From Gene Networks to Gene Function

Thomas Schlitt,1,3 Kimmo Palin,1 Johan Rung,1 Sabine Dietmann,1 Michael Lappe,1 Esko Ukkonen,2 and Alvis Brazma1

1European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; 2Department of Computer Science, FIN-00014 University of Helsinki, Finland

We propose a novel method to identify functionally related genes based on comparisons of neighborhoods in gene networks. This method does not rely on gene sequence or protein structure homologies, and it can be applied to any organism and a wide variety of experimental data sets. The character of the predicted gene relationships depends on the underlying networks; they concern biological processes rather than the molecular function. We used the method to analyze gene networks derived from genome-wide chromatin immunoprecipitation experiments, a large-scale gene deletion study, and from the genomic positions of consensus binding sites for transcription factors of the yeast Saccharomyces cerevisiae. We identified 816 functional relationships between 159 genes and show that these relationships correspond to protein–protein interactions, co-occurrence in the same protein complexes, and/or co-occurrence in abstracts of scientific articles. Our results suggest functions for seven previously uncharacterized yeast genes: KIN3 and YMR269W may be involved in biological processes related to cell growth and/or maintenance, whereas IES6, YEL008W, YEL033W, YHL029C, YMR010W, and YMR031W-A are likely to have metabolic functions.

The function of many genes is still unknown; even for the well studied yeast Saccharomyces cerevisiae, about one-third of all genes are still uncharacterized (Ball et al. 2001). Functions of uncharacterized proteins are usually inferred computationally on the basis of sequence similarities, common structural motifs, gene order, gene fusion events, or similarities in gene expression (Bork and Koonin 1998; Enright et al. 1999; Marcotte et al. 1999; Ge et al. 2001; Ponting 2001; Kemmeren et al. 2002; Valencia and Pazos 2002; Wu et al. 2002; Huynen et al. 2003). Here we introduce a simple and general statistical method for functional predictions based on scoring the similarity of gene neighborhoods in various gene networks. It allows us to utilize recently published biological data from high-throughput technologies. This method allows us to perform functional predictions for proteins independent of homologies in protein structure and sequence and provides a way to characterize proteins that have not been studied previously.

Many biological data sets can be represented as gene networks, where nodes represent genes or proteins, and the connections between the nodes represent relationships between these entities. Directed relationships such as “protein A activates gene B” are represented by arcs (A→B), whereas symmetric relationships such as “protein A and protein B bind to each other” are represented by edges (A—B; Schwickowski et al. 2000; Wallnau and Vidal 2001; Gerstein et al. 2002; Schlitt and Brazma 2002; von Mering et al. 2002).

We compared the neighborhoods of genes in networks derived from microarray experiments on gene deletion mutants (Hughes et al. 2000), the localization of transcription factor binding sites (Pilpel et al. 2001), and chromatin immunoprecipitation (ChIP) experiments for the yeast Saccharomyces cerevisiae (Ren et al. 2000; Iyer et al. 2001; Simon et al. 2001; Lee et al. 2002). By neighborhood of a gene A we mean the set of genes that are directly connected to gene A in the network. If two genes share many neighbors in a network, it suggests that these genes might be functionally related (Fig. 1).

Validation of functional relationships is problematic, because various aspects and meanings are subsumed under the term “function” of a gene or protein. This is mainly due to different experimental approaches that focus either on the effects of mutations or on biochemical activities (Ashburner et al. 2000). Unlike in protein structure prediction, there are no established standards for the evaluation of functional predictions (Blaschke et al. 2002).

We use three approaches to validate the predicted functional relationships: We compare the gene pairs that are predicted to be related (1) with protein–protein interaction data, (2) with protein complexes, and (3) with a literature network. Many biological functions involve protein–protein interactions, and several large protein–protein interaction data sets are available (Uetz et al. 2000; Ito et al. 2001; Gavin et al. 2002; Ho et al. 2002). These data sets are a valuable resource, although they may contain a large number of false positives and are far from being complete (Bader and Hogue 2002; Edwards et al. 2002; von Mering et al. 2002). For protein complexes, a manually annotated data set of high quality is available from the Munich Information Centre on Protein Sequences (MIPS); http://mips.gsf.de/; Mewes et al. 2002). Protein function is not restricted to protein–protein interactions, and for that reason we included knowledge from published scientific articles in our verification procedure. We analyzed the frequency of co-occurrences of gene names in abstracts of scientific articles on S. cerevisiae to construct a literature network. Similar approaches have been used before under the assumption that functionally related genes occur more often in the same abstract than unrelated genes do (Blaschke et al. 1999; Jensen et al. 2001).

Here we describe how the comparison of gene neighborhoods from different gene networks can be used to identify functionally related genes. We provide evidence that gene pairs with similar network neighborhoods occur more frequently together.
in article abstracts and more frequently encode proteins that interact physically than do genes with dissimilar neighborhoods. Our method allowed us to identify 816 functional relationships between 159 genes and to assign biological process annotation to seven previously uncharacterized genes. We examine some of the predictions in detail, and show that for the networks studied here the predicted functions concern biological processes rather than biochemical activities.

RESULTS

Our aim was to study the similarity of genes or proteins by assessing the similarity of their neighborhoods in gene networks (Fig. 2). Here we studied relationships between genes/proteins in six different networks of three different types for the yeast Saccharomyces cerevisiae (Table 1):

1. Mutant network: An arc from a gene A to gene B means that in a mutant where A is deleted, the expression level of B is significantly changed (Rung et al. 2002). The network is derived from microarray studies of yeast mutants by Hughes et al. (2000).

2. In silico network: An arc from gene A to B means that A is a transcription factor, and its binding site is predicted in the putative promoter of B (Palin et al. 2002). The network is derived from the data of Pilpel et al. (2001), who matched binding sites against all upstream sequences in the entire yeast genome computationally. We included only the empirically known binding sites.

3. Four different ChIP networks: These were constructed from genome-wide transcription factor localization data based on ChIP experiments (Ren et al. 2000; Iyer et al. 2001; Simon et al. 2001; Lee et al. 2002). In ChIP networks, an arc from gene A to gene B means that transcription factor A was empirically found to bind to the putative promoter region of B.

All networks listed above are represented as directed graphs. In a directed graph, a node can have incoming and outgoing arcs, and thus we can divide the neighborhood of a node depending on the orientation of the arcs. We call the genes with outgoing arcs source genes, and for every source gene s, we define the target set T_s as the set of genes which have incoming arcs from s (see Figs. 1, 2).

All of the networks described above are asymmetric: Although source genes are an a priori selected subset of the genome (particular for each network), the whole genome is tested for targets. We call such networks comprehensive target networks.

For every pair of source genes s_1 and s_2, we test whether their target sets T_1 and T_2 intersect more than expected by chance, using the hypergeometric distribution (Sokal and Rohlf 1995) and Holm’s correction (Holm 1979) for multiple testing (which leads to some P-values being greater than 1).

We performed 23,758 target-set comparisons for 15,061 source gene pairs within and between the networks (Table 2). For 816 (5.4%) source gene pairs, we found a strong target-set similarity (P < 0.01). We provide the results of our target-set comparisons for all source gene pairs within our Supplemental data (full-table-long.txt), available at www.genome.org.

When we compared target sets for the same source gene from different networks, we found that 34 out of 80 target-set pairs are highly similar. The similarities occur more frequently between the ChIP networks and between the in silico network and the ChIP networks. According to this comparison, the ChIP networks are similar to each other, and to the in silico network, whereas the mutant network is most different from the others. This is consistent with the small intersection of the mutant network and the ChIP networks: They share 16 source genes, but only 78 connections, although there are on average between 51 and 145 connections per source gene in both networks (Table 1).

To test whether the target-set similarity can be used to identify functionally related genes, we used three additional networks as reference networks:

4. Protein–protein interaction networks: Two proteins are connected by an edge if they physically interact. We integrated protein–protein interaction data from several large-scale experiments at www.genome.org.

Table 1. Number of Source Genes, Total Number of Genes, Number of Connections, and the Ratio of Connections per Source Gene in Each Comprehensive Target Network

<table>
<thead>
<tr>
<th>Network</th>
<th>in silico network</th>
<th>mutant network</th>
<th>ChIP network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source genes</td>
<td>38</td>
<td>187</td>
<td>2</td>
</tr>
<tr>
<td>Genes</td>
<td>5583</td>
<td>5555</td>
<td>130</td>
</tr>
<tr>
<td>Connections</td>
<td>23446</td>
<td>27252</td>
<td>131</td>
</tr>
<tr>
<td>Connections per</td>
<td>617.0</td>
<td>145.7</td>
<td>65.5</td>
</tr>
<tr>
<td>source gene</td>
<td>159</td>
<td>567</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>1088</td>
<td>453</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>65.5</td>
<td>4235</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>151.0</td>
<td>51.0</td>
</tr>
</tbody>
</table>

The maximal possible number of target genes in each network is the complete gene set of the yeast Saccharomyces cerevisiae (~6200 genes).
Table 2. Number of Target Set Comparisons Which Have Been Performed (Total) and the Number and Proportion of Highly Similar Target Sets (P ≤ 0.01)

<table>
<thead>
<tr>
<th>Source gene pairs</th>
<th>Total</th>
<th>P ≤ 0.01</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>s₁ = s₂, from different networks</td>
<td>46</td>
<td>17</td>
<td>36.9</td>
</tr>
<tr>
<td>s₁ ≠ s₂, from the same network</td>
<td>7838</td>
<td>741</td>
<td>9.5</td>
</tr>
<tr>
<td>s₁ ≠ s₂, from different networks</td>
<td>10405</td>
<td>143</td>
<td>1.3</td>
</tr>
<tr>
<td>All pairs s₁, s₂</td>
<td>15061</td>
<td>816</td>
<td>5.4</td>
</tr>
</tbody>
</table>

The source genes s₁ and s₂ are chosen from the same or from different networks.
protein–protein interactions (pp1, pp2, mips) and of cocitations (Fig. 3). This means that we can predict protein–protein interaction or a functional relationship based on target-set similarity.

If we base the predictions on target-set comparisons between different networks, we greatly expand the number of source gene pairs for which we perform target-set comparisons, but the false-positive rate also increases (Fig. 3B). This increase in false positives is higher for protein–protein interactions than for cocitations.

The data indicate that for the identification of protein–protein interactions, a comparison of source genes within the ChIP networks yield the best results. However, comparisons of target sets in the mutant network perform best for the identification of interactions in MIPS complexes and literature data. Generally, comparisons between different networks perform worse than comparisons within the same network (see netComparison.pdf in our Supplemental data). It should be noted that there is not enough data available for a reliable analysis of which network combinations yield the best predictions.

The correlation between target-set similarity and functional similarity is evident in the graph representation of the predictions (Fig. 4, fig4.txt in Supplemental data). Genes involved in the same biological processes such as pheromone response or cell-cycle control are linked by several target-set similarities, and are therefore close to each other in the graph. Applying a guilt-by-association approach, we used proximity in the graph to infer gene function (Oliver 2000): We predict function (using a P-value threshold of 10^-12) for four genes (KN3, YEL008W, YEL033W, and YHL029C) which are currently not assigned to a biological process (Table 4). YEL033W is connected to only one other gene, BUD21, which is involved in RNA processing; KN3 shows strong target-set similarity to GAS1, a 1,3-beta-glucanosyltransferase involved in cell wall organization and biogenesis, and to BUD14, which is involved in bud site selection according to SGD (http://www.yeastgenome.org/). This would imply that Kin3p may be involved in cell growth, budding, or related processes.

It is difficult to find terms describing a set of genes appropriately and objectively; therefore we use the “SGD Gene Ontology Term Mapper” (http://db.yeastgenome.org/cgi-bin/SGD/GO/goTermMapper). SGD uses the Gene Ontology (GO) terms from the Gene Ontology Consortium to annotate yeast genes (Dwight et al. 2002; Blake and Harris 2003). GO terms are organized hierarchically, which allows an investigator to find higher-level terms starting from a more specific GO term. GO Term Mapper does exactly this; starting from the GO annotation for each gene in the group, it identifies GO terms which are shared by the whole group (or at least the majority of the genes). When we query GO Term Mapper for the genes that have similar target sets as KN3, it returns “cell growth and/or maintenance” as a common annotation, which is consistent with the conclusion we reached above.

The same approach applied to the target sets of YEL008W and YHL029C suggests that these genes are involved in metabolism (Table 4). Four additional uncharacterized genes show high target-set similarities (at a P-value threshold of 0.01) to several other source genes. Using GO Term Mapper as described above, IES6, YMR010W, and YMR031W-A can be mapped to “metabolism,” whereas YMR269W can be mapped to “cell growth and/or maintenance” (see predictions.pdf in Supplemental data).

Lastly, we examined some source gene pairs with high target-set similarity in detail to illustrate the nature of our predictions: There are 14 source gene pairs for which both genes are present in the protein interaction network pp2, but no interaction between them is reported in this network, although they are connected in cocitation network m3. Of these 14 pairs, six have highly similar target sets (P > 0.01). The pair with the lowest P-value (linked by 11 abstracts in the cocitation network) is MBP1-SWI4. Both genes encode related transcription factors, and each of them can form a complex with Swi6p: MBF (Swi6p-Mbp1p) and SBF (Swi6p-Sw4p; Koch et al. 1993). The second pair consists of the homologous transcription factors Ace2p and Swi4p (linked by three articles in cocitation network m3; Stillman et al. 1994; Measday et al. 2000). The next two pairs are between genes involved in phenome signaling: Ste12p-Ste4p (four abstracts) and Ste4p-Ste7p (four abstracts). The phenome signal in yeast is transmitted from the receptor via a G-protein-complex (Ste4p and others) and a MAP kinase cascade (Ste7p and others) to the transcription factor Ste12p (Sprague and Thorer 1992). The two proteins of the fifth pair—repressor Sum1p and activator Ndt80p (three abstracts)—compete for the transcriptional control of genes containing a middle sporulation element (MSE) in their promoters (Xie et al. 1999; Lindgren et al. 2000). Mig1p (three abstracts) was shown to repress the expression of the SUC2 gene synergistically with the Ssn6p-Tup1p repressor complex (Alepu et al. 1997). Thus, all six source gene pairs with high target-set similarity are in fact functionally related, but do not show protein–protein interactions.

### DISCUSSION

We conclude that the comparison of target sets in gene networks can be used to find functionally related proteins: We predict 816
The nature of the predicted functional relationships is dependent on the nature of the comprehensive target networks. The Gene Ontology consortium differentiates between three major subcategories “cellular localization,” “biological process,” and “molecular function” (Ashburner et al. 2000). For the networks studied here, the predicted functions concern biological processes rather than molecular functions. In this respect our method is similar to other nonhomology methods (Marcotte 2000). We demonstrated that the products of genes with similar network neighborhoods often interact physically, are likely to be part of the same protein complex, and/or are often reported together in the literature. These results are in agreement with the recent finding that protein–protein interactions correlate with protein–DNA interactions (Manke et al. 2003). The proposed method can be used to identify functionally related candidate genes using a guilt-by-association approach.

With the proposed method we did not identify all functional relationships reported in the reference networks. It therefore remains an open question as to how many of the errors are due to limitations of the available data or due to the method. There are several reasons why not all of the target-set pairs derived from the same source gene, or from two genes having a known functional relationship, were highly similar. One reason is that we combined experimental data from different types of experiments, and certain interactions are only observable under very specific conditions not necessarily attained in a given experiment. For example, some transcription factors may bind DNA only if they are phosphorylated.

One advantage of this method is that we can use and integrate a wide variety of different experimental data sets, as long as

Figure 3  ROC plots of true-positive rate (sensitivity) vs. false-positive rate (1 − specificity) for the prediction of protein–protein interaction (ppi1, ppi2), protein complexes (mips), and “co-citation” (mi2, mi3). The source genes $s_1$, $s_2$ are chosen from same (A) or different networks (B). (C) An ROC plot using the union of ppi2, mips, and mi3 as verification network, with source genes $s_1$, $s_2$ chosen from the same network (all-same) or different networks (all-diff).
they can be represented as comprehensive target networks. Even small data sets can be successfully included; unlike clustering of microarray data, there is no need for extensive experiments consisting of tens of microarray hybridizations to provide biologically meaningful results. Our method is versatile; in the present study, for instance, we were able to explore which transcription factor deletions lead to predicted effects on the basis of the localization of its binding sites. We can also look for transcription factors which act in combination with other factors and elucidate possible upstream regulatory mechanisms.

Although sequence information may be important for the design of the experiments which underlie the comprehensive target networks, this is not a prerequisite for our method, which is completely independent of sequence or structural homology. A limitation of this method is that the data sets used for our predictions must be represented as comprehensive target sets. This means that, for example, large-scale protein-interaction networks cannot be used, because of the way these experiments are performed. Only positive interactions are reported, and we do not know which protein interactions do not occur. In contrast, the data sets we included for the predictions always report a signal for all genes in the genome. Therefore, within the limitations of the experimental methods, we always have information regarding the individual behavior of all genes.

The possibility of integrating data derived from different experimental methods and conditions allows the exploration of the complexity of cellular regulatory mechanisms. It is feasible to perform repeated analysis of data from different experimental...
conditions and then use the variations in conditions to explain the changes in interactions predicted. This would lead to a dynamic rather than a static view of protein function.

**METHODS**

**Construction of the Networks**

The mutant network was constructed with data from Hughes et al. (2000). Target sets T contain genes whose expression level changed significantly; that is, the ratio of gene expression in the mutant divided by the background standard deviation in the wild-type strain has an absolute value larger than 2.5 (Runge et al. 2002).

The in silico network was compiled from data reported by Pilpel et al. (2001) on the occurrence of known binding sites in putative promoter regions of yeast genes.

The four ChIP networks were constructed from data published by Ren et al. (2000), Simon et al. (2001), Iyer et al. (2001), and Lee et al. (2002) derived for two, nine, three, and 106 transcription factors, respectively.

Experimental data on yeast protein–protein interactions was retrieved from the following databases and publicly available data sets: DIP (Xenarios et al. 2001), MINT (Zanzoni et al. 2002), MDS (Ho et al. 2002), and cellzome (Gavin et al. 2002). Although DIP and MINT contain binary interactions, the data from the Gavin and Ho studies contain sets of proteins from a number of immunoprecipitations. These were broken down into a complete set of binary interactions.

The MIPS network was derived from manually annotated complexes at MIPS (Mewes et al. 2002) and provided to us by Christian von Mering (von Mering et al. 2002).

The cocitation network: Using a synonym dictionary for gene/protein names in yeast, we scanned over 70,000 journal abstracts from Medline for co-occurrences of genes/proteins, using the SRS server (http://srs.ebi.ac.uk). We compiled a synonym dictionary containing the complete set of unique yeast ORF identifiers, the corresponding gene names, and their synonyms from

**Table 4. Part of the Data Shown in Figure 4**

<table>
<thead>
<tr>
<th>Source gene 1</th>
<th>Network 1</th>
<th>Source gene 2</th>
<th>Network 2</th>
<th>( P ) value</th>
<th>ppi</th>
<th>mi</th>
<th>mips</th>
<th>goTermMapper</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEL033W</td>
<td>mutant</td>
<td>BUD21</td>
<td>mutant</td>
<td>2.18E-14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIN3</td>
<td>mutant</td>
<td>GAS1</td>
<td>mutant</td>
<td>9.16E-26</td>
<td>cell growth/maintenance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIN3</td>
<td>mutant</td>
<td>BUD14</td>
<td>mutant</td>
<td>1.12E-20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEL0008W</td>
<td>mutant</td>
<td>CLB2</td>
<td>mutant</td>
<td>1.41E-20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEL008W</td>
<td>mutant</td>
<td>ERG2</td>
<td>mutant</td>
<td>2.08E-18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEL008W</td>
<td>mutant</td>
<td>ERG28</td>
<td>mutant</td>
<td>3.64E-25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEL008W</td>
<td>mutant</td>
<td>SW4</td>
<td>mutant</td>
<td>1.12E-21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEL008W</td>
<td>mutant</td>
<td>GCN4</td>
<td>mutant</td>
<td>1.61E-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>ANP1</td>
<td>mutant</td>
<td>1.28E-26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>CKB2</td>
<td>mutant</td>
<td>9.49E-18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>CLB2</td>
<td>mutant</td>
<td>7.38E-43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>ERG2</td>
<td>mutant</td>
<td>1.57E-34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>ERG28</td>
<td>mutant</td>
<td>1.18E-50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>ERG3</td>
<td>mutant</td>
<td>1.12E-20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>GCN4</td>
<td>mutant</td>
<td>3.44E-30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>RTC1</td>
<td>mutant</td>
<td>3.44E-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>SP1</td>
<td>mutant</td>
<td>1.52E-13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>SW4</td>
<td>mutant</td>
<td>2.56E-64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YHLO29C</td>
<td>mutant</td>
<td>UBR1</td>
<td>mutant</td>
<td>8.29E-25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The target set similarity for the particular source gene/network combination is indicated by the \( P \) value. yes: the particular interaction is confirmed by a reference network; no: both genes are present in the respective reference network, but no interaction is reported; ppi2, protein-protein interaction network, mi3, co-citation network, mips, mips network; the last column contains the GO term describing the set of genes according to GO Term Mapper (http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermMapper).
publicly available information in the following databases: SGD (http://www.yeastgenome.org/), MIPS (http://mips.gsf.de/yeast/), and EBI Proteome Analysis Database (http://www.ebi.ac.uk/proteome/). Each Medline entry was required to contain at least one ORF/gene name or one of its associated synonyms in the text body of the abstract or in the title; in addition, we required the MESH term ‘Saccharomyces cerevisiae’ to limit the search to our chosen model organism. A co-occurrence between two different gene/protein identifiers was counted if they or any of their respective synonyms were found in the same abstract. This resulted in 41,129 associations, among which about 10285 pairs were co-occurring at least twice for 3616 genes. All networks are available from our Web supplement.

Network Comparison
Assessing the similarity between target sets using the hypergeometric distribution: The null hypothesis for testing the similarity of target sets T1 and T2 is that the genes in the sets are picked from the genome independently, randomly with equal probabilities. Under this null hypothesis, the number of genes in the intersection of T1 and T2 is distributed according to the hypergeometric distribution with the size of the genome, the size of T1, and the size of T2 as parameters (Palin et al. 2002). With this distribution we can compare the probabilities of observing an intersection at least this large, given that the null hypothesis is true. The pairwise P-values need to be corrected, because we evaluate multiple hypothesis tests. For the adjustment of the P-values, we used the sequential Holm’s correction (Holm 1979). In ties. 

Acknowledgments
We thank Michael Ashburner, Aria Baniahmad, Cath Brooks-bank, Frank Holstege, Patrick Kemmeren, Helen Parkinson, and Steve Russell for helpful discussion of the manuscript and Christian von Mering for the MIPS reference set. The project is funded by the European Commission as the TEMBLOR, contract no. QLRI-CT-2001-00015 under the RTD programme “Quality of Life and Management of Living Resources.” K.P. and E.U. are supported by a grant from the Academy of Finland.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

References