nature biotechnology

Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages

Shailesh V Date[1] & Edward M Marcotte[1,2]

We introduce a general computational method, applicable on a genome-wide scale, for the systematic discovery of uncharacterized cellular systems. Quantitative analysis of the coinheritance of pairs of genes among different organisms, calculated using phylogenetic profiles, allows the prediction of thousands of functional linkages between the corresponding proteins. A comparison of these functional linkages to known pathways reveals that calculated linkages are comparable in accuracy to genome-wide yeast two-hybrid screens or mass spectrometry interaction assays. In aggregate, these linkages describe the structure of large-scale networks, with the resulting yeast network composed of 3,875 linkages among 804 proteins, and the resulting pathogenic *Escherichia coli* network composed of 2,043 linkages among 828 proteins. The search of such networks for groups of uncharacterized, linked proteins led to the identification of 27 novel cellular systems from one nonpathogenic and three pathogenic bacterial genomes.

We present a method for the discovery of uncharacterized cellular systems by searching genome-wide networks of predicted protein interactions. The key element of this scheme is the use of computational techniques for the reconstruction of genome-wide protein networks. Computational approaches for finding gene and protein interactions[1–12] complement and extend experimental approaches such as synthetic lethal and suppressor screens[13], yeast two-hybrid experiments[14,15] and high-throughput mass spectrometry interaction assays[16,17]. One promising computational genetics approach involves the use of protein phylogenetic profiles[3] for discovering functional linkages between proteins. Phylogenetic profiles are essentially descriptions of the pattern of distribution of a given gene in the set of organisms with sequenced genomes. Proteins with similar phylogenetic profiles are often components of the same pathway[3,8,18].

Our method to discover cellular systems involves three steps (**Fig. 1**). First, genome-wide protein networks are derived using computational genetics[1,2] approaches after an initial calibration against known systems. Second, the reconstructed networks are examined for discrete clusters composed largely of uncharacterized proteins lacking clear functional assignments. Third, suitable candidate clusters are extended to include proteins appropriately linked to the cluster components, including operon partners and proteins more closely linked to the cluster than to others.

To test this method, we first quantified the utility of phylogenetic profiles for reconstructing pathways in general. After calibrating phylogenetic profiles, we reconstructed the genome-wide protein networks of four bacteria. The method of **Figure 1** was systematically applied to the networks of *Vibrio cholerae* biovar ElTor, *Caulobacter crescentus* CB15, *Staphylococcus aureus* N315 and *Pseudomonas aeruginosa* PA01, resulting in the discovery of 27 novel cellular systems and pathways, of which 7 are described here (see **Supplementary Fig. 1** online for the complete list).

RESULTS

Benchmarking the reconstruction of pathways

To reconstruct protein networks using phylogenetic profiles, we required a quantitative measure of how similar two profiles should be to occur in the same pathway. Logically, such a measure should take into account the complexity of the profiles to avoid simply linking proteins present in or absent from all organisms. Calculation of the mutual information between the profiles[19–21] (treating the occurrences of a gene among genomes as sampled values of a variable) has this property. We performed three tests of this measure and confirmed its ability to effectively reconstruct functional linkages between proteins.

We first evaluated the tendency of proteins' phylogenetic profiles to be similar by chance, using a mutual information metric. The phylogenetic profiles of all pairs of proteins in a genome were compared, and the distributions of mutual information values between pairs of phylogenetic profiles were plotted (**Fig. 2a**). Comparison to results of the same test after shuffling the values of each phylogenetic profile reveals that many more pairs of proteins with significantly similar profiles exist than might be expected at random, although it is possible that shuffling may disrupt phylogenetic as well as functional relationships. Whereas shuffled profiles rarely match each other with scores as high as 0.7, actual profiles show values as high as 1.2, indicating many pairs of significantly coinherited genes, whose precise statistical significance is determined from the corresponding random curves. Very similar results were obtained in this analysis for each of the seven genomes (chosen to represent both prokaryotic and eukaryotic organisms),

[1]Center for Computational Biology and Bioinformatics, Institute for Cellular and Molecular Biology, and [2]Department of Chemistry and Biochemistry, 1 University Station A4800, Austin, Texas 78712-1064, USA. Correspondence should be addressed to E.M.M. (marcotte@icmb.utexas.edu).

Published online 17 August 2003; doi:10.1038/nbt861

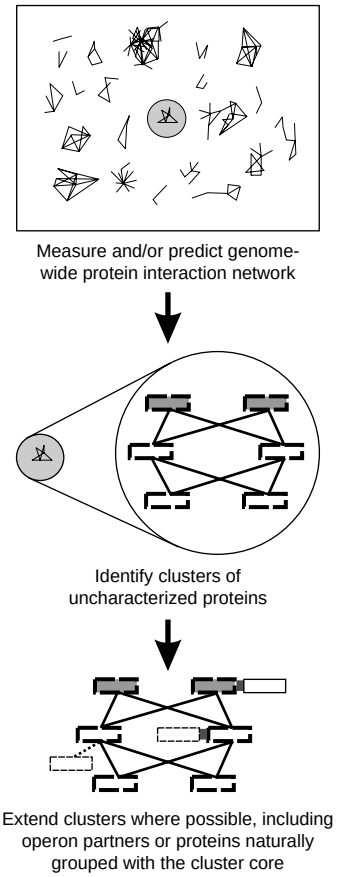

Figure 1 A schematic representation of the systematic method for identification of novel cellular systems. Using computational genetics, the genome-wide protein network of an organism is reconstructed. Suitable candidate clusters that contain three or more linked proteins, at least 50% of which are uncharacterized, are selected for further evaluation. Such core clusters are then extended to include operon partners and other proteins that are naturally linked with the protein cluster. Thick boxes and lines indicate proteins in the core cluster; thin boxes and lines indicate proteins extending the core cluster. Shaded boxes represent homologs; thick gray lines represent links to operon partners. See **Figure 5** for more details.

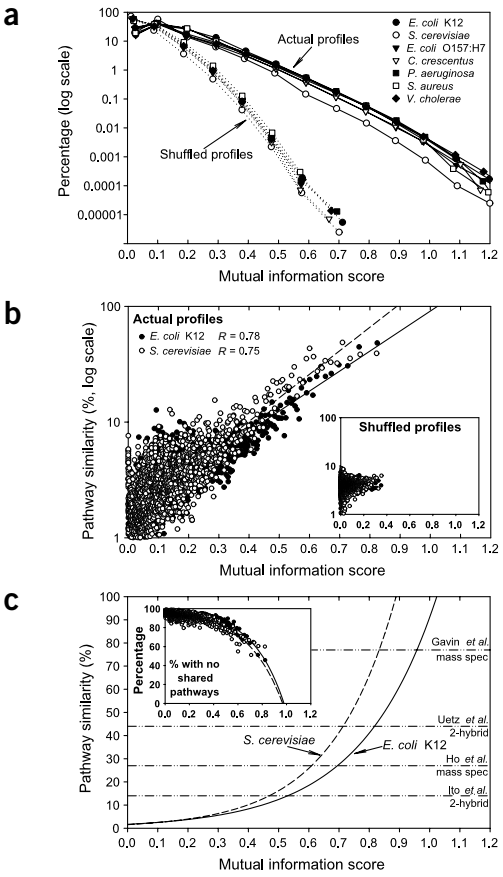

Figure 2 Two measures of the quality of functional linkages are presented. (**a**) The inherent information in phylogenetic profiles can be seen from the distributions of scores from comparisons of all possible protein pairs in each of seven organisms. Pairwise comparisons of actual phylogenetic profiles (solid lines) show significantly more similar profiles (indicated by larger mutual information values) than pairwise comparisons of shuffled profiles (dashed lines). Mutual information scores between shuffled profiles exceed 0.7 at a rate of ~1 in 10^7 pairs, whereas scores between actual profiles are greater than 1.2, indicating that scores above 0.7 are statistically likely to indicate legitimate functional linkages between pairs of genes. Values on the *x* axis represent proportions of comparisons in 0.1 mutual information intervals. (**b**) 1,131 *S. cerevisiae* (open circles) and 1,231 *E. coli* K12 (closed circles) proteins whose functions are precisely known were used to test the quality of the phylogenetic profile linkages. The quality of predicted functional linkages, measured as the mutual information scores between all pairs of phylogenetic profiles, is plotted versus the agreement between the proteins' experimentally known pathways, measured as the Jaccard coefficient between the proteins' pathway memberships in the KEGG database[24]. Each point represents the average values for 1,000 pairs of proteins. The solid and dashed curves indicate the performance of the *E. coli* and yeast phylogenetic profiles, respectively. Shuffled profiles rarely show high mutual information values (inset). (**c**) Mutual information scores plotted versus pathway similarity on a linear scale show increasing trends. The solid and dashed lines represent analytical curves fit to the data of **b** by least squares. Scores of ~0.75 indicate approximately 35–50% accurate predictions by this test, higher scores approach 100% functional accuracy. For comparison, the percentage of proteins that share no pathways in common show a decreasing trend, as mutual information values increase (inset). The accuracies of experimentally determined protein interactions from large scale yeast two-hybrid screens[14,15] indicating ~14% and 44% accuracies, and mass spectrometry experiments[16,17] indicating ~27% and ~76% accuracies are shown with the dot-dashed horizontal lines. As in **b**, each point represents the average values of 1,000 pairs of proteins.

suggesting that mutual information is a very well-behaved measure for the purpose of comparing phylogenetic profiles.

We then calibrated this measure by explicitly measuring its ability to reconstruct known cellular pathways. We compared phylogenetic profiles of all pairs of proteins chosen from among a subset of 1,231 *E. coli* K12 (ref. 22) or 1,131 yeast[23] proteins whose functions were known and recorded in the KEGG[24] pathway database. **Figure 2b** shows the results of this test, plotting the similarity of pairs of protein phylogenetic profiles versus the similarity of their pathway membership. Higher mutual information values correlate with increasing functional similarity; mutual information values of ~0.75 represent ~35–50% functional similarity, whereas mutual information values >~0.95 indicate a 100% chance, within error, of two proteins being in the same pathway. Shuffling the profiles reduces the functional similarity to ~5% on average (**Fig. 2b**, inset). **Figure 2c** shows a comparison with yeast protein interactions[25] experimentally measured with other methods. The two largest genome-wide yeast two-hybrid assays[14,15] showed accuracies in this test of 14% and 44%, respectively, whereas the two largest mass spectrometry interaction assays

showed accuracies of 27% and 76% (refs. 16,17). As a second means of calibration, we measured the percentage of pairs of annotated proteins that showed no functional similarity; as mutual information values increase, the frequency of such protein pairs falls off strongly (**Fig. 2c**, inset).

Finally, we empirically examined reconstructions of individual pathways. An example from the genome of *E. coli* K12 is shown in **Supplementary Table 1** online. When we compared the phylogenetic profile of the flagellar biosynthesis protein FlgD against profiles of all other proteins in *E. coli*, then ranked the proteins in decreasing order of mutual information scores, we reconstruct much of the flagellar biosynthesis system. Of the ten *E. coli* proteins with mutual information values larger than 0.7, all ten (100%) are proteins of the flagellar biosynthesis pathway. Out of the next ten proteins scoring below 0.7, six are also directly involved in flagellar biosynthesis, whereas three of the remaining four are membrane proteins that may still be connected to flagellar biosynthesis.

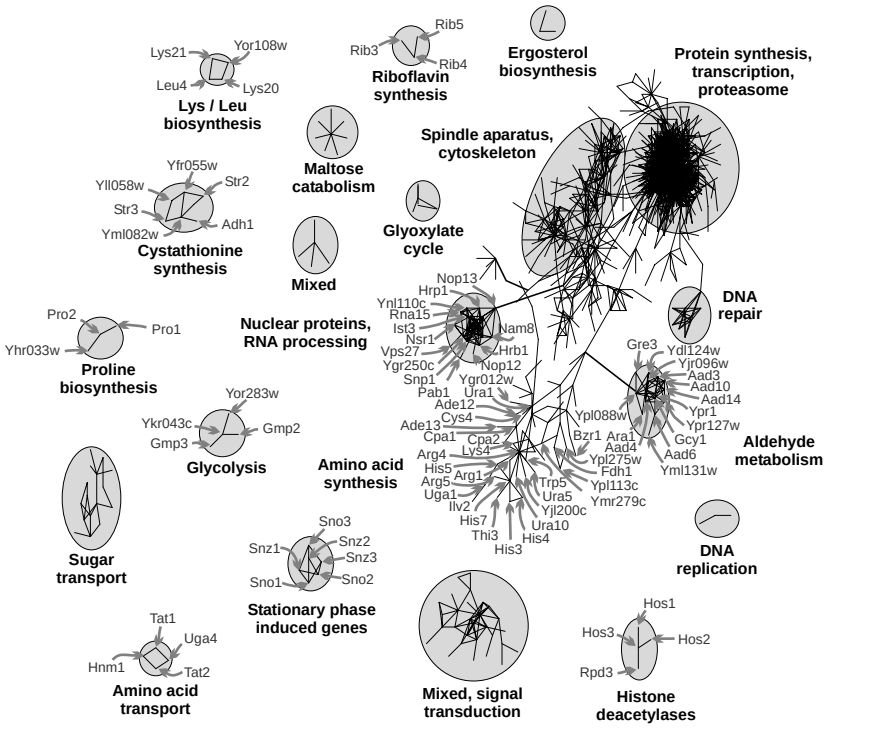

Figure 3 Predicted genome-wide protein networks for yeast[23]. Proteins are represented as vertices, and derived functional linkages are shown as lines connecting the corresponding proteins. All linkages with scores above a mutual information value of 0.75 are drawn, essentially by modeling the linkages as springs that pull functionally linked proteins together on the page. (Thus, the lengths of the lines are not meaningful, only the connections). Groups of proteins sharing functional links are seen to cluster together, representing portions of genetic or functional networks. Systems in gray circles are labeled with their corresponding functions. For visual clarity, small protein networks, including 1 five-protein system, 2 four-protein systems and 31 two-protein systems, have been omitted.

Reconstructing genome-wide protein networks

Using this calibration of phylogenetic profiles, we applied our method to the reconstruction of genome-wide protein networks. First, the phylogenetic profiles of all proteins in yeast or pathogenic *E. coli* O157:H7 were compared against each other. Functional linkages were identified between pairs of proteins whose profiles matched above a given threshold mutual information value. In pathogenic *E. coli*, 7,245 unique linkages between 1,472 proteins were found scoring above 0.75, a value expected at random with a probability $<1\times10^{-8}$ (determined from **Fig. 2a**), whereas 2,043 unique linkages between 828 proteins were found scoring above 0.85, expected at random with a probability $<1\times10^{-9}$. Similarly, in yeast, 3,875 unique linkages between 804 proteins were found scoring above 0.75. **Figures 3** and **4** show the networks of yeast and pathogenic *E. coli* plotted at mutual information value thresholds of 0.75 and 0.85, respectively, corresponding to ~40–50% functional similarity by the test of **Figure 2c**.

Examination of the networks reveals that functionally linked proteins correctly cluster together, representing a portion of the genome-wide network of proteins, with many interconnected functional systems. In both yeast and *E. coli*, many known pathways are reconstructed, such as the amino acid biosynthesis pathways highlighted in detail in both organisms, and the urease subunits found in pathogenic *E. coli* but not in *E. coli* K12. The compact subnetwork of four linked *E. coli* proteins (YaeM, YgbB, GcpE and LytB) involved in the recently established isoprenoid biosynthesis pathway[26–29] is also evident.

As this method is based solely on sequence and is unbiased with respect to knowledge about protein function, it can both assign function to uncharacterized proteins that appear to be linked to systems of known function and uncover potentially new cellular systems and pathways. In the reconstructed yeast network, uncharacterized proteins YDL124W and YML131W can be assigned potential roles in aldehyde metabolism, based on their associations with the Aad proteins (Aad3,4,6,10,14) (ref. 30) and Gre3p[31] (**Fig. 3**). Similarly, proteins YPL113C and YMR279C can be implicated in amino acid synthesis, and YGR250C can be implicated in RNA processing or splicing. We can speculate that one putative *E. coli* system (**Fig. 4**), composed largely of unknown proteins (YeaJ, YhdA, Z2836, YeaI, YliF and YneF) containing the GGDEF domain, is involved in signal transduction, based on the observed participation of GGDEF-domain-containing proteins in two-component signal transduction systems[32].

Systematic search for novel cellular systems

Given the relatively accurate reconstruction of protein networks using this approach, we calculated genome-wide networks for four bacteria by comparing the phylogenetic profiles of all proteins with each other for a given organism. The extent of these networks is described in **Table 1**. Using the method of **Figure 1**, these networks were examined for the presence of linked protein clusters composed of three or more components, in which 50% or more of the component proteins lacked functional assignments. We were able to identify 27 clusters from the reconstructed networks of *V. cholerae* biovar ElTor, *C. crescentus* CB15, *S. aureus* N315 and *P. aeruginosa* PA01 that satisfied the above criteria (see **Supplementary Fig. 1** online for the complete list). In **Figure 5**, we describe seven of the clusters (three from *S. aureus*, two from *P. aeruginosa* and one each from *V. cholerae* and *C. crescentus*), each extended to include operon partners and proteins that were judged to be most appropriately linked to the seed cluster but had mutual information values slightly below the cluster threshold.

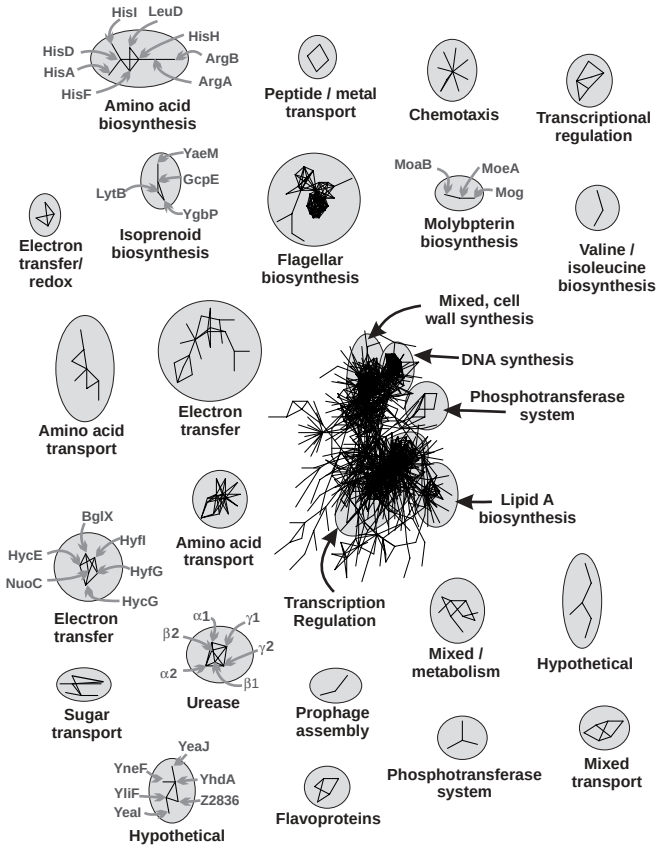

Figure 4 Predicted genome-wide protein networks for pathogenic *E. coli* O157:H7 (ref. 46). (See Figure 3 legend for details.) All linkages with scores above a mutual information value of 0.85 are included. For visual clarity, small protein networks, including 1 six-protein system, 2 four-protein systems, 9 three-protein systems and 40 two-protein systems have been omitted.

V. cholerae cluster A represents one of the smallest clusters described with four uncharacterized proteins, two of which are included in the cluster core. The two characterized proteins in the cluster, VCA0380 and SoxR, are transcriptional regulators. The function of SoxR in regulating genes that respond to oxidative stress[33] might indicate a role in stress response.

Cluster B from *C. crescentus* also includes two unknown proteins in a three-member cluster core. These in turn are linked to five characterized proteins, prominent among which are the ExbD/TolR family, MotA/TolQ/ExbB family, HlyB family and cation efflux family proteins, broadly indicating involvement in membrane transport and uptake of receptor bound substrates like colicin[34–36]. Among other characterized proteins in the cluster, LpxK and KdtA are known to be involved in the biosynthesis of lipopolysaccharides[37,38], supporting the role of this system as a membrane-associated metabolite transport or uptake system.

Three uncharacterized proteins comprise the cluster core of *S. aureus* cluster C, with secondary links to three uncharacterized and four characterized proteins. Whereas the presence of the two homologs of the acetyl-CoA carboxylases accB and accC points toward a possible involvement in fatty acid synthesis[39], secondary links to holin-like proteins[40] may indicate a general association with the cell membrane. Of the two remaining clusters from *S. aureus,* proteins in cluster D appear to be involved in cation transport, whereas the presence of dihydrofolate reductase and DNA polymerase IIIα in cluster E indicates a likely function in nucleic acid synthesis.

P. aeruginosa cluster F, with seven entirely uncharacterized proteins and three weakly characterized proteins, represents the cluster with the maximum number of uncharacterized proteins identified. Secondary links to PA2837, a member of the outer membrane efflux family (OEP family), DadA, a protein involved in alanine catabolism localized to the inner membrane, and a probable outer membrane protein all point toward an association of the cluster members with the cell membrane. PA2838 is tentatively included in the cluster because there are only 37 nucleotides between it and PA2837, which places PA2838 at the outer limits of inclusion as an operon partner[11]. *P. aeruginosa* cluster G is also one of the smallest clusters described, comprising one characterized and two uncharacterized proteins in the cluster core, out of the total six. Three proteins in the cluster had partial functional information, including homology to an aromatic acid decarboxylase, an aromatic hydrocarbon reductase, and UDP-*N*-acetylmuramate: L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase, involved in the synthesis of cell walls and murein-tripeptide recycling[41]. These functions suggest that this pathway operates in the metabolism of cell walls or of nonstandard amino acids.

The phylogenetic distributions of the novel cellular systems (**Fig. 6**) indicate that these systems are broadly conserved among organisms. Each pathway shows a distinct pattern of distribution. Cluster F and G are the most highly conserved systems, present in 34–35 genomes including most of the archaea. In total, the pathways represent functional information for ~1,100 genes from the 57 genomes.

DISCUSSION

A large number of proteins from the available genomes currently have no known functions. For instance, a survey of the *E. coli* genome[22] reveals that 1,342 of the total 4,279 genes (~31%) are currently annotated as encoding hypothetical proteins. Many of these uncharacterized proteins can be implicated in new pathways using the method described here. The method involves searching reconstructed genome-wide protein networks for linked, uncharacterized protein clusters. This approach is in contrast to earlier approaches, which attempted to link uncharacterized genes to characterized systems to determine their function[8]. Although we report 27 novel systems, we expect the actual number to be much higher, as our ability to determine the exact number is limited by the coverage of *in silico* network reconstructions. The method currently finds links between 5–40% of the genes in a genome, leading to reconstructed networks that are reasonably sparse (**Table 1**). The availability of additional interaction and pathway data should considerably improve the mapping of novel cellular systems.

Alternate approaches for constructing phylogenetic profiles can also be imagined[42,43]. For instance, profiles based on the detection of orthologs, rather than homologs, would intuitively seem to offer a better chance of pathway reconstruction. This approach, however, is far from trivial, as high-throughput methods for estimating orthologs are imperfect and often impractical. COGs (clusters of orthologous groups of proteins), one of two primary approaches, is known not to strictly identify orthologs, and bidirectional best hits, the other approach, fails for every Rosetta Stone fusion protein[44]. In contrast, homologs can be easily identified and can be effectively used in the reconstruction of pathways (**Fig. 2**). The incorporation of sequence divergence into the phylogenetic profiles is intended to compensate partly for the lack of knowledge about orthology. Note that this approach gives partial information about orthology, as the absence of homologs implies absence of orthologs, by definition. We suspect that

Table 1 Sizes of gene networks reconstructed for each organism at increasing accuracies, as measured by the minimal mutual information score for each linkage in a network

Genome	Proteins in genome	Proteins with KEGG annotation	Mutual information score threshold	Proteins in network (% of proteome)	Functional linkages	Annotated proteins in network (% total annotated proteins)
E. coli K12	4,405	1,231	0.7	1,751 (39.7%)	12,874	639 (51.9%)
			0.75	1,408 (31.9%)	7,049	528 (42.8%)
			0.85	784 (17.7%)	2,008	291 (23.6%)
E. coli O157H7	5,283	991	0.7	1,827 (34.5%)	13,121	499 (50.3%)
			0.75	1,472 (27.8%)	7,245	403 (40.6%)
			0.85	828 (15.6%)	2,043	241 (24.3%)
S. cerevisiae	6,343	1,131	0.7	1,053 (16.6%)	7,630	354 (31.2%)
			0.75	804 (12.6%)	3,875	286 (25.2%)
			0.85	359 (5.6%)	917	114 (10.0%)
C. crescentus	3,737	734	0.7	1,256 (33.6%)	6,541	391 (53.2%)
			0.75	998 (26.7%)	3,697	328 (44.6%)
			0.85	555 (14.8%)	1,112	197 (26.8%)
P. aeruginosa	5,565	1,205	0.7	2,260 (40.6%)	23,729	672 (55.7%)
			0.75	1,880 (33.7%)	13,193	561 (46.5%)
			0.85	1,145 (20.5%)	3,884	341 (28.2%)
S. aureus	2,594	574	0.7	848 (32.6%)	4,603	266 (46.3%)
			0.75	661 (25.4%)	2,604	211 (36.7%)
			0.85	332 (12.7%)	667	122 (21.2%)
V. cholerae	3,828	898	0.7	1,394 (36.4%)	10,466	463 (51.5%)
			0.75	1,177 (30.7%)	5,876	390 (43.4%)
			0.85	699 (18.2%)	1,712	232 (25.8%)

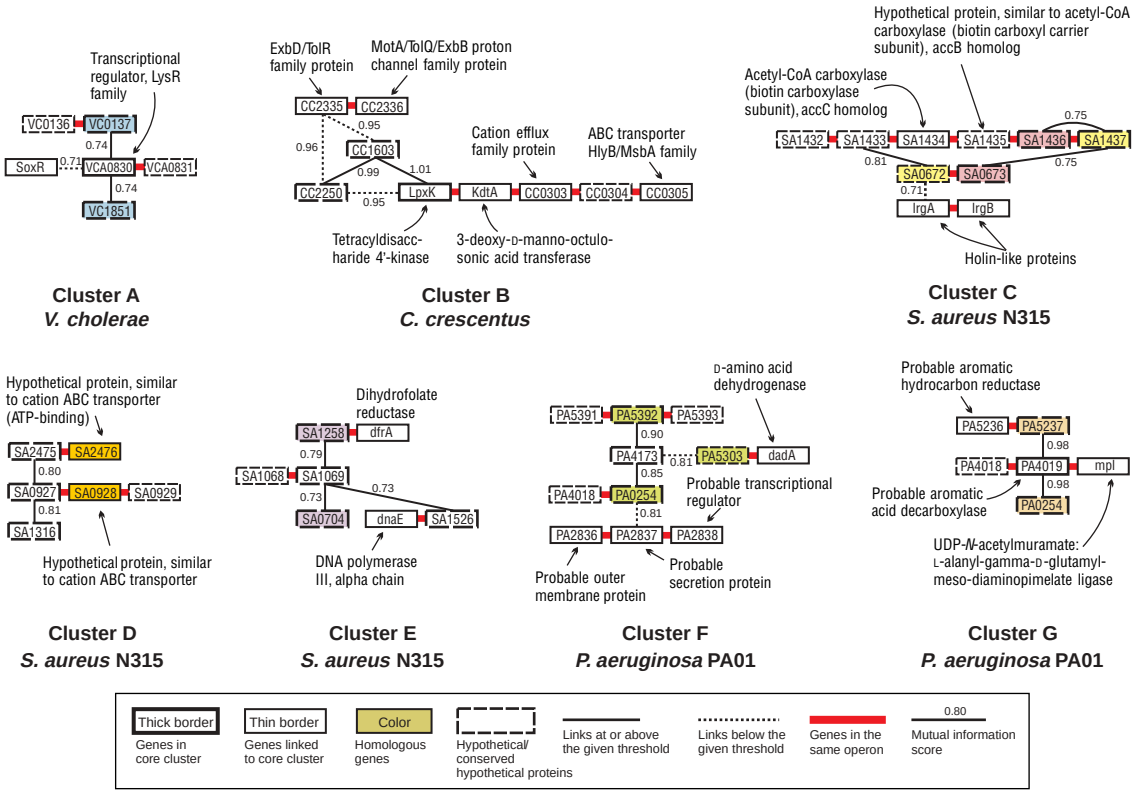

Figure 5 Clusters representing potentially new pathways selected from reconstructions of genome-wide interaction networks of four different organisms. Boxes with thicker borders, and bold lines denote the cluster core. Each cluster was extended to include operon partners, as well as secondarily linked proteins that are naturally grouped with the proteins in the cluster but with a mutual information value less than the selected threshold; these are represented by dotted lines and boxes with thinner borders. Thick red lines represent connections between genes in an operon, whereas colored boxes represent homologous proteins. All selected core clusters are composed of proteins, at least 50% of which lack precise functional assignments. Boxes with dashed outlines represent such uncharacterized proteins.

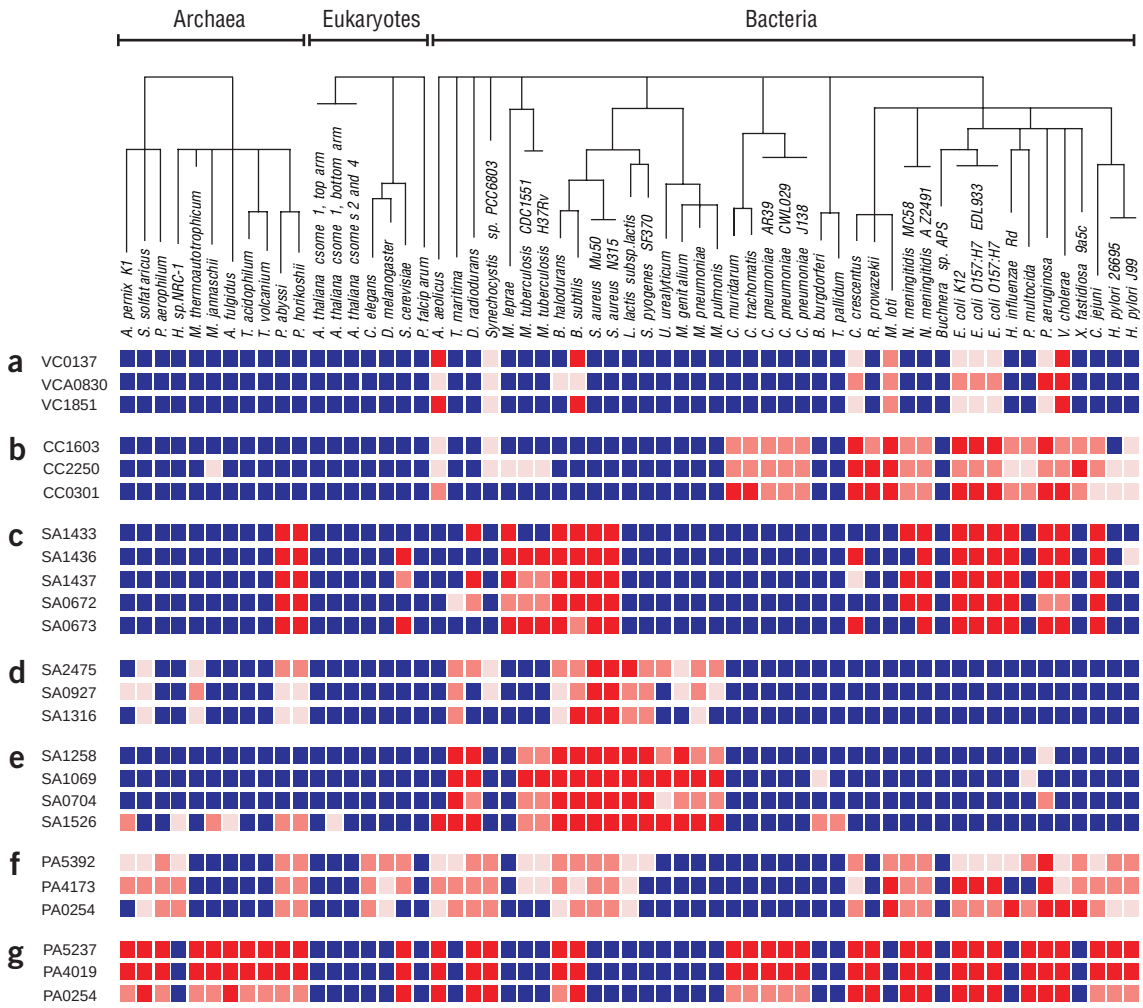

Figure 6 The phylogenetic profiles drawn for the core components of the gene clusters in **Figure 5**. The genes corresponding to proteins within a cluster show similar patterns of presence and absence, indicated by red and blue squares, respectively, among the 57 genomes, labeled across the top. The intensity of red denotes the degree of homology between the protein labeled at the left with the best matching protein sequence of the corresponding genome. Deeper red indicates stronger sequence similarity, blue indicates no detectable similarity (BLAST E-value ≥ 1). To enhance interpretation, genomes are arranged according to broad phylogenetic relationships, indicated above their names (*http://tolweb.org/tree/phylogeny.html*).

cellular systems resulting from recent horizontal transfers will be detected well using this approach.

We can speculate as to why one should expect to find entirely new systems. In well-characterized systems like yeast, ~90% of the uncharacterized proteins are linked in networks to proteins of known function. Most uncharacterized proteins therefore appear to be additional components of known systems. However, in examining the novel cellular systems of **Figure 5**, the few characterized proteins seem to be strongly biased towards metabolic functions, which occur commonly as more or less discrete systems within cells, easily capable of being coinherited or horizontally transferred, and therefore lend themselves to the type of analysis described here. Such an analysis may preferentially reveal new cellular systems of this type, including biosynthetic pathways, systems involved in cell wall and cell envelope synthesis, and degradative systems. However, our analysis does not indicate the precise nature of the systems, which may include pathways, structural complexes and regulatory networks.

A second argument suggests that many more pathways are still unknown. Before the availability of genome sequences, it was the norm to perform experiments on known systems or logical extensions of these systems. The effect of this strategy was to extend biological knowledge gradually across an organism's set of pathways, rather than sampling the pathways evenly. We might therefore expect there to be undiscovered pathways even in well-characterized organisms, especially when the pathways and systems are not obviously connected to known systems. The approach presented here avoids this bias in pregenomic experiments: biological networks are derived independently of whether the proteins have been characterized. In contrast to pregenomic experiments, this allows discoveries about pathways, old and new, to proceed systematically.

METHODS

Calculation and clustering of phylogenetic profiles. The amino acid sequences of all 174,901 known proteins from 57 different organisms were obtained from the National Center for Biotechnology Information (NCBI) Entrez Genome website. The amino acid sequences were first compared with each other using the NCBI Basic Local Alignment Search Tool (BLAST)[45] (174,901^2, or ~31 billion, comparisons). Phylogenetic profiles were constructed as follows: for each protein i, a vector was generated with elements p_{ij}, corresponding to each

organism j in the set of 57 reference organisms, where $p_{ij} = -1/\log E_{ij}$, with values of $p_{ij} > 1$ truncated to 1, to avoid logarithm-induced artifacts. E_{ij} represents the BLAST expectation value of the top-scoring sequence alignment between protein i and all of the proteins in the genome of organism j. Calculating the p_{ij} elements in this manner, rather than using binary values as originally proposed[3], captures different degrees of sequence divergence, providing more information than the simple presence or absence of genes, and requires no minimum threshold of similarity to be specified.

Quantification and assessment of functional linkage quality. As a metric of phylogenetic profile similarity, the mutual information[19–21] was calculated between pairs of phylogenetic profiles. The mutual information $MI(A,B)$ is maximum when there is complete covariation between the occurrences of the genes A and B, and tends to zero as variation decreases or the gene occurrences vary independently, calculated as:

$$MI(A,B) = H(A) + H(B) - H(A,B)$$

where $H(A) = -\Sigma p(a) \ln p(a)$ and represents the marginal entropy of the probability distribution $p(a)$ of gene A of occurring among the organisms in the reference database, summed over intervals in the probability distribution, and $H(A,B) = -\Sigma\Sigma p(a,b) \ln p(a,b)$ represents the relative entropy of the joint probability distribution $p(a,b)$ of occurrences of genes A and B across the set of reference organisms. In practice, mutual information is calculated on histograms of p_{ij} values, binned in 0.1 intervals, with resulting MI values ranging from 0–1.34 (highest mutual information value observed in seven genomes). The sets of phylogenetic profiles and mutual information-ranked linkages for the seven genomes are available at http://bioinformatics.icmb.utexas.edu/pathways/.

The KEGG pathway database[24] categorizes proteins into 158 categories, such as 'Glycolysis,' 'Ribosome' and 'MAPK signaling pathway.' These pathway names were associated with each protein in a genome and used to measure whether two proteins belonged to the same pathway. We calculated the Jaccard coefficient of their KEGG database pathway annotation[24] as follows:

$$\text{Pathway similarity} = 100 \times (|\,\text{KEGG}_A \cap \text{KEGG}_B\,|)/(|\,\text{KEGG}_A \cup \text{KEGG}_B\,|),$$

where KEGG_x is the set of specific KEGG pathways in which protein x is known to participate, and $|\text{KEGG}_x|$ is the number of unique pathways in the set.

For all tests except those of experimental interaction pairs, functional links between homologous proteins (defined as pairs of proteins whose amino acid sequences can be aligned with a BLAST expectation value $\leq 1 \times 10^{-5}$) were omitted. Varying this BLAST threshold between $E \leq 10^{-3}$ to $E \leq 10^{-5}$ showed little effect on algorithm performance (see **Supplementary Fig. 2** online).

Network visualization. To calculate networks, functional linkages were created between all proteins whose phylogenetic profiles matched with a mutual information score above a given threshold. A force-directed layout algorithm[8] was used to position the proteins on the page. First, proteins were placed at random initial coordinates on the page. Links were then modeled as springs. The resulting forces on each protein were calculated, and the proteins were moved in iterations to minimize the forces, with proteins finally settling into equilibrium positions adjacent to their functional partners.

Identification of uncharacterized cellular systems. Reconstructed networks were examined for the presence of discrete clusters containing three or more proteins, at least 50% of which carried no functional assignments. These protein components were designated as core components, linked together by primary links. The lowest mutual information value for a protein pair among all pairs in a cluster represented the threshold for each cluster.

Suitable candidate clusters were then further extended to include operon partners and secondary links to individual proteins or proteins in an operon that were more strongly linked to a protein in the cluster than to any other protein, but where the link strength was below the given threshold. Operon partners were assigned by examining the number of nucleotides separating adjacent genes. Genes separated by less than 40 nucleotides were assigned to the same operon, in accord with the Bayesian operon predictor[11].

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by grants from the Welch Foundation (F-1515), the Texas Advanced Research Program, a Camille and Henry Dreyfus New Faculty Award, National Science Foundation (EIA – 0219061) and a Packard Fellowship.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 14 March; accepted 24 June 2003
Published online at http://www.nature.com/naturebiotechnology/

1. Marcotte, E.M. Computational genetics: finding function by non-homology methods. *Curr. Opin. Struct. Biol.* **10**, 359–365 (2000).
2. Huynen, M., Snel, B., Lathe, W. & Bork, P. Exploitation of gene context. *Curr. Opin. Struct. Biol.* **10**, 366–370 (2000).
3. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
4. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901 (1999).
5. Tavazoie, S., Huges, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Gen.* **22**, 281–285 (1999).
6. Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O. & Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
7. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
8. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
9. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
10. Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yeates, T.O. Protein function in the post-genomic era. *Nature* **405**, 823–826 (2000).
11. Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. & Collado-Vides, J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA* **97**, 6652–6657 (2000).
12. Thompson, H.G.R, Harris, J.W, Wold, B.J, Quake, S.R & Brody, J.P. Identification and confirmation of a module of coexpressed genes. *Genome Res.* **12**, 1517–1522 (2002).
13. Tong, A.H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
14. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
15. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
16. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
17. Gavin, A. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
18. Pavlidis, P., Weston, J., Cai, J. & Grundy, W.N. Learning gene functional classifications from multiple data types. *J. Comp. Biol.* **9**, 401–411 (2002).
19. Shannon, C.E. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423, 623–656 (1948).
20. Krober, B.T.M, Farber, R.M., Wolpert, D.H. & Lapedes, A.S. Covariation of mutations in the V3 loop of human immunodeficiency virus type I envelope protein: an information theoretic analysis. *Proc. Nat. Acad. Sci. USA* **90**, 7176–7180 (1993).
21. Huynen, M., Snel, B., Lathe, W. & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204–1210 (2000).
22. Blattner, F.R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1 3–1474 (1997).
23. Goffeau, A. *et al.* The yeast genome directory. *Nature* **387**, Supplement (1997).
24. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
25. Xenarios, I. *et al.* DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res.* **29**, 239–241 (2001).
26. McAteer, S., Coulson, A., McLennan, N. & Masters, M. The lytB gene of *Escherichia coli* is essential and specifies a product needed for isoprenoid biosynthesis. *J. Bacteriol.* **183**, 7403–7407 (2001).
27. Cunningham, F.X. Jr., Lafond, T.P. & Gantt, E. Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis. *J. Bacteriol.* **182**, 5841–5848 (2000).
28. Takahashi, S., Kuzuyama, T., Watanabe, H. & Seto, H. A 1-deoxy-D-xylulose 5-phosphate reductoisomerase catalyzing the formation of 2-C-methyl-D-erythritol 4-phosphate in an alternative nonmevalonate pathway for terpenoid biosynthesis. *Proc. Natl. Acad. Sci. USA* **95**, 9879–9884 (1998).
29. Herz, S. *et al.* Biosynthesis of terpenoids: YgbB protein converts 4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate to 2-C-methyl-D-erythritol 2,4-cyclodiphosphate. *Proc. Natl. Acad. Sci. USA* **97**, 2486–2490 (2000).
30. Delneri, D., Gardner, D.C., Bruschi, C.V. & Oliver, S.G. Disruption of seven hypotheti-

cal aryl alcohol dehydrogenase genes from *Saccharomyces cerevisiae* and construction of a multiple knock-out strain. *Yeast* **15**, 1681–1689 (1999).
31. Traff, K.L., Jonsson, L.J. & Hahn-Hagerdal, B. Putative xylose and arabinose reductases in Saccharomyces cerevisiae. *Yeast* **19**, 1233–1241 (2002).
32. Galperin, M.Y., Nikolskaya, A.N. & Koonin, E.V. Novel domains of the prokaryotic two-component signal transduction systems. *FEMS Microbiol. Lett.* **203**, 11–21 (2001).
33. Amabile-Cuevas, C.F. & Demple, B. Molecular characterization of the soxRS genes of *Escherichia coli*: two genes control a superoxide stress regulon. *Nucleic Acids Res.* **19**, 4479–4484 (1991).
34. Gentschev, I., Dietrich, G. & Goebel, W. The *E. coli* α-hemolysin secretion system and its use in vaccine development. *Trends Microbiol.* **1**, 39–45 (2002).
35. Braun, V. & Braun, M. Active transport of iron and siderophore antibiotics. *Curr. Opin. Microbiol.* **2**, 194–201 (2002).
36. Bouveret, E. *et al.* Analysis of the *Escherichia coli* Tol–Pal and TonB systems by periplasmic production of Tol, TonB, colicin, or phage capsid soluble domains. *Biochimie* **84**, 413–421 (2002).
37. Garrett, T.A., Que, N.L. & Raetz, C.R. Accumulation of a lipid A precursor lacking the 4'-phosphate following inactivation of the *Escherichia coli* lpxK gene. *J. Biol. Chem.* **273**, 12457–12465 (1998).
38. Tzeng, Y.L., Datta, A., Kolli, V.K., Carlson, R.W. & Stephens, D.S. Endotoxin of *Neisseria meningitidis* composed only of intact lipid A: inactivation of the meningococcal 3-deoxy-D-manno-octulosonic acid transferase. *J. Bacteriol.* **184**, 2379–2388 (2002).
39. Rodriguez, E., Banchio, C., Diacovich, L., Bibb, M. & Gramajo, H. Role of an essential acyl coenzyme A carboxylase in the primary and secondary metabolism of *Streptomyces coelicolor* A3(2). *Appl. Environ. Microbiol.* **9**, 4166–4176 (2001).
40. Grundling, A., Manson, M. & Young, R. Holins kill without warning. *Proc. Natl. Acad. Sci. USA* **98**, 9348–9352 (2001).
41. Mengin-Lecreulx, D., van Heijenoort, J. & Park, J.T. Identification of the mpl gene encoding UDP-N-acetylmuramate: L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase in *Escherichia coli* and its role in recycling of cell wall peptidoglycan. *J. Bacteriol.* **178**, 5347–5352 (1996).
42. Eisen, J.A. & Wu, M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor. Popul. Biol.* **61**, 481–487 (2002).
43. Vert, J.P. A tree kernel to analyse phylogenetic profiles. *Bioinformatics* **1**, 276–284 (2002).
44. Verjovsky Marcotte, C.J. & Marcotte, E.M. Predicting functional linkages from gene fusions with confidence. *Appl. Bioinformatics* **1**, 37–44 (2002).
45. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
46. Perna, N.T. *et al.* Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533 (2001).