Motifs and Modules in Fractured Functional Yeast Networks

J. S. Hallinan and A. Wipat
CISBAN and School of Computing Science
Newcastle University
Newcastle upon Tyne, United Kingdom NE1 7RU

Abstract- **The integration of diverse data sets into probabilistic functional networks is an active and important area of research in systems biology. In this paper we fracture a previously published integrated network into its component networks, and investigate the overlap between the information provided by each data set to the final network. Using three-node network motifs as a surrogate for information about genetic circuits, we find that the same motifs are over-represented in all of the networks, but different genes contribute to the motifs in different data sets. We conclude that the data integration approach is valuable because it clearly does combine different insights into a biological system. However, the fact that the information contained in different data sets is so diverse raises issues of how best to perform data integration so as to accurately estimate error rates for different data sets, whilst including as much data as possible in the integrated network.**

I. INTRODUCTION

The word "genome" was first coined in the 1920s (a combination of "gene" and "chromosome"), and the study of genomes rapidly became the science of genomics. In the 1990s, the term "proteomics" joined the lexicon, describing the study of the full protein complement of an organism, and in the last decade a veritable flood of "omics" terms has come into being, reflecting the flood of omics data. Omics data are generated by whole-cell, high-throughput screening of biological systems and tend to be very large and very noisy. Many different types of omics data can be collected: as well as genomics and proteomics there are metabolomics, transcriptomics, interactomics, and many more. The Genomic Glossaries site lists literally dozens[1]. Although these datasets can be hard to manage, they provide the closest approximation to date of a holistic, unbiased view of the workings of biological systems.

A recent trend in the handling of omics data is to combine diverse types of data into integrated functional networks, often using statistical approaches which take into account the error rates associated with different data sets (e.g. Breitkreutz, Stark & Tyers, 2003; Gopalacharyulu *et al.*, 2005; Li, Li, Su, Chen & Galbraith, 2006; Kohler *et al.*, 2006; Baitaluk *et al.*, 2006). In these networks nodes are usually genes, while edges represent any functional relationship between a pair of genes: physical, regulatory, or any other sort or combination of sorts of interaction. The hope is that integrated networks will provide more information about intracellular interactions than will networks constructed from a single type of data, since each data source used adds information, and the use of statistical integration techniques means that many weak sources of evidence can be combined to produce a stronger link, as different sources of noise cancel each other out.

It is reasonable to ask, however, exactly what is gained by combining data sources in this way. While a network from a single data source—for example, a protein-protein interaction network—is easy to understand (an edge between gene *A* and gene *B* means that *A* physically binds to *B*, and hence the two genes probably perform similar functions (Oliver, 2000)) the meaning of an edge in an integrated network is less conceptually clear, and, indeed, will differ from edge to edge within the same network, depending upon which data sources were utilized in the construction of each edge.

We have previously investigated the relationship between aspects of network topology and dynamics, using abstract computational models of gene regulatory networks. These studies suggest that the topological features of a network influence its dynamics (in the case of gene regulatory networks, the pattern of gene expression over time) (Hallinan & Jackway, 2005), and can provide insights into the functions of genes and sets of genes. Tightly connected sets of nodes ("modules") are often involved in the same biological process (Hallinan, 2004). Network analysis can identify genes which are parts of modules, and those which mediate communication between modules (Hallinan & Wipat, 2006). In an integrated functional network links from different data sources may well provide different insights into the underlying biology. In this paper we attempt to use the approaches we have previously developed to quantify the ways in which different data sets contribute to a network.

One type of information which should be strengthened by the combination of multiple data sets is that regarding the prevalence and structure of cellular control circuits. The circuits controlling the regulation of gene expression are not comprised simply of protein-protein interactions. Protein-DNA, protein-RNA and RNA-RNA interactions are all important, as are interactions with other metabolites. Further, it is becoming increasingly apparent that epigenetic modifications to biomolecules are essential to genetic regulation (Jaenisch & Bird, 2003), as are protein modifications. These data are not yet part of interactome databases, but such interactions may be indirectly captured by an integrated network.

[1] Of which our favourite is the unknome: the complete set of genes within a genome for which there is currently no functional information.

1-4244-0710-9/07/$20.00 ©2007 IEEE

It has recently been suggested that genetic regulatory circuits may be detected in gene networks in the form of network motifs: small sets of nodes with a specific pattern of interaction (Milo *et al.*, 2002). If small functional modules are essential to network functioning, it is argued, they will be preserved by natural selection, and should be found at higher-than-chance levels in evolved networks. Small network motifs have indeed been found to be over-represented in the gene networks of model organisms such as the gut bacterium *E. coli* (Shen-Orr, Milo, Mangan & Alon, 2002; Dobrin, Beq, Barabasi & Oltvai, 2004) and the yeast *Saccharomyces cerevisiae* (Wuchty, Oltvai & Barabasi, 2003; Lee *et al.,* 2002). Interestingly, though, these studies have involved networks constructed from a single data type (transcriptional networks and protein-protein interaction networks, respectively). We suggest that single-source networks can provide only incomplete information about functionally important genetic circuits in intracellular networks, since such circuits must of necessity involve several different types of interactions.

In this paper we compare the motif structure of an integrated functional network of the baker's yeast *Saccharomyces cerevisiae* with that of the single-source networks which underlie the integrated network. By fragmenting the network in this way, and comparing the topology of the fragments, we can investigate not only how much information each data source contributes to the integrated network, but also how relevant that information is to a particular set of questions about biological function.

II. METHODS

A. Data Sets

The integrated functional network we used was compiled and published by Lee, Date, Adai & Marcotte (2004). It consists of data from 11 different sources, combined using a Bayesian statistics approach to yield a network in which nodes represent genes and edges represent functional interactions between genes. The edges are weighted to reflect the data set quality and the probability that an interaction exists, based upon the error rates in each data set as compared with a gold standard (KEGG (Kanehisa & Goto, 2000) and Gene Ontology data (The Gene Ontology Consortium, 2000). The sources used to construct the integrated network are briefly described in Table 1.

TABLE 1.
DATA SOURCE S FOR THE LEE NETWORK

Name	Type	Reference
Gavin	Protein complexes	Gavin *et al*., 2002
Ho	Protein complexes	Ho, 2002
Uetz	Protein – protein interactions	Uetz *et a*l., 2000
Ito	Protein - protein interactions	Ito *et al*., 2001
Tong01	Functional interactions	Tong *et al*., 2001
Tong02	Protein – protein interactions	Tong *et al.,* 2002
Coexp	717 microarrays	Gollub *et al*., 2003
DIP	Protein – protein interactions	Xenarios *et al*., 2000

Phyl	Phylogenetic co-evolution	Pellegrini *et al*., 1999;
		Huynen, Snel, Lathe & Bork, 2000;
		Wolf, Rogozin, Kondrashov & Koonin, 2001

Fusion	Gene fusion	Marcotte *et al.,* 1999;
		Enright, Illiopoulos, Kyrpides & Ouzounis, 1999;
		Yanai, Derti & DeLisi, 2001

Cocite	Literature cocitation	Stapley & Benoit, 2000; Jenssen, Laegrid, Komorowski & Hovig, 2001

The Lee *et a.l* network contains data for all possible pairs of genes, and therefore includes statistically non-significant edges. We extracted the 30,000 most highly weighted interactions from the Lee data set and used them to construct a network consisting of 4,514 nodes and 30,000 edges. We then divided this data set into its components, retaining for each single network only those nodes for which there were edges in that data set. The statistics of each of these networks are shown in Table 2.

TABLE 2
STATISTICS OF THE FRACTURED NETWORK
LCC IS THE SIZE OF THE LARGEST CONNECTED COMPONENT OF THE NETWORK

Data Source	Nodes	Edges	Avg. Conn.	Comp-onents	LCC (%)
All	4514	30000	6.7	82	4302 (95.3)
Coexp.	2241	20061	8.9	149	1862 (83.1)
Cocite	1862	2658	1.4	158	999 (53.6)
Gavin	818	1217	1.5	113	311 (38.0)
Ho	481	476	1.0	92	76 (16.0)
Ito	437	299	0.7	152	38 (8.7)
Phyl.	1271	5902	4.6	98	973 (76.5)
Fusion	330	496	1.5	72	49 (14.8)
DIP	1515	2822	1.8	55	1345 (88.8)
Tong01	17	12	0.7	6	4 (23.5)
Tong02	29	29	1.0	4	21 (72.4)
Uetz	371	252	0.68	132	12 (3.2)

It is immediately apparent from Table 2 that different data sets contribute to the final network to different extents. While the full network s has 4,514 nodes, only 17 of these are present in the Tong01 network, for example. This is because of the selection of the top 30,000 most highly weighted connections from the full Lee data set to construct the network used here. It would be instructive to examine all of the individual networks in full, and we intend to do this in the near future.

It is evident from Table 2 that the statistics of the networks differ considerably. Their visual appearances are also very dissimilar (Fig. 1).

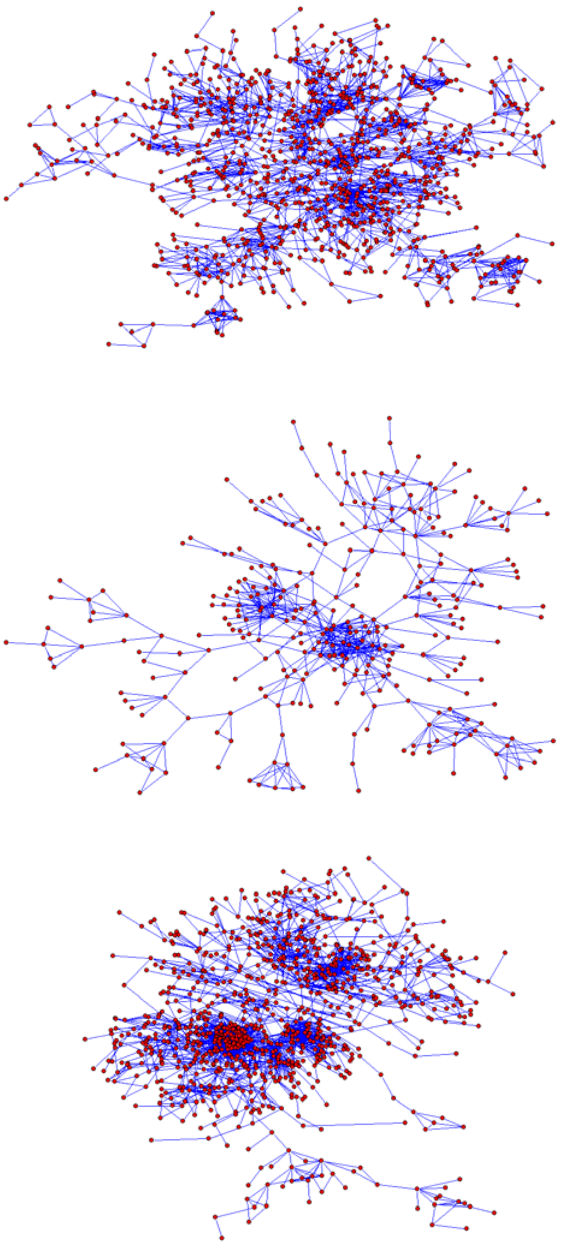


Fig. 1. Topologies of some of the fractured networks. The integrated network is not show, since its high connectivity makes visualization meaningless. a) Largest connected component of the cocitation network; b) Largest connected component of the Gavin network; c) Largest connected component of the phylogenetic network.

B. *Motif Detection*

As a surrogate for small genetic circuits, all possible network motifs of size 3 were exhaustively enumerated in each network using the motif detection application FANMOD (Wernicke & Rasche, 2006). The algorithm implemented by FANMOD assesses the importance of motifs in terms of their frequency of occurrence. The number of each motif occurring in the network of interest is compared with the average count for a suite (default size 1,000) of randomly-generated networks with the same number of nodes and edges, and the same connectivity pattern, and motifs which are statistically over-represented are reported.

C. *Motif Analysis*

The over-represented motifs and their corresponding adjacency matrices were stored in a PostgresSQL relational database using custom Java code and then integrated with information pertaining to gene functional annotations derived from the Gene Ontology database. In order to enable the comparison of motifs between and within networks, the descriptions of genes comprising motifs were re-ordered by reference to a predefined indexed set of genes to ensure consistency in the order of genes in the representation of each of the three motifs. Their corresponding adjacency matrices were also manipulated to reflect this re-ordering, again using custom Java programs to perform the necessary manipulations. Once in the database, the most genes most commonly represented in over-represented motifs from each network, and those motifs common between the motif sets of individual networks, were identified using SQL queries over the database. The relationships between gene functions and their annotations of genes in motifs were identified manually by reference to the database tables. The data relating to *S. cerevisiae* gene function and mutant phenotypes were downloaded from the Munich Information Centre for Protein Sequences (MIPS)[2] .

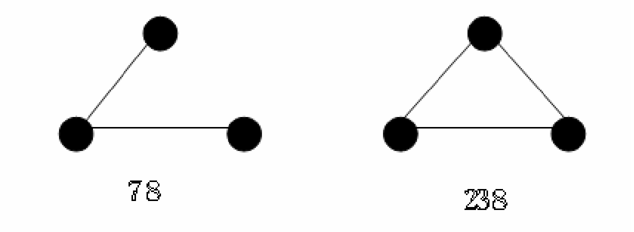
D. *Motif Merging*

In most networks one gene participated in multiple instances of an over-represented motif. In order to investigate the relationship of these genes to each other within a network, we extracted from each network the subnetwork consisting of those genes which participated in over-represented motifs, and examined the characteristics of these subnetworks.

III. RESULTS

E. *Motif Occurrence*

Two motifs (dubbed 78 and 238 by FANMOD) were over-represented in 11 of the 12 networks, and one (78) was the only motif over-represented in the small Tong01 network (Figure 2).



[2] ftp://ftpmips.gsf.de/yeast/archive/

Figure 2. Motifs over-represented in the networks

Although the pattern of connectivity between genes—the functional genetic circuit—was common to all the networks, the genes which were overrepresented in the most common motifs differed from network to network, and within networks (Table 3).

TABLE 3. STATISTICS OF GENES IN OVER-REPRESENTED MOTIFS

Network	Genes	Number of Occurrences			
		Min	Max	Mean	Mode
All	410	3	34,128	1455.7	3
Tong01	7	6	18	11.1	6
Tong02	14	6	90	27.9	6
Phyl	788	6	30,960	1198.1	6
Ito	99	6	66	14.2	6
Ho	189	6	492	39.7	6
Gavin	379	6	1,476	88.0	6
Fusion	144	6	1,074	120.5	6
Uetz	79	6	168	16.9	6
DIP	868	6	4,530	120.5	6
Cocite	787	6	1,770	82.0	6
Coexp	1,444	6	67,320	3,820.0	6

The number of genes involved in motifs varied from seven to 1,444, reflecting, in part, the sizes of the networks. In all of the individual networks genes occurred in at least two motifs (minimum number of occurrences = 6), but in the integrated network the minimum and mode of the number of occurrences was three, indicating that most genes were only represented in a single motif (although a few were far more frequently occurring). The integrated network does indeed seem to smooth out the statistics of the individual networks.

The single gene most frequently occurring in over-represented motifs also varies from network to network (Table 4).

TABLE 4. GENES MOST FREQUENTLY OCCURRING IN OVER-REPRESENTED MOTIFS

Net	Gene	Gene Ontology Function
All	YBR048W	ribosomal protein S11.e.B
Tong01	YDR004W	DNA repair protein
Tong02	YDR388W	reduced viability upon starvation protein
Phyl	YBR191W	ribosomal protein L21.e
Ito	YGR218W	nuclear export factor, exportin
Ho	YGR103W	similarity to zebrafish essential for embryonic development gene pescadillo
Gavin	YBR247C	effects N-glycosylation
Fusion	YBL112C	strong similarity to subtelomeric encoded proteins
Uetz	YDR328C	kinetochore protein complex CBF3, subunit D
DIP	YBR109C	calmodulin
Cocite	YBR160W	cyclin-dependent protein kinase
Coexp	YBR048W	ribosomal protein S11.e.B

The overlap between the genes involved in over-represented motifs in different networks was also minimal. Most of the genes identified occurred in only network, and none occurred in more than eight of the eleven separate networks (Figure 2).

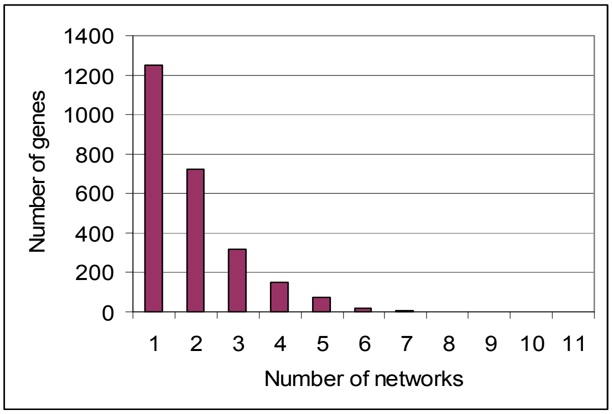


Figure 2. Most genes in over-represented motifs occur in only one network

Since many of the individual networks are protein-protein interaction networks, it is likely that they are providing similar information. To investigate this possibility we looked at the distribution of identical motifs (i.e. those involving the same genes and the same pattern of connectivity) across the protein-protein interaction networks (Table 5.).

TABLE 5. MOTIFS COMMON TO PROTEIN-PROTEIN INTERACTION NETWORKS

Network 1	Network 2	Network 3	Common Motifs
Uetz	DIP		242
Uetz	Ito		234
Ito	DIP		330
Ito	Uetz	DIP	226

Table 5 shows that there is, indeed, considerable overlap in motifs between the protein interaction networks.

We hypothesize that genes in statistically over-represented motifs which are common to several networks are likely to be essential to the viability of the cell. In order to investigate this possibility, we annotated the genes summarized in Table x. with phenotypic data from the *Saccharomyces* Gene Database (SGD) (Table 6).

TABLE 6. NUMBER OF ESSENTIAL GENES PER MOTIF IN PROTEIN-PROTEIN NETWORKS

Networks	Number of motifs with essential genes			
	0 essential	1 essential	2 essential	3 essential
Uetz/DIP	120	20	56	44
Ito/Uetz	104	40	76	12
Ito/DIP	100	140	60	28
Ito/Uetz/DIP	192	32	0	0

The distinction between the individual and the integrated networks is particularly marked in Table 6. In all of the individual networks the number of motifs in which one or more genes is essential is greater than the number in which none are essential, but in the integrated network the reverse is true.

The genes participating in over-represented motifs were merged to form subnetworks for each network. These subnetworks showed considerable variability, with the majority of subnetworks being relatively connected (Fig. 3, 4), while a few are almost completely unconnected (Fig. 5).

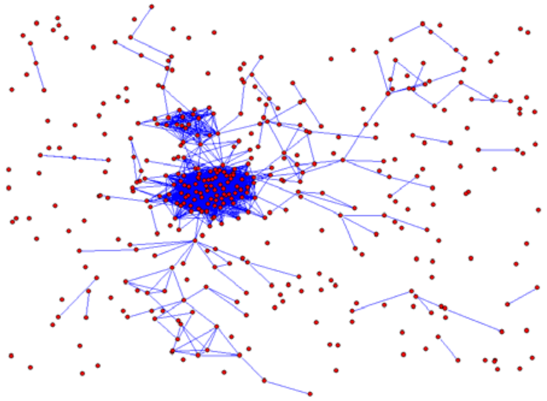


Figure 3. Merged motifs of complete network

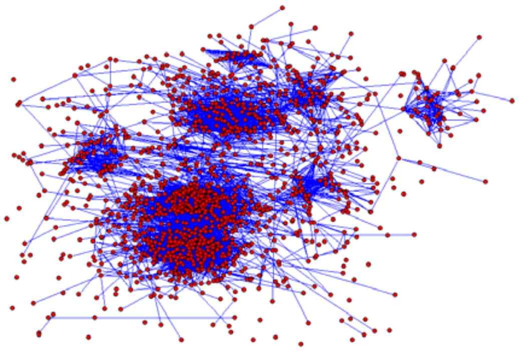


Figure 4. Genes participating in over-represented motifs in the coexpression network

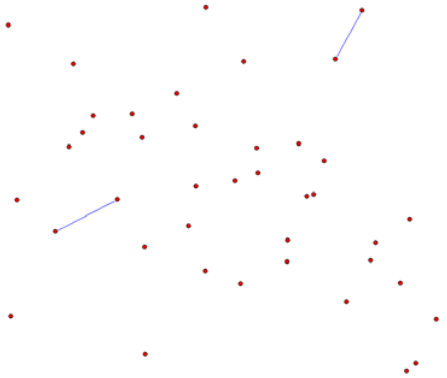


Figure 5. Genes participating in over-represented motifs in the fusion network

IV. DISCUSSION

The average connectivity of the individual networks varies from 0.7 to 8.9. This metric would appear to reflect the specificity of the technique used to generate the data; the low average connectivity networks were generated using approaches such as the detection of functional interactions, while the high-connectivity nets arise from techniques such as microarray analysis, which is known to be noisy and hence will produce many spurious interactions. The high average connectivity of the integrated network clearly reflects the impact of the coexpression network.

The result of a functional analysis of the genes encoding proteins most commonly represented in the motifs varies considerably across the individual networks that comprise the overall integrated network. For example, for those networks representing protein interactions e.g. protein interaction networks (DIP, Ito, Ho, Uetz and Gavin), the proteins most heavily represented in the three motifs in fail to show significant overall correlation between datasets. This variation possibly reflects the fact that even though these networks describe the same phenomenon, the underlying experimental methodology used to derive them varies considerably.

The Uetz network (Uetz, 2000), a general yeast two hybrid generated network, shows two motifs with cyclin specific proteins such as Pcl2 and Pcl9, cyclin dependant kinase regulatory subunits and cyclin like proteins, dominating the top 25 most overrepresented proteins. Interestingly motifs derived from the both DIP and the Ito protein-protein interaction networks show an abundance of proteins localized to the nucleus. In the top 20 DIP dataset (from the database of interacting proteins, derived mostly by small scale protein interaction assays (Xenarios, 2002)) motif proteins show an overrepresentation of nuclear proteins (such as Cdc39), nuclear export factors such as Crm1, in particular nuclear pore proteins such as Nup42, Nup49, Nup60, Nup84 and Nup40.

The comprehensive yeast two-hybrid derived dataset generated by Ito *et al.* (2001) also demonstrates a statistically significant excess of proteins involved in the structure and function of the nucleus, including nuclear pore proteins, Nup57 and Nup58 and also the nuclear export factor Crm1. The dataset generated by Tong and co-workers (2002), was also essentially derived from yeast-two hybrid technology, but with protein baits consisting of cloned peptide fragments using phage display technology. In this case, the most frequent motifs from this source involve proteins with a bias towards cellular structural components such as the actin binding protein Asb1, myosin I (Myo3 and Myo5) and the actin binding protein verprolin, encoded by *vrp1* (Anderson *et al.,* 1998). In addition, there are a number of proteins of unknown function in the mostly frequently represented proteins.

The three motifs derived from a dataset generated by tandem affinity purification (Ho *et al.,* 2002) demonstrate an overrepresentation of proteins that belong to 'functional' complexes, in particular the proteosome (responsible for protein processing and breakdown, proteolysis). The motifs also include structural components of the spliceosome and other proteins involved in its assembly, such as Prp43, together with proteins responsible for RNA processing and maturation. Motifs derived from the Gavin networks (Gavin, 2002), generated by a similar approach, also show similar characteristics including a variety of proteins that are also members of complexes, including proteins that are proteosome subunits, proteins for proteosome regulation, some members of the RNA polymerase complex and proteins required for ribosomal biogenesis.

The Phyl network was generated by Marcotte *et al*., 1999 by phylogenetic profiling based on a comparative genomics approach. This network clearly demonstrates an overwhelming number of motifs that contain genes encoding proteins which are responsible for the structure of ribosomes. Strikingly, more than 50 of the most overrepresented proteins in the three motifs from this network are ribosomal structural proteins.

Genes overrepresented in the most common motifs derived from the co-citation network show wide variation in their functional classification, in contrast to the other networks discussed above, that show a bias in the function of the proteins involved in their most frequent motifs. Analysis of the function of proteins of the three-motifs derived from the co-expression network indicates that this network shows the most similarity in terms of gene function to those motifs derived from the complete integrated network. This observation suggests that the coexpression network contributes most to shaping the relationships within the integrated network, possibly masking out the effect of the other datasets, despite the probabilistic approach to accounting for data quality.

Integrated networks, such as the one investigated here, seem to be biased to particular data sets, such as protein-protein interactions. In some ways, this is a reflection of the types of ‘omics’ experiments that are around at the time. There is a clear requirement for the integration of many other types of interactomes that describe different regulatory layers of the gene to protein to phenotype process if we are to maximise the benefit of the integrated network approach. In particular, epigenetic and protein phosphorylation networks are conspicuous in their absence.

Since the strength of the integrated network approach is argued to be its ability to represent functional, rather than merely physical, interactions between genes and gene products, we chose to concentrate upon the occurrence of three-node motifs in the network;. such motifs have been proposed as being crucial to the dynamics of gene expression. We found that the same two network motifs are statistically over-represented in all of the networks. Since the motifs are so small, they occur many times in each network—tens of thousands of times in the largest networks—and involve large numbers of genes. The genes participating in these motifs, while overlapping, were not identical in each network, an observation which supports the suggestion that different types of data are providing different information about the underlying networks.

Merging the over-represented motifs reveals the connections between the genes which are active in the over-represented motifs. In most of the networks these genes form a small number of connected components, an observation which implies that the functional core of the network is tightly

V. CONCLUSIONS

The rationale behind the construction of integrated functional networks of gene interactions is to make the best use of the vast quantities of omics data currently being generated. By combining data from different sources, and weighting it according to the error rates inherent in the different techniques used, researchers aim to generate a more detailed picture of functional interactions than can be gained from any single experiment.

We find that, as expected, networks generated from single data sources have characteristics very different from those of the network produced by integrating all of the data. The network we analyzed, that produced for *S. cerevisiae* by Lee *et al.* (2004) is heavily dependant upon protein-protein interaction data, with six of the 11 data sources being various measures of protein-protein interaction. Not surprisingly, the protein-protein interaction data have more in common with each other than with the other data sources, although they still differ in their contribution to the network.

integrated. In the fusion network, however, the genes tend to be isolated.

This confirmation that individual networks provide different sorts of information about the underlying biological system raises the issue of how best to combine them. The approach usually taken, which was adopted by Lee *et al.*, is to weight each data set according to how accurately it reproduces a "gold standard" set of interactions. The gold standard is generally a heavily manually curated data set such as the KEGG or Gene Ontology databases. This approach is problematic for two major reasons. Firstly it means that the most reliable data is automatically excluded from the integrated network. The requirement of Bayesian statistics for a calculated prior probability means that a gold standard is essential, but given how noisy most omics datasets are, the exclusion of the highest-quality data from the final network will inevitably degrade its accuracy.

Perhaps more importantly, the variability in information content between the individual datasets casts doubt upon the whole concept of a gold standard dataset. The power of data integration lies in the merging of different lines of evidence with regard to genetic interactions, but the fact that different types of data capture different information means that the use of a single gold standard data set is probably not realistic. This issue could be addressed by defining data type categories and establishing a gold standard for each data type. However, while this may be relatively straightforward for a data category such as protein-protein interactions, for which there is a lot of data available, gold standards for categories such as citation or gene fusion would be much harder to identify. Further, the use of multiple gold standards compounds the problem of omitting valuable data from the integrated network.

Data integration is undoubtedly a powerful tool for computational systems biology, offering a principled way to incorporate large amounts of diverse data into a single unified network of interactions. Integrated networks provide an overview of the functional interactions between genes, and will be increasingly important in the analysis of high-throughput data. We have found that different types of data contribute very different information about genetic control circuits, in the form of small network motifs, with the genes involved in statistically over-represented motifs being different in different data sets, even though the wiring pattern of the motifs are the same.

Although the value of the data integration approach to gene network construction is unarguable, the practical and statistical issues involved in integrating different data types into a single network are complex, and clearly require further investigation.

REFERENCES

[1] Anderson, B. L., Boldogh, I., Evangelista, M., Boone, C., Greene, L.A. & Pon, L.A. (1998). "The Src homology domain 3 (SH3) of a yeast type I myosin, Myo5p, binds to verprolin and is required for targeting to sites of actin polarization."." *J.Cell Biol.* 141(6), 1357-1370.

[2] Baitaluk, M., Qian, X., Godbole, S., Raval, A., Ray, A. & Gupta, A. (2006). PathSys: Integrating molecular interaction graph for systems biology. *BMC Bioinformatics* 7(55) http://www.biomedcentral.com /1471-2105-7-55, Downloaded 01/10/2006.

[3] Breitkreutz, B.-J., Stark, C. & Tyers, M. (2003). Osprey: A network visualization system. *Genome Biology* 4(3): R22.

[4] Dobrin, R., Beq, Q. C., Barabasi, A. L. & Oltvai, S. N. (2004). Aggregation of topological motifs in the *Escherichia coli transcriptional regulatory network. BMC Bioinformatics 5*(10).

[5] Enright, A. J., Illiopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402: 86 - 90.

[6] Gavin, A. -C., Bosche, M., Krause, R., Grandi, P., Marzioch, M. & et. al (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141 - 147.

[7] Gollub, J., Ball, C. A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J. C., Schroeder, M., Brown, P. O., Botstein, D. & Sherlock, G. (2003). The Stanford Microarray Database: Data access and quality assessment tools. *Nucleic Acids Research* 31(1): 94 - 96.

[8] Gopalacharyulu, P. V., Lindfors, E., Bounsaythip, C., Kivioja, T., Yetukuri, L., Hollmen, J. & Oresic, M. (2005). Data integration and visualization system for enabling conceptual biology. *Bioinformatics* 21(Suppl. 1): i177 - i185.

[9] Hallinan, J. (2004). Cluster analysis of the p53 network: Topology and Biology. *2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. San Diego 7-8 October, 2004.

[10] Hallinan, J. & Jackway, P. (2005). Network motifs, feedback loops and the dynamics of genetic regulatory networks. *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE Press. 90 - 96.

[11] Hallinan, J. & Wipat, A. (2006). Clustering and crosstalk in a yeast functional interaction network. *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. San Diego: IEEE Press.

[12] Ho, Y. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415: 180 - 184.

[13] http://www.genomicglossaries.com/content/omes.asp

[14] Huynen, M., Snel, B., Lathe, W. & Bork, P. (2000). Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Research* 10(8): 1204 - 1210.

[15] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the USA* 98(8): 4569 - 4574.

[16] Jaenisch, R. & Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics* 33: 245 - 254.

[17] Jenssen, T. -K., Laegrid, A., Komorowski, J. & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* 28: 21 - 28.

[18] Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28: 27 - 30.

[19] Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlins, C., Verrier, P. & Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* 22(11): 1383 - 1390.

[20] Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* 306(5701): 1555 - 1558.

[21] Li, J., Li, X., Su, H., Chen, H. & Galbraith, D. W. (2006). A framework of integrating gene relations from heterogeneous data sources: An experiment on Arabidopsis thaliana. *Bioinformatics* 22(16): 2037 - 2043.

[22] Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751 - 753.

[23] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* 298: 824 - 827.

[24] Oliver, S. (2000). Guilt-by-association goes global. *Nature* 403: 601 - 603.

[25] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the USA* 96(8): 4285 - 4288.

[26] Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002). Network motifs in the transcriptional network of *Escherichia coli. Nature Genetics* 31: 64 - 68.

[27] Stapley, B. & Benoit, G. (2000). Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pacific Symposium on Biocomputing* 5: 529 - 540.

[28] The Gene Ontology Consortium (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics* 25: 25 - 29.

[29] Tong, A. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294: 2364 - 2368.

[30] Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W. V., Fields, S., Boone, C. & Cesarini, G. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295: 321 - 324.

[31] Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W. V., Bussey, H., Andrews, B., Tyers, M. & Boone, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294: 2364 - 2368.

[32] Uetz, P., Glot, L., Cagney, G., Mansfield, T. A., Judson, R. S. & et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* 403: 623 - 631.

[33] Wernicke, S. & Rasche, F. (2006). FANMOD: A tool for fast network motif detection. *Bioinformatics* 22(9): 1152 - 1153.

[34] Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research* 11(3): 356 - 372.

[35] Wuchty, S., Oltvai, Z. N. & Barabasi, A. L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics* 35(2): 176 - 179.

[36] Xenarios, I., Rice, D. W., Salwinsky, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. (2000). DIP: The Database of Interacting Proteins. *Nucleic Acids Research* 28(1): 289 - 291.

[37] Yanai, I., Derti, A. & DeLisi, C. (2001). Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. *Proceedings of the National Academy of Sciences of the USA* 98(14): 7940 - 7945.