BMC Bioinformatics

BioMed Central

This Provisional PDF corresponds to the article as it appeared upon acceptance. The fully-formatted PDF version will become available in approximately two weeks, from the URL listed below.

Automated Modelling of Signal Transduction Networks

BMC Bioinformatics 2002, 3:34

Dr Martin A Steffen (steffen@rascal.med.harvard.edu)
Allegra A Petti (petti@fas.harvard.edu)
Prof John Aach (aach@genetics.med.harvard.edu)
Dr Patrik D'haeseleer (patrik@genetics.med.harvard.edu)
Prof George M Church (church@arep.med.harvard.edu)

ISSN 1471-2105

Article type Research article

Submission date 26 Aug 2002

Acceptance date 01 Nov 2002

Publication date 01 Nov 2002

Article URL http://www.biomedcentral.com/1471-2105/3/34

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

http://www.biomedcentral.com/info/publishing_adv.asp

© 2002 Steffen *et al*, licensee BioMed Central Ltd. BioMed Central Ltd. This article is published in Open Access: verbatim copying and redistribution of this article are permitted in any medium for any purpose, provided this notice is preserved along with the article's original URL.

Automated Modelling of Signal Transduction Networks

Martin Steffen[1], Allegra Petti[1], John Aach[1,2], Patrik D'haeseleer[1,2], George Church[1,2]*

[1]Dept. of Genetics and [2]Lipper Center for Computational Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts, 02115, USA. *To whom correspondence should be addressed. E-mail: church@arep.med.harvard.edu

Abstract

Background

Intracellular signal transduction is achieved by networks of proteins and small molecules that transmit information from the cell surface to the nucleus, where they ultimately effect transcriptional changes. Understanding the mechanisms cells use to accomplish this important process requires a detailed molecular description of the networks involved.

Results

We have developed a computational approach for generating static models of signal transduction networks which utilizes protein-interaction maps generated from large-scale two-hybrid screens and expression profiles from DNA microarrays. Networks are determined entirely by integrating protein-protein interaction data with microarray expression data, without prior knowledge of any pathway intermediates. In effect, this is equivalent to extracting subnetworks of the protein interaction dataset whose members have the most correlated expression profiles.

Conclusion

We show that our technique accurately reconstructs MAP Kinase signaling networks in *Saccharomyces cerevisiae*. This approach should enhance our ability to model signaling networks and to discover new components of known networks. More generally, it provides a method for synthesizing molecular data, either individual transcript abundance measurements or pairwise protein interactions, into higher level structures, such as pathways and networks.

Background

Signal transduction is the primary means by which cells coordinate their metabolic, morphologic, and genetic responses to environmental cues such as growth factors, hormones, nutrients, osmolarity, and other chemical and tactile stimuli. Traditionally, the discovery of molecular components of signaling networks in yeast and mammals has relied upon the use of gene knockouts and epistasis analysis. Although these methods have been highly effective in generating detailed descriptions of specific linear signaling pathways, our knowledge of complex signaling networks and their interactions remains incomplete. New computational methods that capture molecular details from high-throughput genomic data in an automated fashion are desirable and can help direct the established techniques of molecular biology and genetics.

DNA microarray technology has evolved to the point where one can simultaneously measure the transcript abundance of thousands of genes under hundreds of conditions, producing hundreds of thousands of individual data points. Similarly, high-throughput yeast two-hybrid experiments have identified thousands of pairwise protein-protein interactions. Once a core pathway is established, these data can readily be integrated into model refinements, as a recent study in systems biology elegantly demonstrates [1]. However, synthesizing these data *de novo* into models of pathways and networks remains a significant challenge.

How can one bridge the gap from transcript abundances and protein-protein interaction data to pathway models? Clustering expression data into groups of genes that share profiles is a proven method for grouping functionally related genes, but does not order pathway components according to physical or regulatory relationships. Here we present an automated approach for modelling signal transduction networks in *S. cerevisiae* by integrating protein-protein interaction [2-4] and gene expression data. Our program, NetSearch, draws all possible linear paths of a specified length through the interaction map starting at any membrane protein and ending on any DNA-binding protein. Microarray expression data [5-7] is then used to rank all paths according to the degree of similarity in the expression profiles of pathway members. Linear pathways that have common starting points and endpoints and the highest ranks are then combined into the final model of the branched networks.

Our approach is calibrated using the yeast MAPK (mitogen-activated protein kinases) pathways involved in pheromone response, filamentous growth, and maintenance of cell wall integrity (Fig.1). These pathways are activated by G protein-coupled receptors and characterized by a core cascade of MAP kinases that activate each other through sequential binding and phosphorylation reactions; they are among the most thoroughly studied networks in yeast and are therefore excellent benchmarks against which to test our approach.

Results

Input data and parameters

Recent papers [2-4] have used the yeast-two-hybrid technique and literature surveys to identify and assemble over 7000 non-redundant protein-protein interactions among more than 4000 proteins. While two-hybrid screens efficiently identify fusion proteins that are able to interact, the biological significance of the interaction for native proteins acting *in vivo* generally requires verification, because the technique is susceptible to a high rate of false positives [8]. To assess the possible contribution of false-positive protein-protein interactions to the combined interaction dataset, we analyzed the connectivity of each protein and found that a small fraction of proteins had a very high number of interactions (highlighted in red, Fig. 2). With these highly connected proteins included in our data set, NetSearch generates 17 million candidate signaling pathways of length seven or less, 95% of which involve one of these twenty-two highly-connected proteins. We excluded the highly interacting proteins from the interaction dataset based on their nonspecific inclusion in the predicted pathways and evidence of their susceptibility to systematic error. This yielded an interaction map that contains 5560 interactions among 3725 proteins, an average of three interactions per protein.

Using the NetSearch algorithm, this protein interaction network was queried for paths up to length eight that begin at membrane proteins and end at transcription factors. The search generated approximately 4.4 million candidate pathways of length eight or less whose biological plausibility was assessed using gene expression data.

To score the pathways, we first used a *k*-means algorithm to cluster all yeast genes into clusters based on their expression profiles. NetSearch then assigned each pathway a statistical score [10] according to the number of pathway members that clustered together. For example, a path with six members in one cluster would score higher than a path that only had five members in that cluster. Cluster size influenced path scoring such that a path that had three members from a cluster of 30 elements would score higher than a path that had three members from a cluster of 100 elements. Also, a path with four elements in one cluster and three elements in a second cluster would score higher than a path that had four elements in cluster one, but no more than two elements in cluster two.

Pathways were scored using NetSearch's 'sumprob' scoring metric: Assuming N proteins total and a partitioning of proteins into k clusters C_1, C_2,…C_k, with N_1, N_2,…N_k members, respectively, and a pathway p of L proteins $p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_L$, where $c_p(i)$ = number of proteins in p in cluster C_i, the sumprob score is computed as follows:

$$\mathrm{prob}_p(i) = -\log_{10}\left(\sum_{h=c_p(i)}^{L} \frac{\binom{N_i}{h}\binom{N-N_i}{L-h}}{\binom{N}{L}} \right)$$

$$\mathrm{sumprob}(p) = \sum_{\substack{i=1 \\ c_p(i)>=2}}^{k} \mathrm{prob}_p(i)$$

$\mathrm{prob}_p(i)$ scores a pathway for a cluster C_i such that pathways which are more concentrated in C_i have higher scores. The summation in $\mathrm{prob}_p(i)$ computes the cumulative hypergeometric probability of pathway p containing $c_p(i)$ or more members of C_i. $\mathrm{prob}_p(i)$ assesses co-clustering of pathway members in the single cluster C_i. sumprob(p), the sum of $\mathrm{prob}_p(i)$ values over all clusters for which $c_p(i)>=2$, is a simple measure of co-clustering across the entire collection of available clusters. The rationale for the restriction $c_p(i)>=2$ is that without it a pathway could get a high score simply from having single members in one or more rare clusters, in which case the score would no longer reflect *co*-clustering.

The exact composition of paths discovered using NetSearch depend on the parameters used in path drawing and path scoring. To ensure that NetSearch reproducibly generates statistically significant, biologically plausible paths, we combinatorially varied every parameter value in the path-drawing and path-scoring algorithms, and selected parameter combinations that generate the most statistically significant pathways. Statistical significance was measured by drawing pathways from membrane proteins to DNA-binding proteins through the experimentally determined protein-interaction map (henceforth called "real pathways") and comparing these pathways with pathways drawn through control interaction maps that were created by randomizing all pairwise interactions in the original dataset. (The randomization procedure was performed three times and statistics were calculated on the average output of these runs. Paths produced using these interaction maps are henceforth referred to as "random pathways"). We ultimately chose parameters that maximized the number of high-scoring pathways produced with real interactions, while minimizing high-scoring pathways from the randomized interactions.

The parameters we varied included the number of clusters into which the genes were grouped, the microarray expression datasets used in clustering, the maximum path length, and the scoring metric. Expression data were clustered into 12, 25, 50, 100 and 250 clusters, and NetSearch best discriminated between real pathways and random pathways when genes were grouped into 25 clusters. Three *S. cerevisiae* expression datasets were examined individually, including the

"Compendium" set, composed of expression profiles in response to 300 diverse mutations and chemical treatments [5]; the "MAPK" set, composed of 56 conditions chosen to probe the behavior of MAPK signal transduction [6]; and the "Cell Cycle" set, composed of 77 conditions relevant to the cell cycle [7]. Combinations of these datasets were also examined, for a total of five different sets that allowed us to compare the utility of data that probes specific biological processes, such as MAPK signaling or the cell cycle, and that which probes the state of the cell more broadly, such as the Compendium set and the combined sets. The composite data set that combined all three individual sets (for a total of 433 conditions) provided the best discrimination between real pathways and random pathways, although the other sets performed comparably.

The final input parameter that required evaluation was the maximum path length allowable for NetSearch paths. While short path lengths risk omission of key path members, longer path lengths increase the likelihood of including false-positive interactions. As a first step towards determining the optimal maximum path length, we examined the path lengths connecting every possible pair of the 3725 proteins in the interaction dataset, regardless of subcellular localization. The minimal path length between any two proteins chosen at random contains on average 7.4 members. Secondly, we examined the fraction of pathways with high coclustering ratios for various path lengths. Consistent with our finding that the average path length between any two proteins is 7.4, this fraction peaks at eight, which we set as our maximum, unless otherwise noted.

NetSearch output

Using a maximum path length of eight, and 25 gene clusters from 433 conditions [5-7], NetSearch generated ~4.4 million pathways each for the real and randomized protein interaction datasets. From the experimental ("real") data, 4059 pathways had a coclustering score $\geq$ 16 (Fig. 3). At this cutoff, randomized interaction data produced on average only ~1% this number of pathways (32 pathways, $P = 7 \times 10^{-6}$). However, we emphasize that NetSearch selects paths based on their rank relative to all paths between selected starting and endpoints. The absolute score depends on the particular expression data set used, and varies from network to network depending on the degree of coregulation in the cell under the conditions tested in the expression data.

The signaling network models generated by NetSearch for the pheromone response, cell wall integrity and filamentation pathways are depicted in Fig. 4. In each case, the starting protein (receptor, depicted in blue) and ending protein (transcription factor, depicted in red) were selected as inputs, and NetSearch draws all possible paths between these points. The size of each vertex is proportional to the sum of scores of the paths in which that protein is found, providing a useful visual clue to the potential importance of a protein in the given network. Comparison with

Fig.1 shows that NetSearch reproduced many of the essential elements of these MAPK pathways, while providing a detailed account of the experimentally determined interconnections among network elements. Of the three network models, the one generated for the pheromone response pathway originating at Ste3p (Fig. 4A) exhibited the highest co-clustering scores. Every protein NetSearch included in this network model has a description in the Yeast Proteome Database (YPD) [9] consistent with a known or plausible role in mating. Of the nineteen proteins we have included in our depiction of the pheromone response network, eighteen are annotated as playing a role in the fungal cell differentiation by MIPS [10]. The probability that this selection would have occurred by chance was calculated with the hypergeometric distribution was found to be $P = 5 \times 10^{-24}$. Our model does differ in several respects from the canonical pheromone response pathway depicted in Fig 1. It includes more members of the heterotrimeric G protein complex, including the alpha, beta, and gamma subunits, the GDP-GTP exchange factor, and the GTPase-activating protein (Gpa1p, Ste4p, Ste18p, Cdc24p, and Sst2p, respectively). It includes Far1p, a protein necessary for pheromone-induced cell cycle arrest in G1 [11], Mpt5p, a protein necessary for recovery from cell cycle arrest [12], and Bem1p and Sph1, both of which are necessary for establishment of cell polarity during shmooing and budding [13,14]. In our protein-interaction map there is no direct interaction between a pheromone receptor (Ste2p or Ste3p) and any component of the heterotrimeric G protein complex (Ste4p/Ste18p/Gpa1p), so NetSearch drew indirect paths through Akr1p, a known inhibitor of signaling in the pheromone pathway [15]. The predicted network does not include the GTPase Cdc42p (paths were instead drawn preferentially through its cofactor Cdc24p, which physically interacts with Ste4p) or Ste20p, because of missing interactions in the protein-interaction map.

Fig. 4B depicts the pheromone response network at several different score cutoffs, and demonstrates how higher co-clustering score cutoffs reduces the complexity of the protein-interaction map. NetSearch detects 354 paths of length eight from Ste3p to Ste12p, and incorporates 70 different proteins into those paths. The top graph in Fig. 4B shows the network constructed from all 354 paths (with each protein arranged on the perimeter of an ellipse for clarity). In the middle graph, all paths that scored below the median have been eliminated, leaving only 27 proteins. On the bottom of Fig. 4B, only the highest scoring paths (those used to construct the network in Fig. 4A) with 19 proteins, are depicted. Comparison of these networks indicates that most proteins are eliminated by simply excluding the pathways that score in the bottom half; further modifications to the cutoff affect the results incrementally. In setting a precise cutoff for pathway inclusion in the final network models, one seeks to strike a balance between the inclusion of false-positives and the omission of true-positives. We set the cutoff such that the top fifteen paths for each network were included.

The network model generated for the cell wall integrity pathway is depicted in Fig. 4C. Membrane proteins in particular may fail to produce interactions when forced into the nucleus by the requirements of the standard two-hybrid technique. We observed this to be the case for the cell wall integrity pathway, as neither Wsc1p, Wsc2p, Wcs3p or Mid2p were observed to interact in any of the high-throughput screens. To reconstruct this network, we therefore started with the momomeric GTPase Rho1p, and restricted our search to a length of seven because of the omission of the initial signal sensor. Of the 18 proteins included in this network model, all but Smd3p have descriptions consistent with a role in cell wall maintenance. NetSearch included both GTPase constituents of this pathway, Rho1p and Cdc42p, as well as associated GAPs and other interactors, including Rdi1p, Rga1p, and Gic2p. Other included network elements are Fks1p, the 1,3-βglucan synthase of which Rho1p is a subunit [16], the actin protein Act1p, and the proteins Bni1p, Bud6p, and Sph1p, which are associated with Rho-mediated signal transduction, actin filament organization, cell polarity establishment, and bud growth. Smd3p forms a complex with the Sm core spliceosomal proteins [17], and we are not aware of any role it may play in maintaining cell wall integrity. Its inclusion is most likely a result of its expression correlation with *BUD6* in one of the microarray datasets, but it seems unlikely that the observed interactions of Smd3p with Spa2p and Slt2p have biological significance. In the NetSearch-generated model, Bck1p is downstream of Mkk1p because, although it interacts with both Mkk1p and Mkk2p, it has been shown specifically not to interact with Pkc1p in two hybrid assays [18].

The network model for filamentous growth (Fig. 4D) involves 21 proteins, 20 of which are known to play a role in filamentous growth, or have functions consistent with that role, with the exception of Fus1p. As in the pheromone response and cell wall integrity network models, key components of the Ras GTPase are included, such as Cdc25p (the Ras guanine nucleotide exchange factor), Cyr1p (the Ras-associated adenylate cyclase), and Srv2p, which enables the activation of adenylate cyclase by Ras2p. Several proteins with roles in actin filament organization, cell polarity establishment, bud growth, and GTPase-mediated signal transduction are shared with the cell wall integrity pathway, including Bni1p, Spa2p, Bud6p, and Act1p. NetSearch depicts interactions between Abp1p and both Srv2p and Act1p, consistent with the function of Abp1 in tethering Srv2p to the cytoskeleton. The adenylate cyclase and associated proteins mentioned above, along with Hsp82p and Hsc82p, activate the cAMP pathway [19], a pathway that acts in parallel with the MAPK pathway to promote filamentation. Hsp82p is a chaperone protein known to interact with a number of signaling pathway components [20]. It is required for activation of the pheromone signaling pathway [21], and for the general response to amino acid starvation [22]. It may play a similar role in response to nitrogen (ammonia) starvation, a trigger for filamentation. Fus1p, included in our predicted network, does not have a documented role in filamentation; it is required for cell fusion during pheromone initiated mating. Its transcript levels are significantly

upregulated in response to pheromone, but are unchanged in *tec1*Δ strains [6]; that study notes, however, that in *dig1*Δ*dig2*Δ cells, fus1 is constitutively activated, and both mating and invasive growth are observed. Tec1p, conspicuously absent in our model, has not been observed to interact with any proteins in high-throughput two-hybrid screens.

Discussion

The utility of yeast protein-protein interaction maps for generating signaling network models has previously been suggested [23], and they have been used to predict metabolic pathways [24]. Expression data has been used to generate and refine models for genetic regulatory networks without the benefit of protein-protein interaction data [25]. In this study, we have used expression data to rank candidate pathways of interacting proteins. This approach has a strong biological and experimental rationale: proteins used in the same signaling network must exist simultaneously with its activation. The genes encoding these proteins must be transcribed at approximately the same time, and under the same environmental conditions in which the signaling network is required. Furthermore, experimental evidence suggests that when a signaling network is activated, positive feedback mechanisms upregulate the expression of genes that encode pathway proteins [26], implying that this rationale is also applicable to "surveillance" pathways, whose protein components may need to be constitutively present in small quantities, but whose concentration increases with activation. This biological rationale is borne out by evidence that interacting proteins have more highly correlated expression profiles than do non-interacting proteins [27]. However, if a single component of a signaling network is independently (and differentially) regulated, it would not necessarily be excluded using our approach, if for instance, it connected two halves of a pathway which had similar average expression profiles.

NetSearch can be used to predict new signaling pathways, identify previously unknown members of documented pathways, or identify smaller clusters of interacting proteins. Until we have a more complete protein-interaction set, a user who wishes to explore a particular pathway (http://arep.med.harvard.edu/NetSearch) needs to specify pathway starting points and ending points (such as membrane and DNA-binding proteins, respectively). This selection can be based on a known genetic interaction, a shared mutant phenotype, a shared functional classification, or signature expression profile. This is the approach we have followed in constructing the networks depicted in Fig. 4. Those networks are comprised of all highest ranking linear paths connecting the receptors and transcription factors for that pathway.

The pheromone response pathway is commonly depicted as a simple, linear transmission of the mating signal from the membrane receptor, Ste2p (for alpha-factor) or Ste3p (for a-factor), to the nuclear effectors, Ste12p and Mcm1p, via a

MAPK cascade. However, mating pheromone exposure also induces other cellular processes such as those required for polarized growth, cell cycle arrest, and recovery from cell cycle arrest. Furthermore, the topology of the protein interactions required for these processes is considerably more complicated than a series of pairwise interactions. In addition to accurately depicting the MAPK cascade, our predicted pheromone response network identifies many proteins necessary to execute the coordinated processes of growth polarization and cell cycle arrest, and reflects the complex topology of the interaction network.

The complexity of these interactions are observed in large, multifunctional complexes of possibly dynamic composition. For example, products of Ste18, Ste4, Cdc42, Cdc24, Far1, Bem1, Ste20, Ste5, and other proteins are thought to constitute a complex that has numerous interactions among components, and that mediates many different cellular processes [14,28]. The complex may coordinate mating pheromone detection with (1) cell cycle arrest via Far1p, (2) MAPK signal transduction via Ste5p and Ste20p, and (3) cell polarity via Bem1p and Far1p (among others) [29,30].

Given that several of these networks share components of the MAPK cascade, the mechanism by which input-output specificity is maintained remains one of the most important questions in the field of molecular signal transduction. One well accepted hypothesis is that scaffolding proteins such as Ste5p and Pbs2p tether the MAPK module to the appropriate input and output components [31]. The recent identification of numerous Ste5p analogs in yeast and mammals makes this hypothesis even more intriguing [26]. Beyond scaffolding proteins, higher-order protein complexes have been hypothesized to play a role in maintaining signal specificity [32]. Our computational results suggest that this may indeed be the case. When comparing the minimal pathways for pheromone response and filamentation as depicted in Fig. 1, it appears that maintaining signal specificity would be a considerable challenge. But when comparing the two network predictions depicted in Fig. 4, one notes many differences, all of which may help ensure specificity. The network perspective suggests not a single scaffolding protein, but many scaffolding proteins - in fact, a “scaffolding network.” The possibility exists that relatively nonspecific kinases function simply as “phosphorylation modules,” operating inside insulating networks that are the primary determinant of signaling specificity.

Because our protein-protein interaction data is only a small fraction of a truly complete interaction map, one finds portions of a network that cannot be connected using available protein-protein interaction data. This was the case in our attempts to model the HOG network. While NetSearch correctly identified the upstream elements of this pathway (Sln1p $\rightarrow$ Ypd1p $\rightarrow$ Ssk1p $\rightarrow$ Ssk22p), it was unable to form any connections to Pbs2p or Hog1p that ended in a transcription factor. In some cases, a missing interaction can be circumvented, however. In the model for the pheromone response network, NetSearch inserted Akr1p, a known inhibitor of the pheromone pathway [15], between Ste3p and the G protein complex (Ste4p/Ste18p/Gpa1p).

Although the protein-interaction dataset we used contained no direct interaction between Ste2p/Ste3p and Ste4p, Ste2p-Ste4p has been shown to interact in a targeted yeast two-hybrid study [33].

Our failure to model the HOG pathway underscores the fact that, for the purposes of this algorithm, missing interactions (false-negatives) are a more significant obstacle than are false-positive interactions. Missing interactions cannot be “created” by the algorithm, but false-positive interactions are de-emphasized as a result of the bias imposed by ranking paths according to the similarity of expression profiles. Bearing this out, of the fifty-eight proteins included in our networks, only Smd3p seems to be included as a result of false-positive interactions. (This is distinct from the case of Fus1p, which may be misplaced in the filamentation pathway, but whose interactions with Act1p and Ste7p are real.) This highlights a general observation on the integration of genomic technologies. Two-hybrid and microarray expression studies are both known to have a sizable fraction of systematic errors (for instance, self-activators in two-hybrid experiments, and cross-hybridization in microarrays), but when looking at the intersection of the two, the true signals tend to reinforce one another, whereas the systematic errors in the two tend to be different and are reduced further into the noise. These effects may help explain why we observe so few false-positive proteins inserted into our predicted networks.

In addition to using more complete interaction datasets, such as those found in Ho [35] and Gavin [36], one could improve this approach by integrating more types of data. Homology modelling could be used to differentially weight the inclusion of molecules likely to be involved in signal transduction (e.g. kinases), and genetic interactions could weight the inclusion of the two proteins in the same path. Signaling motif identification [36] and data from protein kinase chips [37] could also easily be incorporated into this framework. Based on the interaction data available, the networks depicted in Fig. 4 are static, with all interactions given equal weight, and without information on the direction of information transfer. In reality, signaling networks are dynamic and vectorial complexes, with interactions of varying strengths among component proteins [38]. The technology necessary to generate data which will allow modelling of these network properties are beginning to emerge. Kinase chips [37] will allow one to incorporate information about the direction of information flow. The strength of protein interactions (with DNA) has been measured on chips in a highly parallel manner [39] and the same could be done for protein-protein interactions [40]. Data on the spatial and temporal co-localization of signaling components is being generated by new imaging techniques [41], which will yield insight into the mechanism with which the cellular response to a signal is modulated by the intensity and the duration of the signal [42], and the interplay with parallel pathways.

Conclusions

The approach we have presented allows one to query the intersection of two enormous sets of functional genomic-derived molecular data. One can, in effect, simultaneously browse protein-protein interaction and gene expression data. It allows one to extract a group of highly-connected, highly-correlated proteins from global data to isolate a sub-network of particular interest. Significantly, this approach does not require prior knowledge of pathway intermediates. The interaction data determines the pathways that are considered, and gene expression data is used to rank the pathways. Although we have focused on signaling pathways, this approach should be applicable to modelling the relationships among any group of interacting proteins that cooperate to perform a given function within a cell, and the web-version of the software allows for these queries. As many genomic techniques are generating increasingly large amounts of molecular data, new tools such as this will be required for the synthesis of "parts into pathways" in order that we may understand how cells regulate the many processes necessary for growth and development.

Authors' contributions

M.S. conceived of the study, performed the network modelling and drafted the manuscript. A.P. wrote program code, analyzed the network models and drafted the manuscript. J.A. devised algorithms, wrote and refined program code and constructed the associated web pages. P.D. performed statistical analyses and examined the protein interaction maps. G.C. guided the study and coordinated the project. All authors read and approved the final manuscript.

Acknowledgements

We thank Lisa Pacella, Aimee Dudley and Vasudeo Badarinarayana for excellent advice and assistance and all members of the Church and Winston labs for helpful discussion. We also thank the Lipper Foundation, ONR, NSF and DOE grant DE-FG02-87ER60565.

References

1. T Ideker, V Thorsson, JA Ranish, R Christmas, J Buhler, JK Eng, R Bumgarner, DR Goodlett, R Aebersold, L Hood: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001 292: 929-934
2. B Schwikowski, P Uetz, S Fields: **A network of protein-protein interactions in yeast.***Nat Biotechnol.* 2000**18** : 1257 1261
3. P Uetz, L Giot, G Cagney, TA Mansfield, RS Judson, JR Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, *et al.*: **A comprehensive analysis of**

protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000 **403**: 623-627

4. T Ito, T Chiba, R Ozawa, M Yoshida, M Hattori, Y Sakaki: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci U S A* 2001 98: 4569-4574
5. TR Hughes, MJ Marton, AR Jones, CJ Roberts, R Stoughton, CD Armour, HA Bennett, E Coffey, H Dai, YD He, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000 102: 109-126
6. CJ Roberts, B Nelson, MJ Marton, R Stoughton, MR Meyer, HA Bennett, YD He, H Dai, WL Walker, TR Hughes, *et al.*: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles.** *Science* 2000**287** : 873-880
7. PT Spellman, G Sherlock, MQ Zhang, VR Iyer, K Anders, MB Eisen, PO Brown, D Botstein, B Futcher: **Comprehensive Identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell.* 1998 **9**: 3273 3297
8. C von Mering, R Krause, B Snel, M Cornell, SG Oliver, S Fields, P Bork: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002**417** : 399-403
9. YPD™ http://www.incyte.com/sequence/proteome/databases/YPD.shtml
10. http://mips.gsf.de/proj/yeast/catalogues/funcat/fc14_04_03.html
11. M Peter, I Herskowitz: **Direct inhibition of the yeast cyclin-dependent kinase Cdc28-Cln by Far1.** *Science* 1994**265** : 1228-1231
12. T Chen, J Kurjan: **Saccharomyces cerevisiae Mpt5p interacts with Sst2p and plays roles in pheromone sensitivity and recovery from pheromone arrest.** *Mol Cell Biol.* 1997**17** : 3429 3439
13. K Madden, M Snyder: **Cell polarity and morphogenesis in budding yeast.** *Annu Rev Microbiol.* 1998 **52**: 687-744
14. D Pruyne, A Bretscher: **Polarization of cell growth in yeast. I. Establishment and maintenance of polarity states.** *J Cell Sci.* 2000 **113**: 365-375
15. PM Pryciak, LH Hartwell: **AKR1 encodes a candidate effector of the G beta gamma complex in the Saccharomyces cerevisiae pheromone response pathway and contributes to control of both cell shape and signal transduction.** *Mol Cell Biol.* 1996**16** : 2614-2626
16. H Qadota, CP Python, SB Inoue, M Arisawa, Y Anraku, Y Zheng, T Watanabe, DE Levin, Y Ohya: **Identification of yeast Rho1p GTPase as a regulatory subunit of 1,3-beta-glucan synthase.** *Science* 1996**272** : 279-281
17. J Roy, B Zheng, BC Rymond, JL Woolford Jr.: **Structurally related but functionally distinct yeast Sm D core small nuclear ribonucleoprotein particle proteins.** *Mol Cell Biol.* 1995**15** : 445-455

18. G Paravicini, L Friedli: **Protein-protein interactions in the yeast PKC1 pathway: Pkc1p interacts with a component of the MAP kinase cascade.** *Mol Gen Genet.* 1996 251: 682-691
19. M Geymonat, L Wang, H Garreau, M Jacquet: **Ssa1p chaperone interacts with the guanine nucleotide exchange factor of ras Cdc25p and controls the cAMP pathway in Saccharomyces cerevisiae.** *Mol Microbiol.* 1998 **30**: 855-864
20. WB Pratt: **The hsp90-based chaperone system: involvement in signal transduction from a variety of hormone and growth factor receptors.** *Proc Soc Exp Biol Med.* 1998**217** : 420-434
21. JF Louvion, T Abbas-Terki, D Picard: **Hsp90 is required for pheromone signaling in yeast.** *Mol Biol Cell.* 1998**9** : 3071 3083
22. O Donze, D Picard: **Hsp90 binds and regulates Gcn2, the ligand-inducible kinase of the alpha subunit of eukaryotic translation initiation factor 2.** *Mol Cell Biol.* 1999**19** : 8422 8432
23. CL Tucker, JF Gera, P Uetz: **Towards an understanding of complex protein networks.** *Trends Cell Biol.* 2001**11** : 102-106
24. A Zien, R Kuffner, R Zimmer, T Lengauer: **Analysis of gene expression data with pathway scores.** *Proc Int Conf Intell Syst Mol Biol.* 2000**8** :407-417
25. AJ Hartemink, DK Gifford, TS Jaakkola, RA Young: **Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.** *Pac Symp Biocomput* 2001 422-433.
26. EA Elion: **Pheromone response, mating and cell biology.** *Curr Opin Microbiol.* 2000 **3**: 573-581
27. P Kemmeren, NL van Berkum, J Vilo, T Bijma, R Donders, A Brazma, FC Holstege: **Protein interaction verification and functional annotation by integrated analysis of genome-scale data.** *Mol Cell.* 2002 9: 1133-1143
28. T Leeuw, A Fourest-Lieuvin, C Wu, J Chenevert, K Clark, M Whiteway, DY Thomas, E Leberer: **Pheromone response in yeast: association of Bem1p with proteins of the MAP kinase cascade and actin.** *Science* 1995**270** : 1210-1213
29. A Nern, RA Arkowitz: **A Cdc24p-Far1p-Gbetagamma protein complex required for yeast orientation during mating.** *J Cell Biol.* 1999 144: 1187-202
30. AC Butty, PM Pryciak, LS Huang, I Herskowitz, M Peter: **The role of Far1p in linking the heterotrimeric G protein to polarity establishment proteins during yeast mating.** *Science* 1998 **282**: 1511 1516
31. TP Garrington, GL Johnson: **Organization and regulation of mitogen-activated protein kinase signaling pathways.** *Curr Opin Cell Biol.* 1999 **11**: 211-218
32. HD Madhani, GR Fink: **The riddle of MAP kinase signaling specificity.** *Trends Genet.* 1998**14** : 151-155

33. L Ongay-Larios, AL Savinon-Tejeda, MJ Williamson Jr, M Duran-Avelar, R Coria: **The Leu-132 of the Ste4(Gbeta) subunit is essential for proper coupling of the G protein with the Ste2 alpha factor receptor during the mating pheromone response in yeast.** *FEBS Lett.* 2000**467** : 22-26
34. Y Ho, A Gruhler, A Heilbut, GD Bader, L Moore, SL Adams, A Millar, P Taylor, K Bennett, K Boutilier, *et al.*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002 **415**: 180-183
35. AC Gavin, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, JM Rick, AM Michon, CM Cruciat, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002 **415**: 141-147
36. MB Yaffe, GG Leparc, J Lai, T Obata, S Volinia, LC Cantley: **A motif-based profile scanning approach for genome-wide prediction of signaling pathways.** *Nat Biotechnol.* 2001**19** : 348-353
37. H Zhu, JF Klemic, S Chang, P Bertone, A Casamayor, KG Klemic, D Smith, M Gerstein, MA Reed, M Snyder: **Analysis of yeast protein kinases using protein chips.** *Nat Genet.* 2000 **26**: 283-289
38. D Endy, R Brent: **Modelling cellular behaviour.***Nature* 2001**409** : 391-395
39. ML Bulyk, X Huang, Y Choo, GM Church: **Exploring the DNA-binding specificities of zinc fingers with DNA microarrays.** *Proc Natl Acad Sci U S A.* 2001 **98**: 7158-7163
40. A Lueking, M Horn, H Eickhoff, K Bussow, H Lehrach, G Walter: **Protein microarrays for gene expression and antibody screening.** *Anal Biochem.* 1999 **270**: 103-111
41. N Mochizuki, S Yamashita, K Kurokawa, Y Ohba, T Nagai, A Miyawaki, M Matsuda: **Spatio-temporal images of growth-factor-induced activation of Ras and Rap1.** *Nature* 2001 **411**: 1065 1068
42. CJ Marshall: **Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation.** *Cell* 1995 **80**: 179-185
43. T Cormen, C Leiserson, R Rivest: **Introduction to Algorithms**. *Cambridge, MA, MIT Press* 1990
44. V Batagelj, A Mrvar:**Pajek - Program for Large Network Analysis**. *Connections* 1998 **21:** 47 57

Figures

Figure 1 - MAPK signal transduction pathways in yeast

Membrane proteins are depicted in blue, transcription factors in red, and intermediate proteins in green. Figure adapted from [6].

Figure 2 - Histogram of the number of proteins with a given number of protein-protein interactions

Interaction data obtained by high-throughput two-hybrid assays [2-4]. The highly interacting proteins in red were removed from the interaction dataset (see text for details).

Figure 3 - Histogram of pathways with a given coclustering score for experimental and randomized protein-protein interactions

Histogram of the number of pathways with a given coclustering score for experimental and randomized protein-protein interactions. Shown here is the tail of the distribution with the highest coclustering scores. The paths were drawn with a depth-first search algorithm [43] from membrane to DNA-binding proteins. It is evident that at high coclustering scores, pathways from the experimentally observed interactions (blue) outnumber those generated from randomized interactions (red – an average of three separate randomizations). The total number of paths for experimental and randomized interaction data (averaged) were within 5% of each other.

Figure 4 - Network models produced by NetSearch

Pathways predicted by NetSearch for (A,B) pheromone response, (C) cell wall integrity, and (D) filamentation pathways, with the starting membrane protein for path drawing (blue), intermediate proteins (green) and transcription factor (red). In each case, the fifteen highest ranked paths between common endpoints were combined to form the signaling network. For the cell wall integrity pathway, the sensor proteins that initiate signal transduction Wsc/1/2/3p and Mid2p did not have any productive interactions. For this pathway, we began our searches at Rho1p and searched for a path length of seven. The size of each vertex is proportional to the sum of the scores of the paths in which it was included. Network graphs were produced with PAJEK graph drawing software [44, http://vlado.fmf.uni-lj.si/pub/networks/pajek].

Additional files

Supplementary website – http://arep.med.harvard.edu/NetSearch

Web interface for NetSearch: http://arep.med.harvard.edu/NetSearch/runprog.html

Figure 1

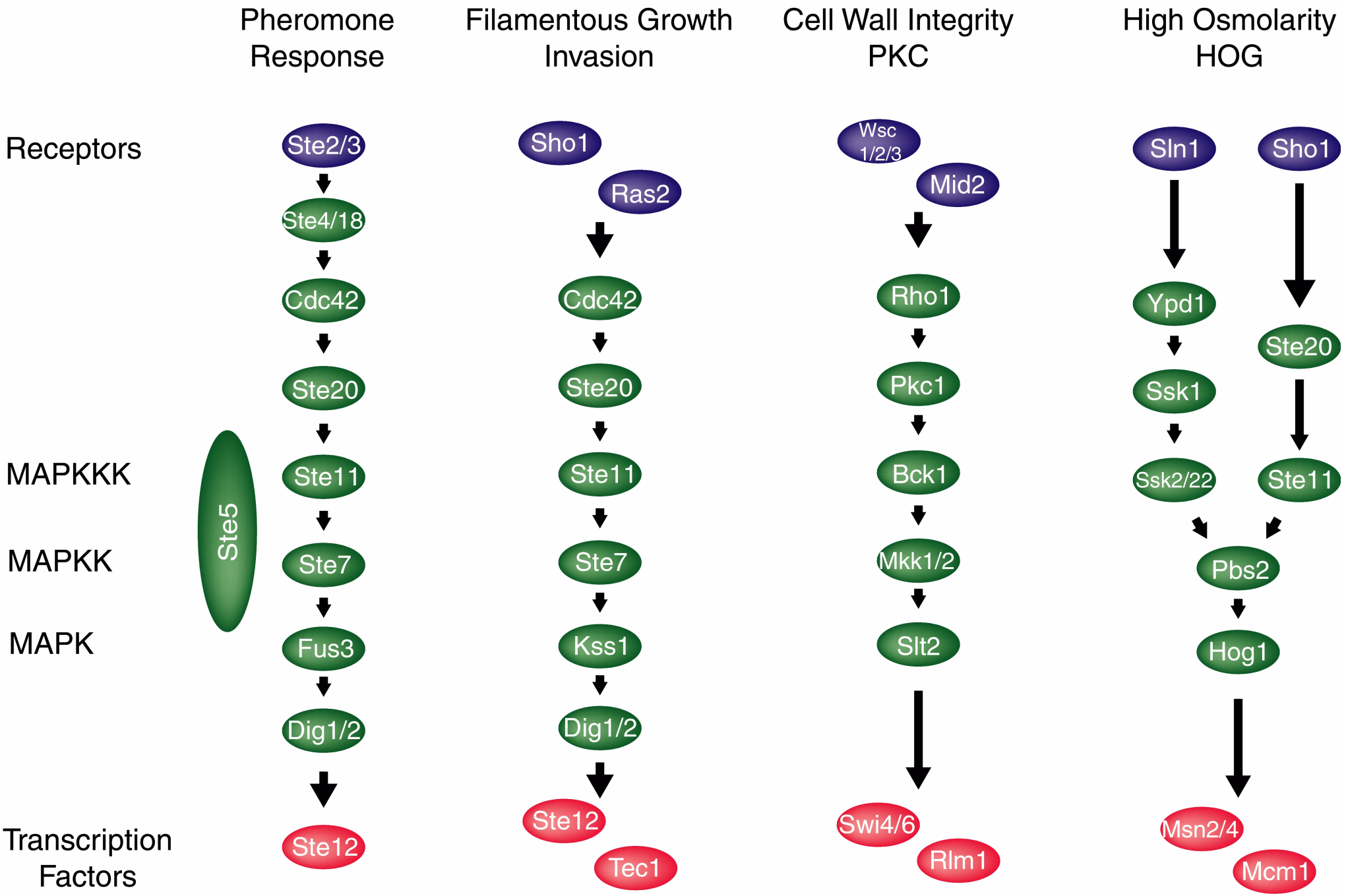

Pheromone Response
Filamentous Growth Invasion
Cell Wall Integrity PKC
High Osmolarity HOG
Receptors
MAPKKK
MAPKK
MAPK
Transcription Factors
Ste2/3
Ste4/18
Cdc42
Ste20
Ste11
Ste7
Fus3
Dig1/2
Ste12
Ste5
Sho1
Ras2
Cdc42
Ste20
Ste11
Ste7
Kss1
Dig1/2
Ste12
Tec1
Wsc 1/2/3
Mid2
Rho1
Pkc1
Bck1
Mkk1/2
Slt2
Swi4/6
Rlm1
Sln1
Sho1
Ypd1
Ste20
Ssk1
Ssk2/22
Ste11
Pbs2
Hog1
Msn2/4
Mcm1

Figure 2

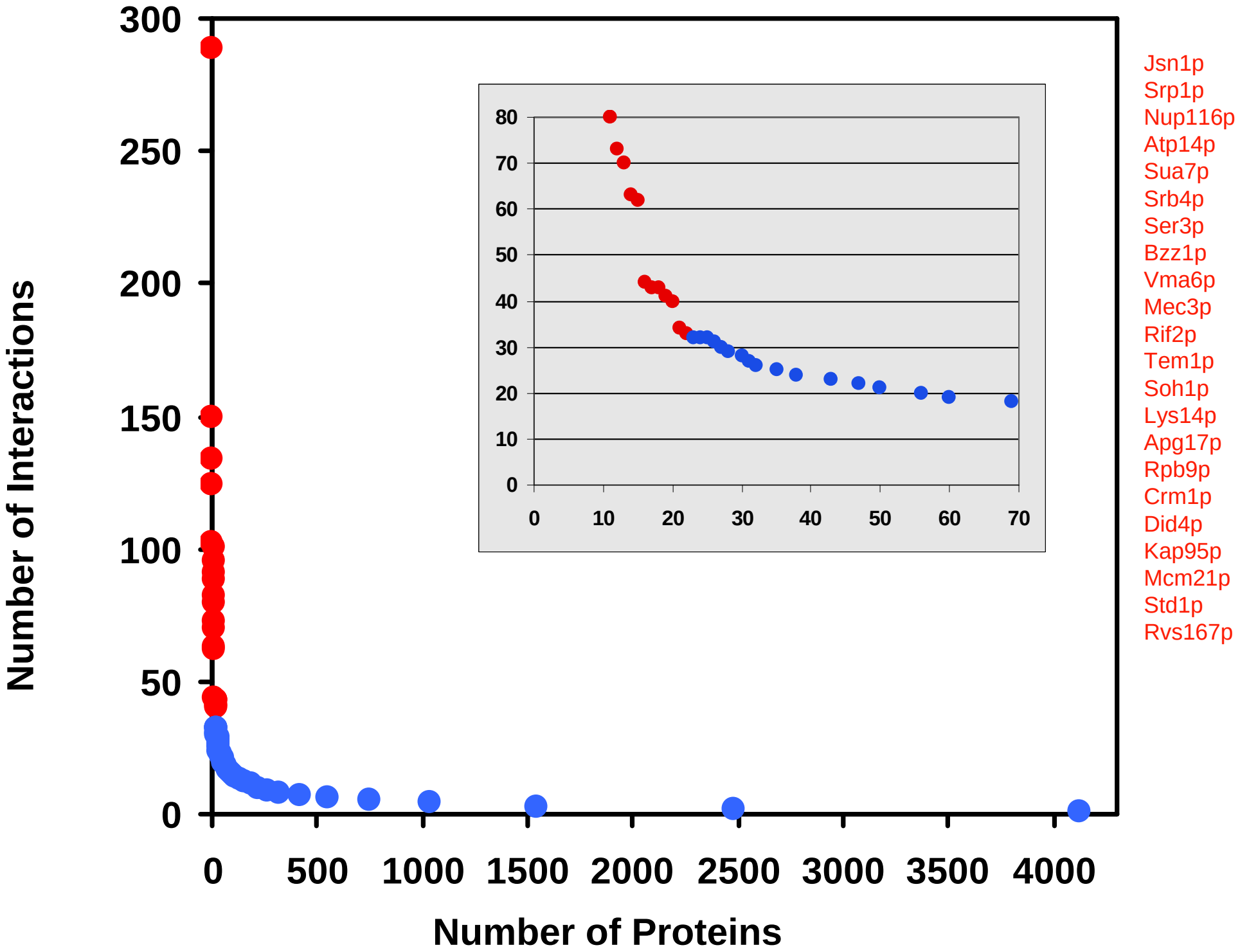
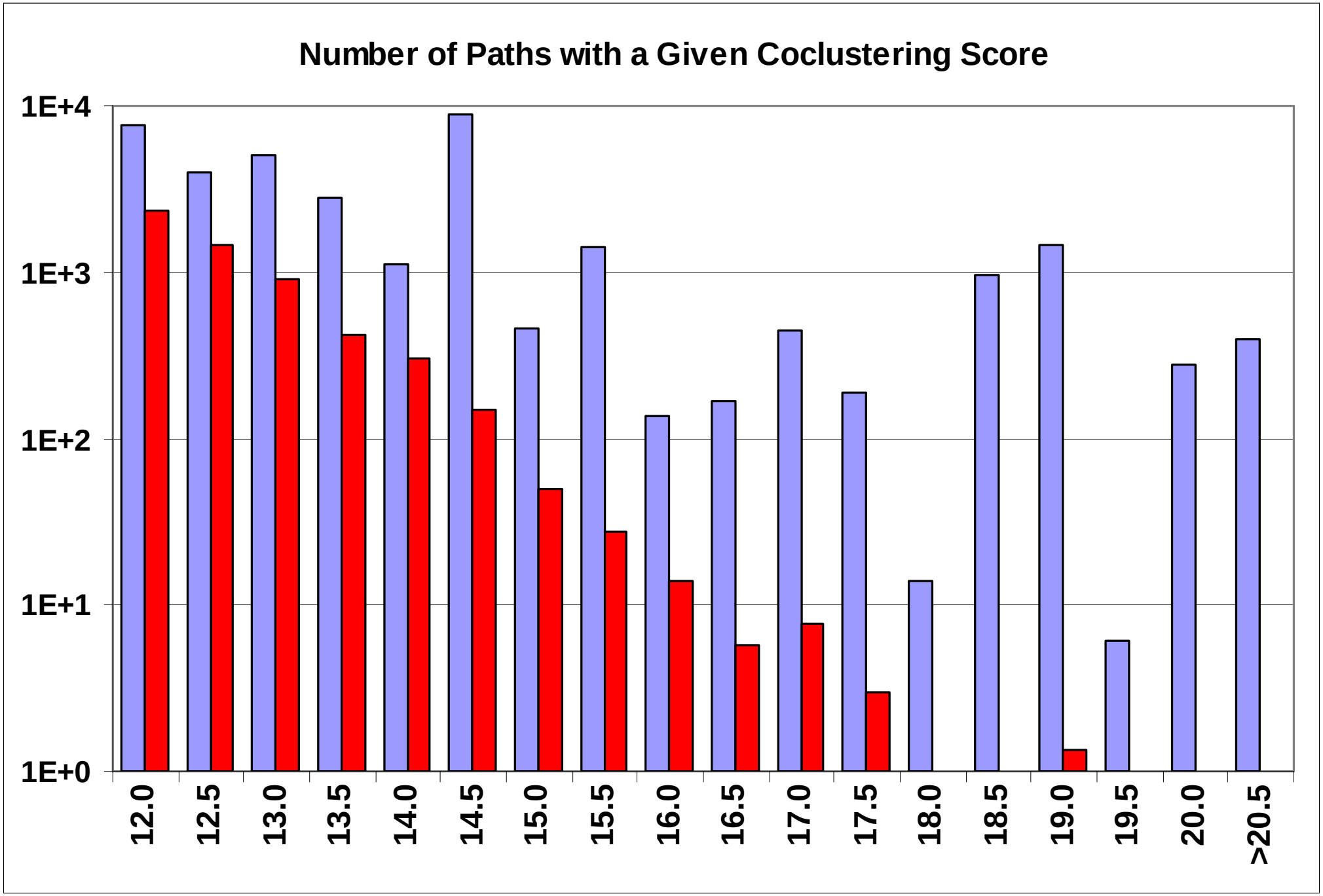

Number of Interactions
300
250
200
150
100
50
0
0
500
1000
1500
2000
2500
3000
3500
4000
Number of Proteins
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
Jsn1p
Srp1p
Nup116p
Atp14p
Sua7p
Srb4p
Ser3p
Bzz1p
Vma6p
Mec3p
Rif2p
Tem1p
Soh1p
Lys14p
Apg17p
Rpb9p
Crm1p
Did4p
Kap95p
Mcm21p
Std1p
Rvs167p

Figure 3


Number of Paths with a Given Coclustering Score
1E+4
1E+3
1E+2
1E+1
1E+0
12.0
12.5
13.0
13.5
14.0
14.5
15.0
15.5
16.0
16.5
17.0
17.5
18.0
18.5
19.0
19.5
20.0
>20.5

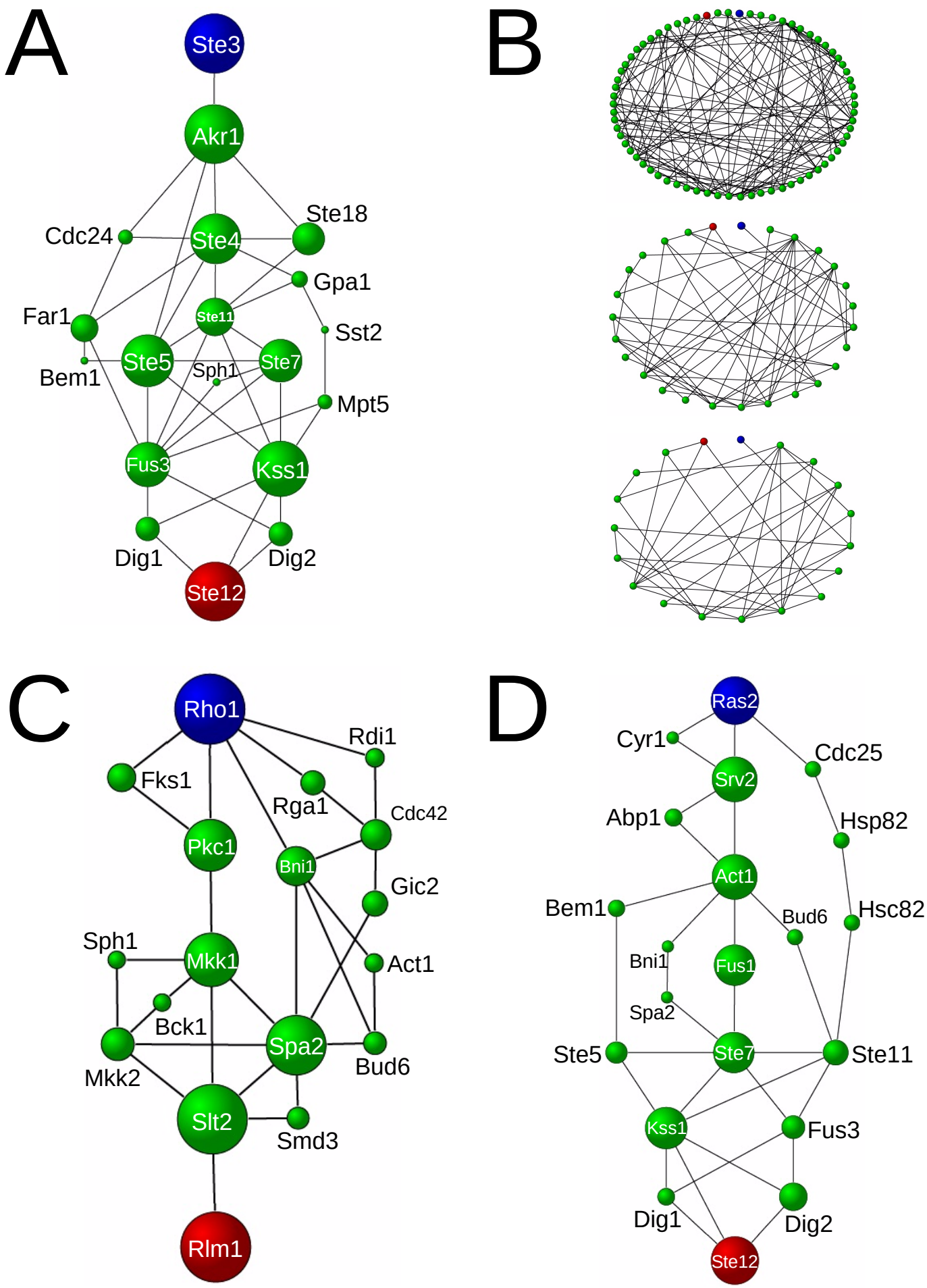

A
Ste3
Akr1
Ste18
Cdc24
Ste4
Gpa1
Far1
Ste11
Sst2
Ste5
Sph1
Ste7
Bem1
Mpt5
Fus3
Kss1
Dig1
Dig2
Ste12
B
C
Rho1
Rdi1
Fks1
Rga1
Cdc42
Pkc1
Bni1
Gic2
Sph1
Mkk1
Act1
Bck1
Spa2
Bud6
Mkk2
Slt2
Smd3
Rlm1
D
Ras2
Cyr1
Cdc25
Srv2
Abp1
Hsp82
Act1
Bem1
Bud6
Hsc82
Bni1
Fus1
Spa2
Ste5
Ste7
Ste11
Kss1
Fus3
Dig1
Dig2
Ste12

Figure 4