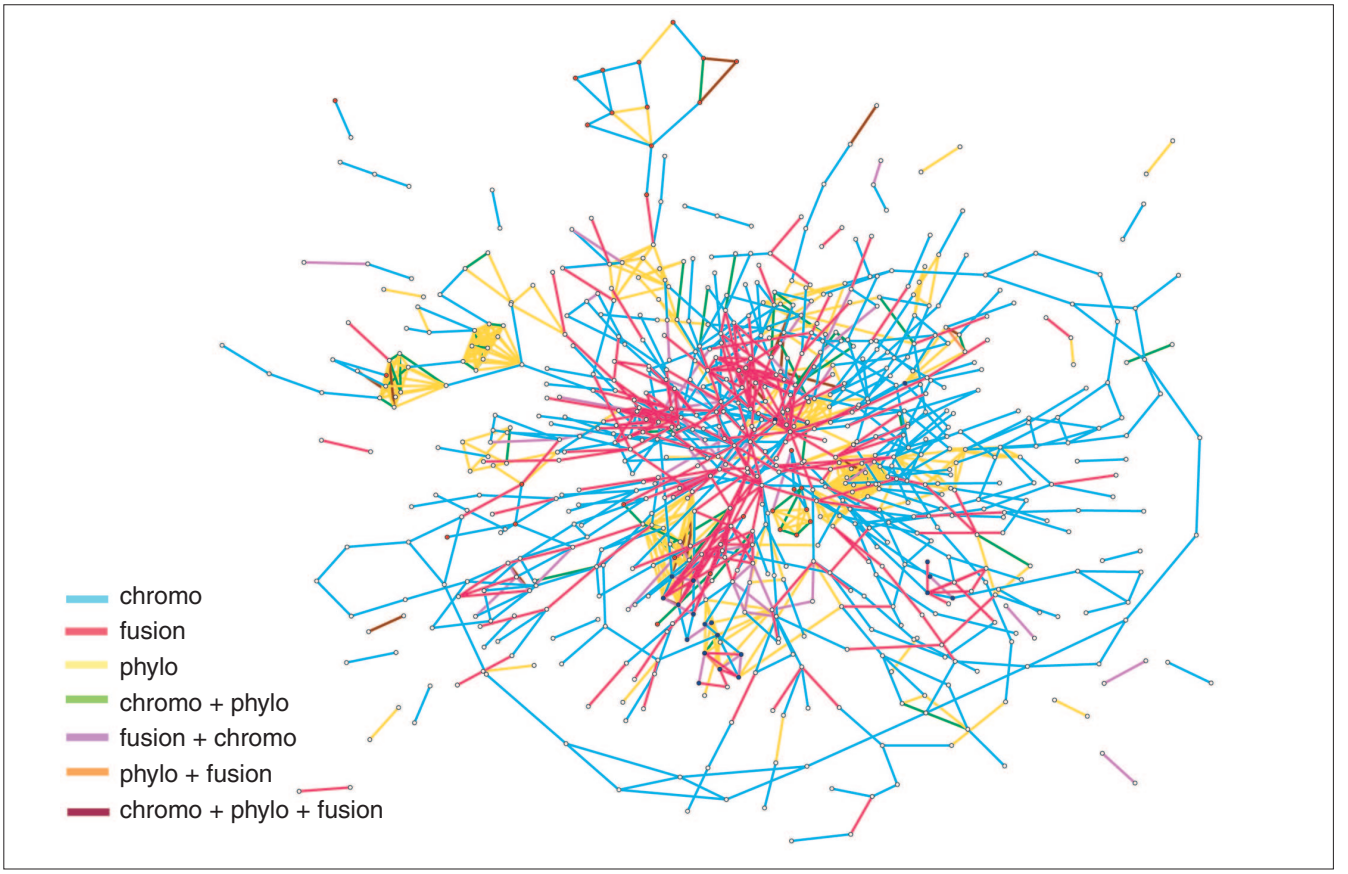

Figure 2
The combined network in *H. pylori*. The networks of the three individual methods (shown in Figure 1) are superimposed. Links colored yellow were obtained by phylogenetic profiling; links colored blue by conserved chromosomal proximity, and links colored red by domain fusion. The remaining links are coded as composites of the three primary colors: purple, links found by both fusion and chromosomal proximity; green, links found by chromosomal proximity and phylogenetic profiling, orange, links found by phylogenetic profiling and fusion; and brown, links found by all three methods. The nodes highlighted in red and blue identify genes that participate in oxidative phosphorylation and aromatic amino-acid biosynthesis, respectively (see Figure 4).

To evaluate the significance of a network we first produced 100 shuffled versions of this network (see Materials and methods). For each of these we calculate the percentage of linked pairs whose members are present in the same functional category (according to COGs) or pathway (according to KEGG) and then estimate the average and standard deviation for the population of random networks. We find a statistically significant difference between the observed networks of all types (fusion, chromo, phylo and combined), and the shuffled networks (p = 2.2e-16 in a t-test). On average, 33 standard deviations of separation distinguish an observed network from its 100 shuffled networks. As can be seen in Figure 3, the observed networks have high functional correlation centered around 60% agreement with COGs and 70% with KEGG, whereas the shuffled networks form a tight distant cluster with a low background functional correlation - a noise level in the range of 10 to 20%.

The methods described here uncover a large number of functional systems such as citrate cycle, purine metabolism, and fructose and mannose metabolism. To provide two specific examples, the highlighted nodes in the network shown in Figure 2 correspond to genes that encode proteins involved in oxidative phosphorylation and phenylalanine, tyrosine and tryptophan biosynthesis. We find that all but four of the 35 *H. pylori* genes annotated as participating in oxidative phosphorylation are present as six clusters in the combined network for this organism (Figure 4a). The clusters found by the three methods reflect the functions subserved by different groups of oxidative phosphorylation genes. In particular, the six clusters shown in Figure 4a, ranging in size from 1 to 12 are found in the network corresponding to five of the six receptor complexes that act in oxidative phosphorylation: NADH dehydrogenase, ATP synthase, succinate dehydrogenase, cytochrome *bc*, and cytochrome *d*.

As can be gleaned from Figure 4a, seven of the genes involved in ATP synthase (HP1131 to HP1137) form a conserved cluster of genes (detected in almost all known bacterial genomes). The chromo links among HP0828, HP1212 and

Table 1

Overlap of nodes and edges between the four types of network

(a) Nodes	Chromo	Fusion	Phylo	Combined
Chromo		41 (6)	31 (3)	100 (0)
Fusion	72 (6)		30 (3)	100 (0)
Phylo	75 (10)	41 (6)		100 (0)
Combined	80 (6)	46 (6)	34 (4)	

(b) Edges	Chromo	Fusion	Phylo	Combined
Chromo		14 (2)	11 (1)	100 (0)
Fusion	20 (4)		7 (2)	100 (0)
Phylo	14 (7)	6 (2)		100 (0)
Combined	43 (10)	30 (7)	38 (11)	

Element in row *i* column *j* is the percentage of average nodes **(a)** or average edges **(b)** in network of type *i* that are also found in network of type *j*, with the standard deviation in parentheses. For example, 75% of the nodes found by phylogenetic profiling are contained in the chromo networks. This should, however, be contrasted with the finding that only 14% of the functional links in the phylo networks are in chromo networks.

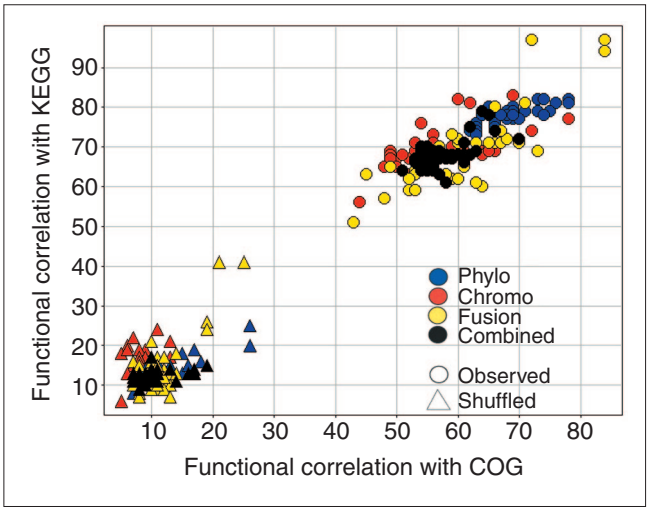

Figure 3
Functional correlation of networks in terms of COG functional correlations and KEGG pathways. Each circle corresponds to a network of one of the four types of the observed networks (differently colored) for one of the 43 genomes. Each triangle corresponds to the mean of 100 shuffled versions of each of the observed networks (see Methods). Note the clear separation of the functional correlation between the observed and shuffled networks.

HP1137:HP1136 (the last two are paralogs), are based on conserved proximity in other genomes. Six of the genes also have identical phylogenetic profiles, and the profile is unique to these six genes. The link between genes HP1135 and HP1137:HP1136 is strengthened because it is established by both chromosomal proximity and fusion in two *Mycobacterium* genomes. The five (of six) *H. pylori* genes that form the cytochrome *bc* and *c* complexes are linked. Essentially, the string of links is based upon two separate chromosomal sections: HP0144, HP0145 and HP0147; and HP1539 and HP1540. Within these two sections, two links are further strengthened by phylo links. The two sections are linked by a fusion link between two of the genes (HP1539 and HP0147). Three genes (HP0191, HP0192 and HP0193) compose the fumarate reductase complex and are connected by two chromo links. Chromosomal links build a skeleton of links that unite all the genes involved in NADH dehydrogenase. This cluster is further supported, however, by numerous phylo and fusion links. Two links are supported by all three methods. Two genes, HP1010 and HP1420, involved in oxidative phosphorylation are not connected to other genes involved in this functional system.

Figure 4b shows the functional links among the *H. pylori* genes involved in phenylalanine, tyrosine and tryptophan biosynthesis. Genes HP0402, HP0403 and HP0774 correspond to tRNA synthetase domains of phenylalanine and tyrosine linked by fusion events elsewhere as well as the proximity between HP0402 and HP0403 (HP0402 corresponds to two nodes, one for each of its domains, see Materials and methods). HP1249, HP1277, HP0663 and HP0401 share the same phylogenetic profile. HP0401, HP1249, HP0157 and HP0283 are found fused together as the yeast *Aro1* gene and are thus all fusion linked. The proximity of HP1277 and HP1278, tryptophan synthase α and β chain respectively, is conserved in other genomes. These two genes are also found fused together in some genomes. The relationship between HP1279 (represented twice for each of its domains), HP1281 and HP1282 is enforced by phylo, chromo and fusion links for each pair of the three genes, with the exception of a lack of a fusion link between HP1279 and HP1282. HP0154 and HP0672 are present in the network but are not linked to other genes known to act in this pathway.

Overall, in *E. coli*, 26 clusters of seven or more genes corresponding to distinct KEGG pathways are identified in the network; where a cluster is a minimally connected subgraph such as those shown in Figure 4. The ability to reconstruct pathways depends on an integration of the three methods. In other words, the clusters shown would become fragmented if the methods were used individually. To illustrate this, for each of the 26 *E. coli* clusters we ask what fraction is obtained by the individual methods alone (Figure 5). Although some of the pathway clusters can be completely recovered by one of the methods - for example, the ribosomal system can be completely accounted for by chromo links - most can only be found by the integration of the methods. As an example, in the cluster in the combined network of 10 genes involved in the citrate cycle, 30%, 60% and 30% can be associated by only the fusion, chromo and phylo links, respectively.

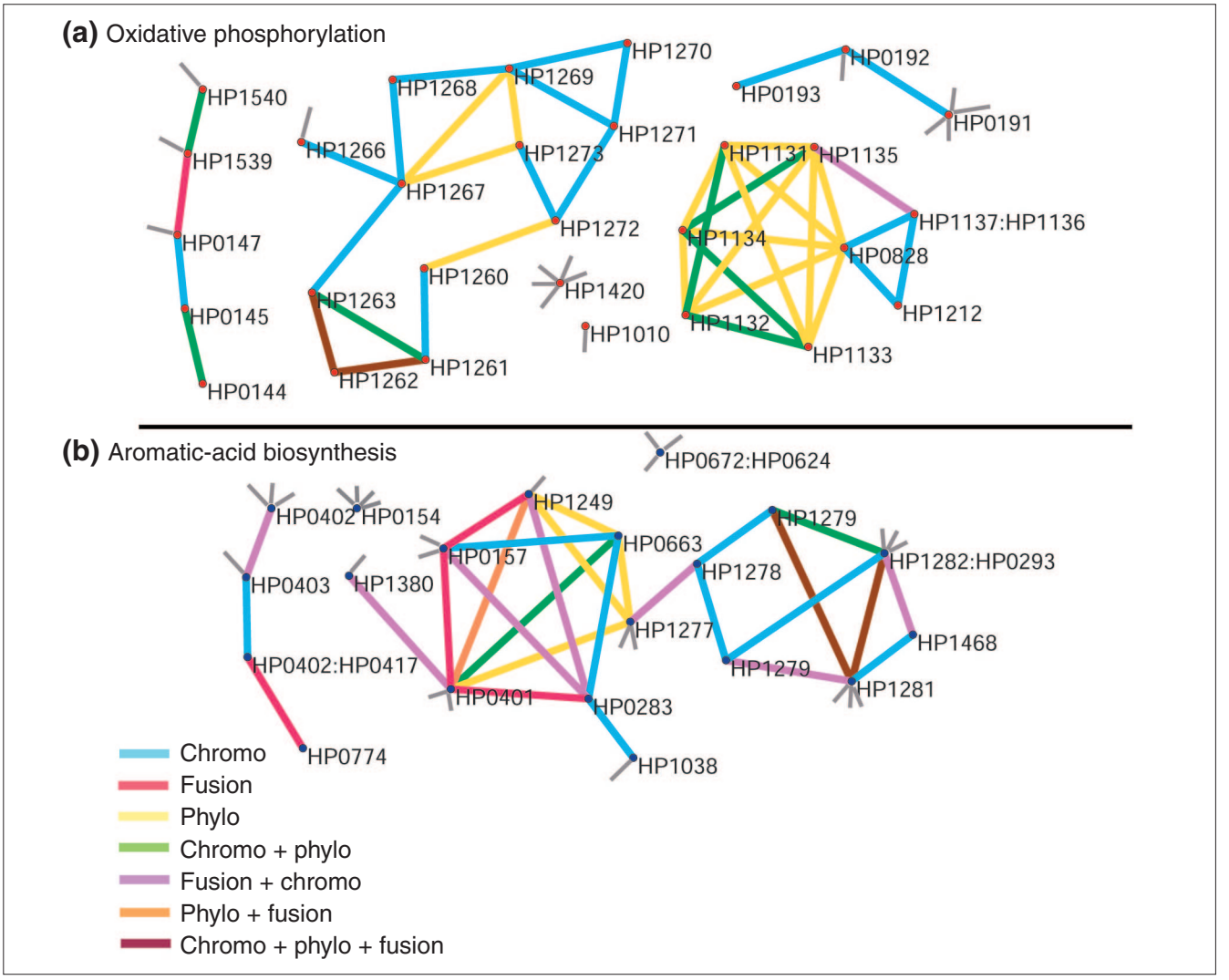

Figure 4
Local structure of the *H. pylori* combined network captures functionally related genes. **(a)** Oxidative phosphorylation genes; **(b)** genes involved in phenylalanine, tyrosine and tryptophan biosynthesis. The color of the links between genes is the same as in Figure 2. Gray lines indicate links with genes not ascribed to that pathway.

New functional information on particular genes can be predicted on the basis of the combined networks. Naturally, the most fundamental unit of prediction in the network corresponds to a link between two nodes. For example, *M. jannaschii* genes MJ1313 and MJ1407, whose functions are completely uncharacterized, are linked by fusion, chromo and phylo links. From this superlink we can confidently conclude that, although the actual functions of the genes in question remain elusive, a functional link probably exists between the two genes and they need to be experimentally studied together.

By superimposing known pathway information onto the combined networks, functional predictions can be made on the basis of a gene's location in the network. For example, Figure 6 shows a fraction of the *Thermotoga maritima* network. TM0885 and TM1367 have no characterized function, but their position in the network suggests that they have a role in energy production. These two genes form a four-gene clique of phylo links with TM0397 (a ferrodoxin-like domain) and TM0034 (an uncharacterized protein with a putative Fe-S center from the COGs database). The context of the completely uncharacterized genes (TM0885 and TM1367) within the network further strengthens the inference of their functional association with their neighbors (Figure 6). TM0397 is fusion linked to TM0396, a dehydrogenase with a Fe-S cluster, which is in turn fusion linked to an oxidoreductase protein, TM1640. From this network locus of TM0885 and TM1367, we predict for them a role in energy production.

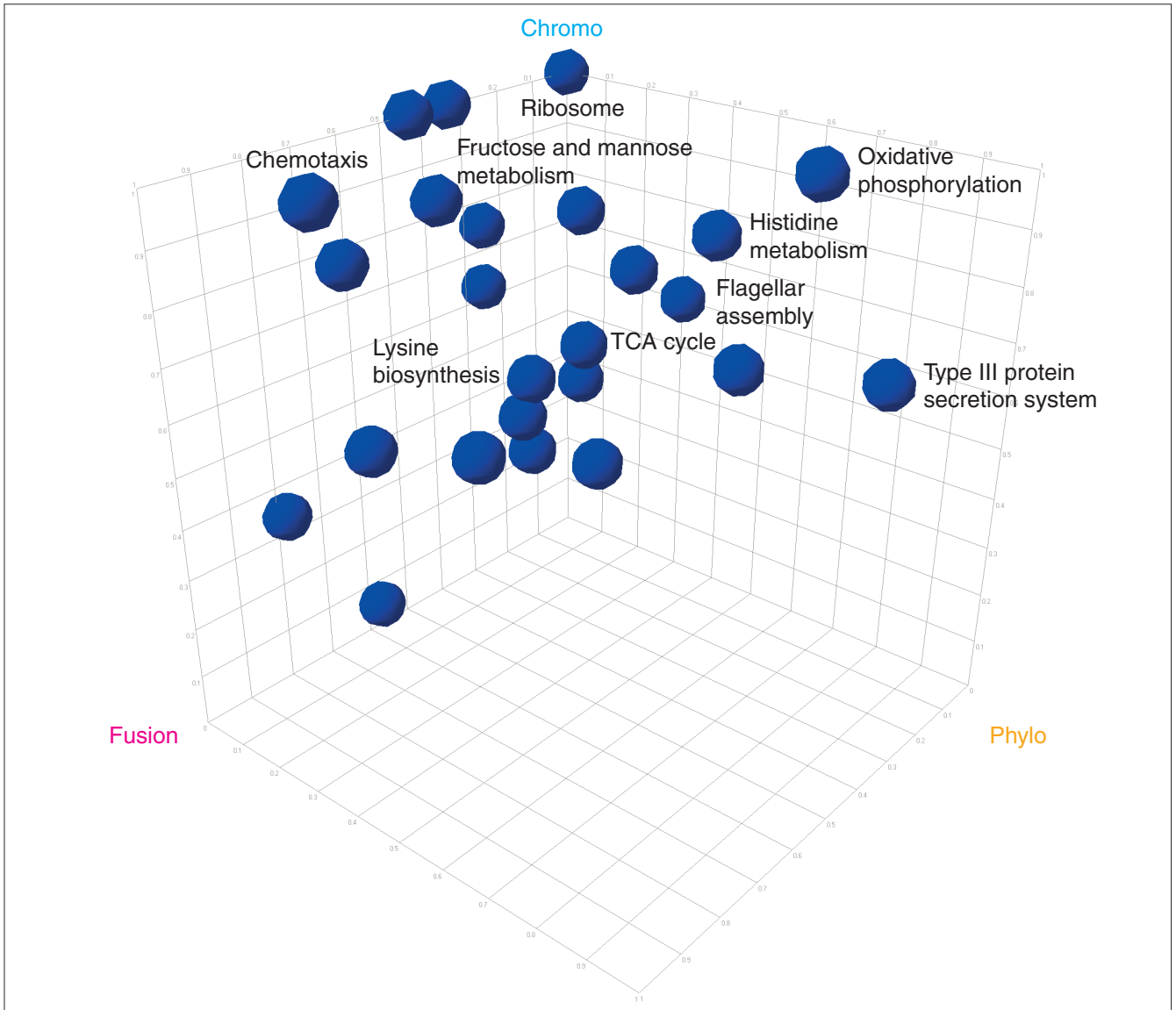

Figure 5
Combined networks reconstruct portions of known pathways that cannot be obtained by applying the methods independently. The blue spheres correspond to clusters of genes ascribed to a particular functional pathway (such as the ones described in Figure 4). The three-dimensional coordinates of the spheres correspond to the fraction of the clusters (in terms of nodes) that could have been recovered by each of the methods (the axes). The names of some of the pathways are shown. The *E. coli* genome was used and only clusters of seven or more genes are shown.

Structural properties of the combined comparative genomics networks

A striking property that emerges when the networks are combined is the formation of a giant cluster. This occurs for all but three genomes (*Mycoplasma pneumoniae*, *Mycoplasma genitalium*, and *Ureaplasma urealyticum*) perhaps because the number of links in these short genomes is very small. Although the number of nodes in the combined networks range from 400 to 1,600 (Figure 7a), depending on the genome, the percentage of nodes contained in a genome's largest cluster is relatively invariant, ranging from 80 to 90% (Figure 7b). These observations suggest that the properties of the combined networks have stabilized, and their general characteristics, as described below, should be relatively invariant against further increases in the number of functional links discovered. We note that random networks generated using the same number of nodes and edges as the chromo, fusion or combined networks, also form giant clusters, as is theoretically expected. In contrast to the very large clusters we find in chromo and fusion links, the largest cluster of phylogenetic links contains 19 ortholog families. This is not surprising because, by definition, all genes in a

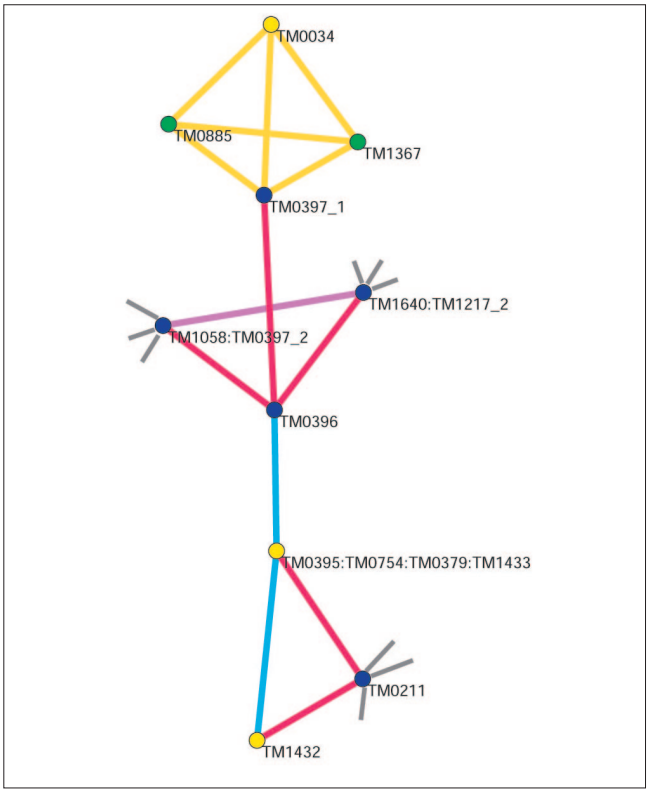

Figure 6
Ascribing function to uncharacterized genes on the basis of their locus in the network. A portion of the *T. maritima* network is shown. Genes of uncharacterized function are highlighted in green and genes with only a partial annotation in yellow. Genes involved in energy production are in blue. The color of the links is the same as in Figure 2. Gray lines indicate links with genes not shown. From the network locus of TM0885 and TM1367, we predict for the genes shown in blue a role in energy production.

phylo cluster must have the same phylogenetic profile. Thus, the probability of finding a cluster with a very large number of genes is low.

In general, many paths of different lengths connect each pair of nodes in the giant cluster. A useful measure of the global characteristics of the cluster is the minimum path length between each pair of nodes. It seems intuitively plausible that such minimum paths are the most biologically relevant of all paths connecting a pair. For each genome, the set of shortest path lengths has a Gaussian distribution. Its average is referred to as the characteristic path length. The characteristic path length averaged over the 43 genomes that have a giant cluster is 7, and the standard deviation about this average is 3. In other words, on average, seven comparative genomic links separate any two genes in a giant cluster. This is in contrast to a characteristic path distance of 5 ± 1 for random networks simulated with the same number of nodes and edges. Thus, the characteristic path distances of the combined networks are slightly larger than those of random networks but much smaller than the characteristic path of uniform lattice networks.

An example of a minimum path is shown in Figure 8. The *E. coli* gene *tyrA* of the phenylalanine, tyrosine and tryptophan biosynthesis pathway is separated from the *aspS* gene of the aminoacyl-tRNA biosynthesis and alanine and aspartate metabolism pathways by five links. This shortest path proceeds by way of four histidine metabolism genes. Each link along this path relates genes with a common functional pathway. However, as many genes are mapped to multiple functional pathways, paths in the network typically traverse many pathways. Snel *et al.* have proposed that such 'linker' genes be used to mark the boundaries between functional modules of genes in such networks [30].

Although global properties appear to be nearly random, local properties are not. In particular we expect local properties to be structured as each biological pathway that subserves a particular function invariably has several members, and the relationships between these members are likely to emerge through the comparative genomics methods applied here. One descriptive characteristic of a local environment is the clustering coefficient: the average probability that two genes linked to a common gene are also linked to each other [2]. This number is 1 for systems that are fully transitive (for example, networks constructed using phylogenetic profiling as applied here; see Figure 1) and becomes very small for random networks.

Not surprisingly, the combined networks for all 43 genomes show a significant amount of local clustering - or 'cliquishness' - when compared to random graphs. The chromosomal proximity and fusion networks both have a similar clustering coefficient of 0.24 and 0.25 respectively, as calculated for the giant component of the networks. When shuffling the networks (see Materials and methods) the clustering coefficients for these random graphs are 0.02 and 0.06 for chromo and fusion networks, respectively. The clustering coefficient increases to 0.36 for the combined network (0.02 for the shuffled networks), reflecting the high coefficient of the fully transitive phylogenetic graphs. The combination of high clustering coefficients and random-like characteristic path distance places these biological networks among the well studied class of so-called 'small-world' networks, to which social nets often conform [2,31].

Quantitative measures of the local (clustering coefficient) and global (characteristic path distance) properties of networks allow us to analyze the similarities and differences among all of the networks. Figure 7c shows the properties of the network mapped onto a two-dimensional space defined by the characteristic path distance and clustering coefficient. We find that networks of the same type (phylo, chromo, fusion or combined) cluster together, demonstrating the universality of network structures of each method. The

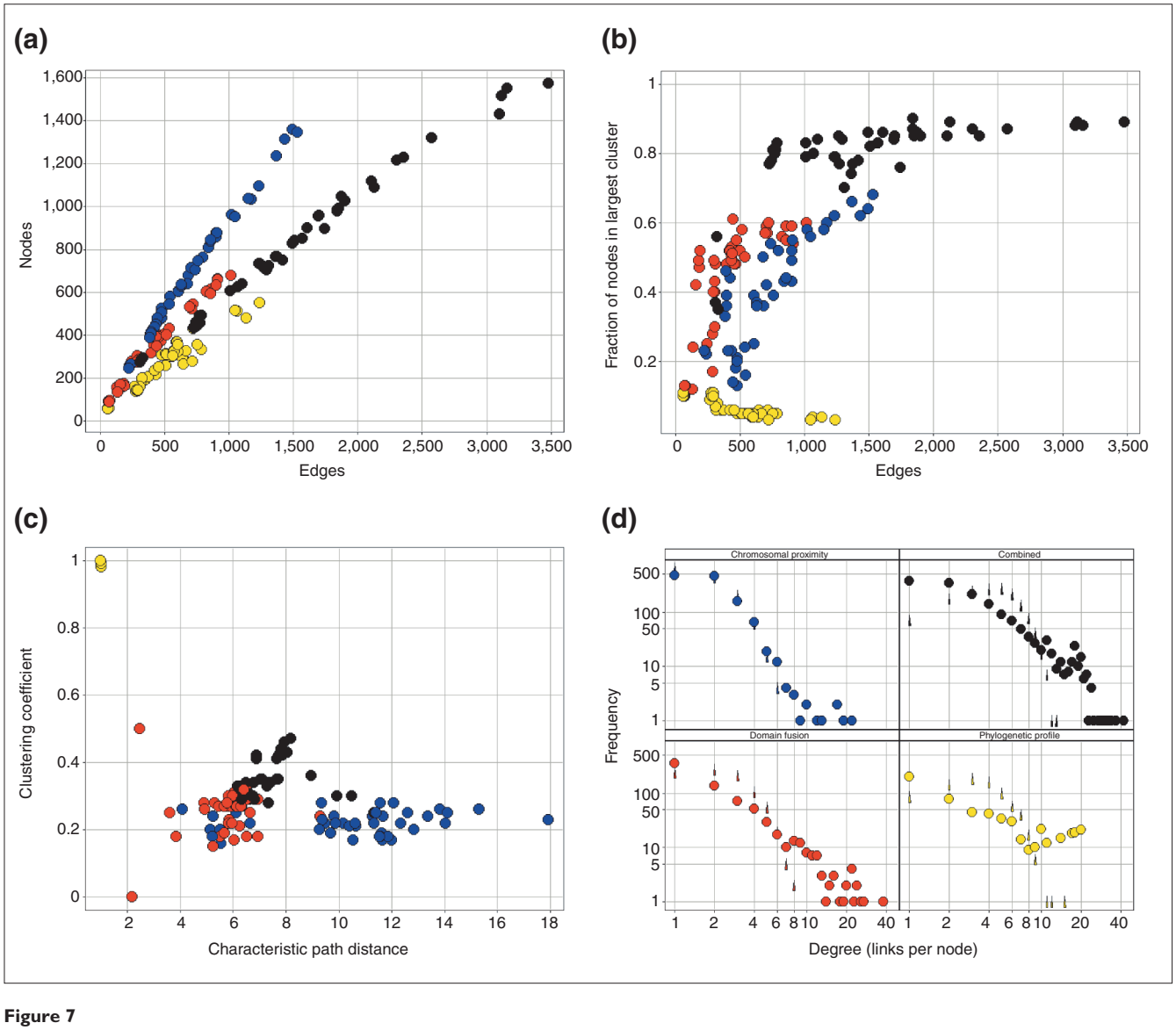

Figure 7
Network properties. **(a)** Basic characteristics. For each of the 43 genomes the number of nodes and edges of the four different networks based upon the three methods - chromosome proximity (red), domain fusion (blue), phylogenetic profiling (yellow) - and the combined networks (black) are shown. **(b)** Giant clusters. The networks are characterized by a giant cluster that relates a dominant fraction of the nodes in the network. In the combined networks of most of the genomes, the giant cluster accounts for 80-90% of the nodes, on average. **(c)** Universality. The characteristic path distance (global property) and clustering coefficient (local property) of the giant cluster of each network is mapped. Note that networks of the same method tend to cluster together. **(d)** Degree distributions. The histogram of the number of edges per node is shown on a log-log scale for each network type for *Pseudomonas aeruginosa*; similar distributions are observed for the other organisms. As in Figure 3, the circles denote values from the observed distributions while the triangles denote the degree distributions of the *de novo* random networks (see Materials and methods) with the same number of nodes and edges.

combined networks have a characteristic path that is roughly the average of the fusion and chromo networks and a clustering coefficient that is greater than both.

Furthermore, the networks appear to be scale-free in terms of the number of links per node (degree): a few nodes have many connections while most have few connections. As shown in Figure 7d, the distribution of links per node for the chromo, fusion and combined networks differs markedly from those of random graphs with the same number of nodes and edges (Figure 7d). While the power law distribution (linear regression on a log-log scale) holds well for the fusion and phylo graphs, the number of nodes of degree 1 in the chromo graph is less than would be expected from a power law, and is, indeed, not significantly larger than those of degree 2. The explanation lies in the operon organization of genes on bacterial and archaeal chromosomes, making two the typical number of chromo links of a conserved

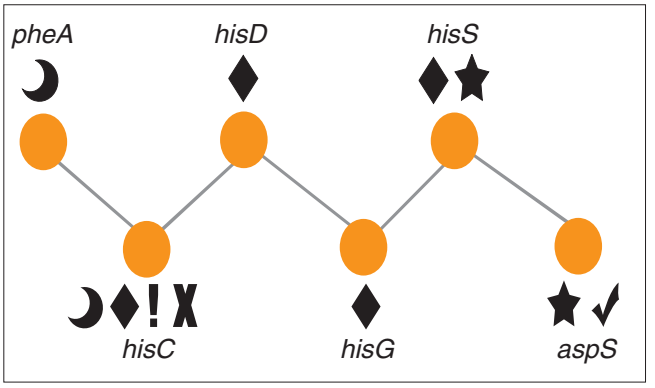

Figure 8
Global path through the networks. The nodes (genes) along this particular path of five links in the *E. coli* network are shown as circles. The symbols associated with each node represent the functional pathways in which the gene is annotated. Moon, phenylalanine, tyrosine and tryptophan biosynthesis; diamond, histidine metabolism; exclamation mark, phenylalanine metabolism; X, tyrosine metabolism; star, aminoacyl-tRNA biosynthesis; check mark, alanine and aspartate metabolism.

proximity of a gene with other genes - one on each end (see also chromo links in Figures 1 and 2).

We stress that the comparative genomic links do not necessarily correspond to direct protein-protein interactions, and thus are not expected to be detected by yeast two-hybrid methods. Links made *in silico* have been shown to be indicative of broader functional relationships [19,22]. In a recent study, Rain *et al.* identified over 1,200 interactions using a high-throughput application of the yeast two-hybrid method using approximately 15% of the *H. pylori* genes as 'baits' [32]. Of the 1,200 interacting pairs detected, only 17 correspond to links also found in our comparative genomic network. We note that low overlap between sets of links is even observed between high-throughput datasets [33,34] and may be explained by a lack of saturation of the complete functional relationship space, a high false-positive rate, and/or a bias towards a certain type of link [35].

Uncovering functional links between genes is a major step towards deducing the function of individual genes. Here we describe the properties of networks generated by the combination of three comparative genomics methods for 43 microbial genomes representing the three domains of life. We find that a giant 'small-world' cluster consistently includes 80% to 90% of the nodes in these networks, so that the average minimum path between any two genes is small, but local cliquishness is frequent relative to random networks. This structure for a society of genes, along with the observation that local order corresponds to genes of the same functional pathways, supports the notion that the network of relationships among a cell's genes is a set of highly cliquish functional systems interlinked by genes that are common to multiple pathways.

Materials and methods

Identification of comparative genomics links

We used three previously published methods for detecting functional links between proteins by comparative genomics. In this section we describe the implementation of these methods. As a pair of proteins can be coded as distinct genes in some genomes but fused as a multidomain protein in others (the basis for the fusion method), a comparative genomics study must adopt protein domains (instead of whole proteins) as the unit of analysis. Thus, multidomained proteins are split into domains (as they are at present in the COGs database, which clusters orthologous protein domains that are present in three or more lineages) and these, in turn, form the units of the comparative genomics links generated by the three methods. All the links are available in the Predictome database [36] and as additional data files (see Additional data files). These methods all depend upon detecting orthologs across genomes and, consequently, each node in any network corresponds to a cluster of orthologous groups [12,37] - a COG - with a representative sequence in that genome. Thus, the term 'gene' which, for brevity, is used throughout as indicative of a single node in the network, may in some cases be represented by multiple nodes.

Domain fusion links

Fusion links between protein domains were detected by a BLASTP [38] search of the 43 genomes included in the COGs database against nrdb90 [39], a non-redundant protein database. We deemed two protein domains - assigned to different COGs - to be fusion linked if each had an alignment of at least 80 residues to the same nrdb90 protein with a maximum expectation (E) value of 10^{-10} and with a maximum overlap of the two alignments of 20 residues. We then extrapolated the link between the domains to a link between their respective ortholog families (COGs) by applying the link to the common representatives of the two sets; that is, each member of the first COG is linked to those members of its partner COG that are of the same genome. This final step assigns links to domains that were undetected as the result of the high-stringency cutoff, and it therefore greatly increases the number of links. The COGs database is threshold-free and is instead build upon clusters of bidirectional best intergenomic matches [12]. Thus, the availability of this reliable ortholog family's database allows for the generalization of a fusion link from the level of domains to domain families while minimizing the possibility of false-positive alignments between the domains and their fusion.

Chromosomal proximity links

Chromo links were identified as recently described [19]. Genes A and B in genome X are linked by a chromosomal proximity link if they satisfy either of the two following conditions. They have a direct link, in which A and B are proximate (within 300 bp and transcribed in the same direction [16]) in X and their orthologs are also proximate in at least two other genomes corresponding to different phylogenetic

groups (as defined by COGs). Or, they have an inferred link, in which A and B are not proximate in X but their orthologs are proximate in at least three other genomes corresponding to different phyletic groups.

Phylogenetic profiling links

The 43 organisms are organized into 26 phyletic groups as defined by the COGs database. Two ortholog groups (COGs) are phylo linked if their phyletic distributions are identical (zero bit difference in [14]); that is, their 26-bit profiles of presence and absence are the same. Finally, two domains are phylo linked if their respective COGs are linked. The reliability of a phylo link is inversely related to how frequently the linked patterns are observed (data not shown). Thus, setting a lower threshold for the number of ortholog families that can have the same phyletic pattern can increase the quality of the links. However, a lower threshold, by definition, reduces the number of links. To strike a balance between the number of links and their functional correlation, a threshold of 29 was used. In other words, if a particular phyletic pattern corresponds to more than 29 orthologous groups, the pattern is considered uninformative for phylogenetic links.

Functional annotation of the networks

Functional correlation of the sets

To determine the functional correlation of a given network, observed (phylo, chromo, fusion or combined) or shuffled, with the classification of a given database (COGs or KEGG) we begin by collapsing the network to a list of links and selecting those links where both members are annotated (broad functional category in COGs [12] or pathway in KEGG [40]). The correlation of the network with the functional annotation is taken as the percentage of links in this set that is in the same category or pathway.

Functional pathways

All pathway information used to investigate local network structures was derived from the KEGG database [40].

Random networks

Two different methods for generating random networks were used. Both methods compare random networks with observed networks having the same number of nodes and edges.

Shuffling of existing network

To preserve the unique degree distribution of the observed network in its random counterpart, we shuffled the edges of the observed network according to the following algorithm [41]. We begin with the observed network and repeatedly (10,000 times) randomly choose two links in the observed network, $x_1 \leftrightarrow y_1$ and $x_2 \leftrightarrow y_2$, and rewire them to: $x_1 \leftrightarrow y_2$ and $x_2 \leftrightarrow y_1$.

De novo synthesis of random network

When analyzing the degree distribution of the observed networks, random networks are required to have the same number of nodes and edges as the observed networks but have a randomly generated degree distribution. We begin with N unlinked nodes and proceed by randomly choosing two nodes and adding an edge between them unless it already exists. The simulation ends when there are E edges in the random graph.

Additional data files

All the network files are available as additional data files with the online version of this paper, along with a text file description of the COG clusters (COGs), and instructions for interpreting the network files. The networks are organized according to type and genome. The 172 network files correspond to four networks (chromo, phylo, fusion and composite) for each of 43 genomes. Each network file lists the edges of the network. In the composite files there is an additional column to specify the nature of the link: 1, fusion; 2, chromo; 3, phylo; 4, chromo + fusion; 5, phylo + fusion; 6, chromo + phylo; 7, fusion + chromo + phylo.

Acknowledgements

We thank Adnan Derti, Ron Ophir, Carlos J. Camacho, Daniel Segré and Todd Silverstein for critical readings and helpful discussions. We thank Ivy Lee for work on an early stage of this study. This work was funded by a Whitaker Graduate Fellowship and by a Koshland Postdoctoral Fellowship to I.Y.

References

1. Bar-Yam Y: *Dynamics of Complex Systems*. London: Addison Wesley Longman; 1997.
2. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393:**440-442.
3. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286:**509-512.
4. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403:**623-627.
5. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
6. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
7. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415:**180-183.
8. Galperin, MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol* 2000, **18:**609-613.
9. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10:**366-370.
10. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405:**823-836.
11. Marcotte, EM: **Computational genetics: finding protein function by nonhomology methods.** *Curr Opin Struct Biol* 2000, **10:**359-365.
12. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
13. Gaasterland T, Ragan MA: **Constructing multigenome views of whole microbial genomes.** *Microb Comp Genomics* 1998, **3:**177-192.
14. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome**

analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.
15. Lawrence, JG: **Selfish operons and speciation by gene transfer.** *Trends Microbiol* 1997, **5:**355-359.
16. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
17. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
18. Fujibuchi W, Ogata H, Matsuda H, Kanehisa M: **Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping.** *Nucleic Acids Res* 2000, **28:**4029-4036.
19. Yanai, I, Mellor JC, DeLisi C: **Identifying functional links between genes using conserved chromosomal proximity.** *Trends Genet* 2002, **18:**176-179.
20. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
21. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.
22. Yanai I, Derti A, DeLisi C: **Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes.** *Proc Natl Acad Sci USA* 2001, **98:**7940-7945.
23. Enright AJ, Ouzounis CA: **Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions.** *Genome Biol* 2001, **2:**research0034.1-0034.7.
24. Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C: **Predictome: a database of putative functional links between proteins.** *Nucleic Acids Res* 2002, **30:**306-309.
25. Wuchty S: **Scale-free behavior in protein domain networks.** *Mol Biol Evol* 2001, **18:**1694-1702.
26. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes.** *J Mol Biol* 2001, **310:**311-325.
27. Apic G, Gough J, Teichmann SA: **An insight into domain combinations.** *Bioinformatics* 2001, **17:**S83-S89.
28. Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10:**1204-1210.
29. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402:**83-86.
30. Snel B, Bork P, Huynen MA: **The identification of functional modules from the genomic association of genes.** *Proc Natl Acad Sci USA* 2002, **99:**5890-5895.
31. Amaral LA, Scala A, Barthelemy M, Stanley HE: **Classes of small-world networks.** *Proc Natl Acad Sci USA* 2000, **97:**11149-11152.
32. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al.*: **The protein-protein interaction map of *Helicobacter pylori*.** *Nature* 2001, **409:**211-215.
33. Hazbun TR, Fields S: **Networking proteins in yeast.** *Proc Natl Acad Sci USA* 2001, **98:**4277-42778.
34. Tucker CL, Gera JF, Uetz P: **Towards an understanding of complex protein networks.** *Trends Cell Biol* 2001, **11:**102-106.
35. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417:**399-403.
36. **Predictome** [http://predictome.bu.edu]
37. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.
38. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
39. Holm L, Sander C**: Removing near-neighbour redundancy from large protein sequence collections.** *Bioinformatics* 1998, **14:**423-429.
40. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
41. Newman ME, Strogatz SH, Watts DJ: **Random graphs with arbitrary degree distributions and their applications.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 2001, **64:**026118.
42. **Pajek: package for large network analysis** [http://vlado.fmf.uni-lj.si/pub/networks/pajek/]