doi:10.1016/j.jmb.2007.03.013 J. Mol. Biol. (2007) 369, 1126–1139

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

The Identification of Similarities between Biological Networks: Application to the Metabolome and Interactome

Adrian P. Cootes[1], Stephen H. Muggleton[2] and Michael J.E. Sternberg[1]*

[1]*Division of Molecular Biosciences, Imperial College London, South Kensington London SW7 2AZ, UK*

[2]*Department of Computing Imperial College London South Kensington London SW7 2AZ, UK*

**Corresponding author*

The increasing interest in systems biology has resulted in extensive experimental data describing networks of interactions (or associations) between molecules in metabolism, protein–protein interactions and gene regulation. Comparative analysis of these networks is central to understanding biological systems. We report a novel method (PHUNKEE: Pairing subgrapHs Using NetworK Environment Equivalence) by which similar subgraphs in a pair of networks can be identified. Like other methods, PHUNKEE explicitly considers the graphical form of the data and allows for gaps. However, it is novel in that it includes information about the context of the subgraph within the adjacent network. We also explore a new approach to quantifying the statistical significance of matching subgraphs. We report similar subgraphs in metabolic pathways and in protein–protein interaction networks. The most similar metabolic subgraphs were generally found to occur in processes central to all life, such as purine, pyrimidine and amino acid metabolism. The most similar pairs of subgraphs found in the protein–protein interaction networks of *Drosophila melanogaster* and *Saccharomyces cerevisiae* also include central processes such as cell division but, interestingly, also include protein sub–networks involved in pre-mRNA processing. The inclusion of network context information in the comparison of protein interaction networks increased the number of similar subgraphs found consisting of proteins involved in the same functional process. This could have implications for the prediction of protein function.

Keywords: network comparison; metabolic networks; protein interactions; systems biology

Introduction

There is an increasing interest in how molecules combine in the cell to form complex biological systems. For many species, we now have an increasingly comprehensive picture of their genes and the molecules for which each of those genes code. Researchers have been building on this success of genome sequencing efforts by identifying interactions or functional associations between pairs of molecules or genes.[1] There is now a considerable amount of such data in the literature with more being rapidly generated *via* high-throughput experiments. The complex network of relationships involved in biological systems such as protein–protein interactions,[2] small-molecule metabolism,[3] and gene regulation,[4] can now be studied in increasing detail. There has been a considerable number of studies into general biological network properties, such as their scale-free nature.[5,6] However, the development of methods to compare biological networks has been quite limited.[7–14] This is in stark contrast to the variety of methods for the comparison of sequences and the number of ways used to assess the significance of a sequence match. Here, we introduce a novel method by which similar subgraphs in a pair of networks can be identified.

Most studies of biological networks have concentrated on comparing their connectivity properties to theoretical or other types of well-studied graphical systems.[5,6] However, the comparison of biological networks to one another will yield more evolutionary information than simply examining each net-

E-mail address of the corresponding author: m.sternberg@imperial.ac.uk

work in isolation. Furthermore, comparing the network contexts of nodes in different biological networks may reveal information about the entities represented by those nodes. For example, the function of proteins can be predicted from information in the surrounding interaction network.[15] Locating statistically significant similar subgraphs in different networks would give greater confidence to assigning function to uncharacterised proteins or interactions between proteins within those subgraphs.

Thus far, there have been a number of approaches to the comparison of biological networks. Several of these methods compare the general topological statistics of subgraphs[7] or compare the statistical prevalence of different types of three or four node connection patterns.[8] Particular types of small network structures have been shown to be over-represented in the transcription regulation networks of *Saccharomyces cerevisiae* and *Escherichia coli*, such as three node "feed forward loops" and four node "bi-fan" patterns (where two transcription factors each regulate the same two genes).[16] A more recent approach locates approximately matching small motifs.[12] Clearly, it would be desirable to be able to also consider the similarity in higher-order connectivity patterns in networks. The PATHBLAST algorithm was used to search for larger similar subgraphs by first comparing short linear "strings" of nodes within a pair of protein–protein networks,[9] and has since been extended to also search for similar "dense clusters" of interacting proteins (NetworkBlast).[13] Both approaches allowed for gaps in similar strings or clusters. These strings and clusters were then grouped to search for larger related subgraphs. In the initial study,[9] similar subgraphs (based on short strings) in the protein interaction networks of *S. cerevisiae* and *H. pylori* were found to be involved with protein synthesis and cell rescue, protein fate and targeting, cell envelope and nuclear transport, proteolytic degradation and rRNA transcription. The latter study[13] searched for similar subgraphs (based on both strings and clusters) in multiple species: *Drosophila melanogaster*, *S. cerevisiae* and *Caenorhabditis elegans*. The most prevalent processes found in related subgraphs in this case were protein degradation, RNA polyadenylation and splicing, protein phosphorylation and signal transduction. NetworkBlast bases its initial search on particular types of graphical form modelled on signalling pathways (strings) and protein complexes (clusters). The Græmlin algorithm extended this approach by searching efficiently for similar subgraphs of arbitrary topology across multiple networks.[14] Another recent approach searched for general graphical structures by clustering nodes according to their distance from one another in each network.[10] This technique was used to identify clusters of genes that code for enzymes that are also clustered in metabolic networks. This type of arrangement often implies that the expression of the enzymes involved in that part of metabolism are co-regulated. Hundreds of such enzyme clusters were identified for a range of species. The approach used to produce these clusters finds regions in each network with many similar nodes but does not consider explicitly whether those nodes are connected to one another (and other nodes) in a similar way. For many types of network system, one would like a method to consider the similarity of the edges, as well as the proximity of nodes, in comparing subgraphs. Such a method has been demonstrated for the comparison of metabolic networks.[11] Frequently occurring metabolic subgraphs were sought across a large number of species. The most similar subgraphs found participate in pyrimidine, glutamate and alanine and aspartate metabolism. However, this approach did not consider gaps in the network structure. Given that biological systems often have elements inserted or deleted in the course of evolution, it is important for a network comparison method to permit approximate matches between subgraphs.

Here, we have developed a method (PHUNKEE: Pairing subgrapHs Using NetworK Environment Equivalence†) for the comparison of biological networks that searches for similar subgraphs of general structure, allowing for gaps, and determines the statistical significance of the match. The similarity of subgraphs was determined by considering explicitly the similarity of the edges as well as the similarity of the nodes. Furthermore, we depart from previous work in considering the set of all edges adjoining nodes belonging to a subgraph (referred to as the network context (Figure 1)), rather than simply those edges connecting nodes within the subgraph. We examined whether considering network context improved subgraph comparison and applied this technique to the comparison of pairs of metabolic and protein–protein interaction networks. We also explored two different methods for calculating the statistical significance of similar subgraphs and the suitability of each for analysing various types of biological networks. In this way, we identified similar subgraphs in pairs of metabolic networks and pairs of protein interaction networks for a number of different species. As expected, similar subgraphs were found in biological processes central to all life, such as amino acid, purine and pyrimidine metabolism, but also in processes not expected to be as strongly conserved, such as pre-mRNA processing.

Our Approach

A brief description of PHUNKEE (Figure 2) is given here and a more detailed description is given in Materials and Methods.

PHUNKEE consists of two basic steps. First, corresponding (shared) nodes and edges were identified. Second, the most similar network regions of a given, user-defined size were sought, allowing

† A server based on PHUNKEE can be found at www.sbg.bio.ic.ac.uk/phunkee

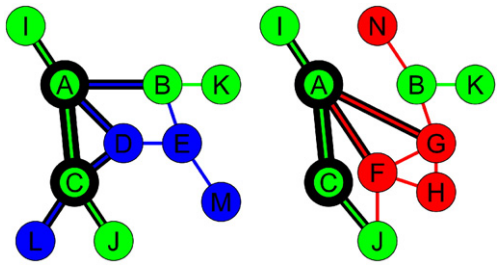


Figure 1. Network context and the similarity of subgraphs. A pair of networks from two species (one coloured red, the other blue, features common to both species are coloured green) are represented. Nodes (circles) represent a biological entity, such as a protein or an enzyme function. Edges (lines) joining nodes represent an interaction or association between those entities. Corresponding nodes are indicated in green and have the same label. Green edges are shared between the two networks. Blue and red nodes and edges are not shared between the species. Here, the similarity of subgraphs containing nodes AC is shown. When determining the similarity of the subgraphs, all edges adjoining these nodes (edges with a thick outline) are considered. This set of edges is referred to as the network context. Internal edges (thick outline) and external edges (thin outline) are weighted differently. An internal edge is any edge connecting nodes that are both members of the subgraph (the A-C edges in this Figure). An external edge is any edge connecting a subgraph node and another node that is not a member of the subgraph (for example, the A-I and C-J edges in this Figure). Internal and external edges are given weights w_i and w_e respectively. The similarity of the subgraphs is given by the shared-edge ratio. The shared-edge ratio is the weighted proportion of highlighted edges that are shared between species. In this case, there are two internal edges (two shared) and ten external edges (four shared). In the simplest scenario, where $w_i = w_e = 1$, the resulting shared-edge ratio is 6/12 (0.5).

for gaps and considering explicitly the similarity of edges as well as nodes.

In order to find shared nodes, we first considered relationships between nodes in different networks. For some types of biological network, there is a clear one-to-one correspondence between nodes. For example, a node that represents a particular compound in the metabolic network of a given species will correspond to the node representing that compound in another species. However, other types of biological networks have a many-to-many relationship between nodes. For instance, a protein in an interaction network may be similar in sequence to many proteins in another network. In this study, we employed two different methods to determine corresponding nodes in a pair of protein interaction networks. Firstly, we applied a graph-matching scheme (known as the Hungarian method[17]) to find the optimal correspondence between nodes according to their sequence similarity. A drawback of this method was that a small proportion of corresponding node pairs were found to have different functions (according to the COG database).[18,19] An alternative scheme determined corresponding nodes by finding the most sequence-similar proteins with the same function (known in this study as the COG function-matching method). This approach found a smaller number of corresponding node pairs but ensured that they had the same function. For both of these approaches, nodes with a corresponding node in the other species were deemed to be shared nodes. Edges of the same type (for example, representing the same metabolic function)

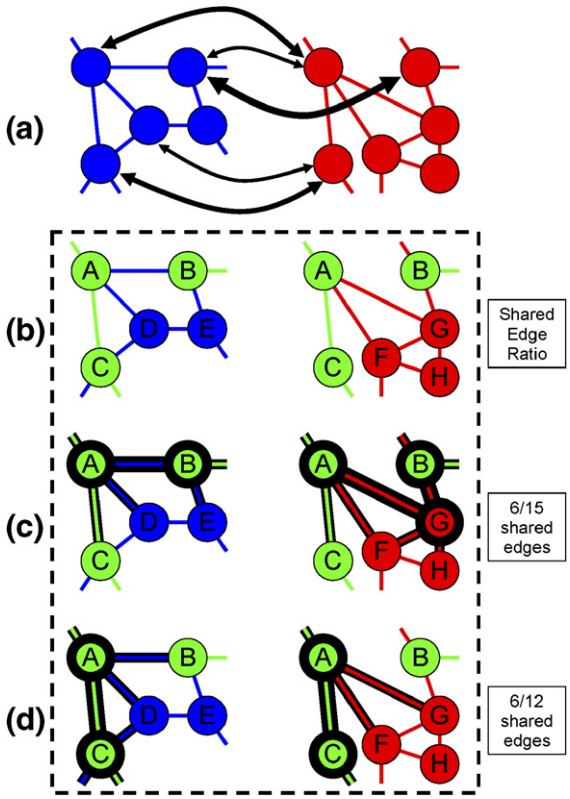


Figure 2. Locating similar subgraphs in a pair of networks. A pair of networks from two species are represented in a fashion similar to that in Figure 1. In general, there may be many relationships of different weights (arrows) between nodes in the different networks (a) (for example, protein interaction networks). For pairs of networks where there is already a one-to-one node correspondence between nodes (for example, identical compounds in metabolic networks), then only the steps contained within dotted lines (b), (c) and (d) are necessary. For the general case, a matching scheme is used to assign an optimal one-to-one correspondence between nodes (b). Nodes are then grouped so as to optimise the shared-edge ratio. All shared nodes within a user-defined radius are considered for grouping. For simplicity, let us give external and internal edges each a weight of 1 ($w_i = w_e = 1$). In (c), the shared node A is grouped with shared node B *via* the shortest path in each species (including the unshared node G). In this case, the shared-edge ratio is 6/15 (0.4). However, if A is grouped with C (d), the shared-edge ratio is 6/12 (0.5). The ratio is higher in the latter case; therefore, the AC subgraph will be selected and the grouping of a further node will be conducted in a fashion similar to that above. This process of grouping nodes is continued until the size of the subgraph reaches a limit set by the user.

connecting the same pair of shared nodes in both species were deemed to be shared edges.

PHUNKEE then searched for pairs of subgraphs with similar network contexts (see Figure 1). The network context of a subgraph is the set of all edges that adjoin all nodes that belong to that subgraph. The similarity in network context of a pair of subgraphs was measured using the shared-edge ratio. The shared-edge ratio was the weighted proportion of the adjoining edges of the subgraph pair that were shared. Edges were weighted by w_i if they were "internal" edges (an edge connecting two nodes both belonging to the subgraph) and by w_e if they were "external" edges (an edge connecting a node belonging to the subgraph to one that does not). Weights w_i and w_e were defined by the user. By considering the similarity of the network context, subgraph pairs with equally similar internal connections can be distinguished on the basis of their connections to the rest of the network. This approach differs from simply considering external nodes part of a larger subgraph because connections between external nodes are not considered and connections between internal and external nodes may be weighted differently to internal connections. A pair of subgraphs with similar internal connections and similar connections to its external neighbours may have quite different connections between those neighbours. Including external edges also has the advantage of automatically introducing a gap penalty. The inclusion of an unshared node will increase the number of unshared edges and hence lower the shared-edge ratio.

Each pair of corresponding nodes acted in turn as the starting point for a subgraph pair search. Other corresponding node pairs were grouped progressively in such a way as to maximise the shared-edge ratio of the resulting subgraph pair until one (or both) reached a user-defined maximum size.

The standard method (for example, see NetworkBlast[13]) of calculating the statistical significance of the similarity of a subgraph pair is by comparison to similar subgraphs found by the same algorithm in a pair of randomised networks. The randomised networks have the same connectivity as the networks of interest but have all node and edge labels reassigned randomly. We refer to this measure of significance as the "global" significance of the subgraph match (see Materials and Methods for more details). However, in pairs of biological networks that are known to be very similar overall, very many subgraphs will be significant according to this test. In this case, we may be more interested to know whether local network regions are significantly more similar than regions within the networks are generally. In order to do this, we developed a novel measure of significance that we label the "local" significance of subgraph match. The local statistical significance of the subgraph pair similarity was determined by generating subgraph pairs randomly from the same pair of networks without permuting the connectivity or the identities of the nodes and edges. In this way, the subgraph pair similarity will be determined to be locally significant if it is much greater than the overall similarity of the networks. This approach is appropriate for assessing the significance of local variations and allows for the peculiar connectivity properties of biological networks.

Results for Metabolic Networks

Using the approach outlined in the previous section, we first compared the metabolic networks of four organisms (*E. coli*, *S. cerevisiae*, *Helicobacter pylori* and *Homo sapiens*). Nodes in each network represented a metabolite. Each edge linking a pair of metabolites represented an enzyme that catalyses a reaction involving those compounds. In this study, we searched for subgraphs of similar function rather than similar evolutionary history. Unrelated enzymes may exist in the two species that nevertheless perform the same function. Hence, each edge was labelled with the Enzyme Classification (EC) number of the enzyme. Each pair of metabolite nodes may be linked by a number of edges labelled with different enzyme functions.

We searched for similar subgraphs consisting of ten shared nodes for each possible pairwise network comparison of the four organisms listed above.

Statistical significance

First, the global statistical significance of resulting subgraph pairs was examined. In this case, all subgraph pairs found for each possible pair of species were deemed to be significant (data not shown). The shared-edge ratios of subgraph pairs were often very high for metabolic networks (sometimes close to 1.0), probably due to many enzymes in KEGG being assigned to reactions *via* comparison to model species. However, this result was also due to the generated random networks (with randomised node and edge labels) used to measure global significance having few, if any, shared edges. This was probably due to the specificity of the edge labels (enzymatic functions) in metabolic networks. In other types of networks with unlabelled edges (for example, protein interaction networks), edges are shared if they simply connect pairs of nodes with the same label. For metabolic networks, we are also interested in the types of enzyme that catalyse reactions between metabolites, and we label each edge accordingly. In this case, shared edges have to meet the additional constraint of having identical edge labels. Clearly, this constraint makes it less likely that shared edges will be found in randomised networks.

The metabolic networks compared here are clearly very similar, and all subgraphs found were deemed to be globally significant. However, the similarity between the networks was not necessarily uniform. To test this, and to determine which subgraph pairs (if any) were significantly more similar than the networks were generally, we employed the test of

local significance (described previously) in the following analysis of metabolic networks.

The influence of external edges

To examine the effect of including network context in assessing subgraph similarity, external edge weights of $w_e=0$ and $w_e=0.1$ were used in two separate searches. When discounting external edges ($w_e=0$), no locally significant similar subgraphs of ten shared nodes could be found for any pair of species considered. However, the inclusion of external edges ($w_e=0.1$) revealed many locally significant similar subgraphs for each species pair. For the comparison of *S. cerevisiae* and *E. coli*, 205 of 617 (33%) searches resulted in subgraph pairs with a p value of <0.01. Many of these resulting 205 subgraph pairs overlapped to some extent or were identical. Therefore, we considered the number of non-redundant significant subgraph pairs (those with <25% of nodes in common with other significant subgraphs). Eleven such non-redundant similar subgraphs were found for *S. cerevisiae* and *E. coli*. One of the subgraph pairs with the highest shared-edge ratio (1.0) involved in purine metabolism can be seen in Figure 3. For the comparison of *E. coli* and *H. pylori*, only 65 of 371 (18%) pairs had p values <0.01, four of which were non-redundant. This was despite the subgraph pairs having relatively high shared-edge ratios (the highest being 0.99). Thus, the networks were very similar overall but had few regions with unusually high levels of local similarity. Over all species comparisons, 18–33% of searches resulted in locally significant similar subgraphs with p values of <0.01 when accounting for edges external to the subgraphs. These searches revealed between four and 11 non-redundant significant similar subgraphs. Thus, the comparison of metabolic networks was enhanced by considering the external network context of a pair of subgraphs as well as their internal connectivities.

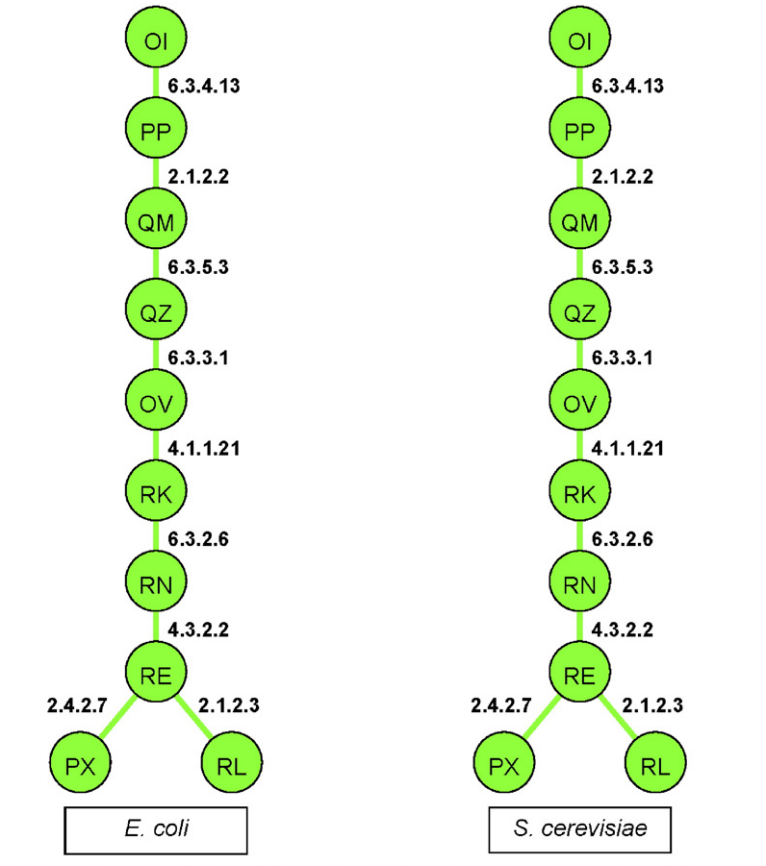


Figure 3. Similar subgraphs from a comparison of *E. coli* and *S. cerevisiae* metabolic networks. Nodes correspond to metabolites and edges correspond to enzymes that catalyse reactions involving those metabolites. The name of the metabolite corresponding to each node label is listed. The edge labels give the EC numbers of the functions. Only edges connecting nodes belonging to the subgraph are shown. This was one of the most similar subgraph pairs found in *E. coli* and *S. cerevisiae* with ten shared nodes, $w_e=0.1$ and $w_i=0.1$. The subgraph pair had a shared-edge ratio of 1.00 and a p value of 1.66×10^{-4}. The subgraphs correspond to part of the purine metabolism pathway.

The biological function of resulting subgraphs

Metabolism is often discussed in terms of pathways of enzymatic reactions that correspond to key biochemical processes (for example, purine metabolism). We examined the pathways (as defined by KEGG) in which the most similar subgraphs of functions found above participate (Table 1). The KEGG pathways are defined manually; however, they give a general indication of which overall biological processes a given set of enzymes are involved in. Predominantly, the most similar subgraph pairs consist of functions involved in purine, pyrimidine and amino acid metabolism pathways. This is perhaps unsurprising, given the importance of these processes to all organisms. In general, the subgraph pairs located did not always neatly correspond to part of a single KEGG pathway. Surprisingly, when comparing *S. cerevisiae* and *E. coli*, some of the most similar subgraphs were involved with porphyrin metabolism.

Table 1. Biochemical pathway composition of the most similar ten-node subgraphs in pairs of species

Species compared	Subgraph number	KEGG pathways
S. cerevisiae and *E. coli*	1	Purine metabolism
	2	Cyanoamino acid metabolism
		Taurine and hypotaurine metabolism
		Cysteine metabolism
	3	Porphyrin and chlorophyll metabolism
E. coli and *H. pylori*	1	Phenylalanine, tyrosine and tryptophan biosynthesis
		Cysteine metabolism
		Aminoacyl-tRNA biosynthesis
		Glycine, serine and threonine metabolism
		Sulfur metabolism
E. coli and *H. sapiens*	1	Pyrimidine metabolism
	2	Purine metabolism
H. sapiens and *H. pylori*	1	Selenoamino acid metabolism
		Aminoacyl-tRNA biosynthesis
		Methionine metabolism
S. cerevisiae and *H. pylori*	1	Phenylalanine, tyrosine and tryptophan biosynthesis
		Tryptophan metabolism
		Aminoacyl-tRNA biosynthesis
S. cerevisiae and *H. sapiens*	1	Purine metabolism
	2	Pyrimidine metabolism
		Fructose and mannose metabolism
		Purine metabolism
	3	Selenoamino acid metabolism
		Aminoacyl-tRNA biosynthesis
		Methionine metabolism

The subgraph pairs listed were those with the highest shared-edge ratio for a given pair of species using edge weights $w_i = 1$ and $w_e = 0.1$. Each subgraph pair consisted of 20 shared nodes and no unshared node in total. For several of the species comparisons, more than one subgraph pair had the equal highest shared-edge ratio. For example, three types of subgraph pair had equal highest score for *S. cerevisiae* and *E. coli*. Subgraph pairs are listed with their overlapping KEGG pathways. Each compound may belong to a number of pathways. Only those pathways with at least one pair of connected nodes within the subgraph pair are listed. All subgraph pairs represented here had ("local" significance) p values < 0.05.

Results for Protein–protein Interaction Networks

Protein–protein interaction networks were compared for *D. melanogaster* and *S. cerevisiae*. The network nodes represented proteins and edges represented physical interactions between pairs of proteins. The interaction networks compared here are the largest in the DIP[20] database and contained 7066 proteins for *D. melanogaster* and 4733 proteins for *S. cerevisiae*. This represents significant coverage of the *D. melanogaster*[21] and *S. cerevisiae*[22] genomes (13854 and 5749 protein-coding genes respectively).

Node matching using the Hungarian method

In the first comparison of these networks, we determined corresponding nodes using the Hungarian method. A total of 1743 corresponding node pairs was found for the *D. melanogaster* and *S. cerevisiae* interaction networks. Of the 1057 of these pairs with both proteins function defined by COG,[18,19] 871 (82%) had the same function. Incorporating the sequence similarity of the network neighbours of prospective corresponding nodes yielded no significant increase in this figure (data not shown). Searches were then conducted for subgraph pairs with a maximum of four shared and four unshared nodes with internal edge weight $w_i = 1$ and external edge weights $w_e = 0$, 0.1, 0.5 and 1.0. Searching for relatively small subgraphs with up to four unshared nodes (gaps) ensured that a reasonable number of significant subgraphs were found. We examined whether the external network context influenced the composition of protein functions within the subgraphs. In particular, we tested whether the subgraphs found contained proteins involved in the same biological processes. In order to do this, the function of each protein in the *D. melanogaster* and *S. cerevisiae* interaction networks was taken from the COG database.[18,19] Each specific COG function also belongs to at least one of 23 more general functional classes (for example, RNA processing and modification). Here, we defined significant subgraph pairs with >50% of nodes belonging to the same COG functional class to be functionally consistent. For each subgraph search, we determined the number of functionally consistent subgraph pairs found (Table 2). Given that subgraphs located in different searches often overlap, we considered also the number of non-redundant, functionally consistent subgraph pairs (those with <25% of nodes in common with other functionally consistent subgraphs). We also determined the number of nodes in each interaction network that belonged to at least one non-redundant, functionally consistent subgraph pair.

The influence of external network context

Clearly, a greater number of functionally consistent subgraph pairs are found with an external edge

Table 2. The influence of external edges on determining similar subgraphs

w_e	Measure of statistical significance	Number of significant functionally consistent similar subgraphs	Number of non-redundant significant functionally-consistent similar subgraphs	Number of nodes found in non-redundant significant functionally-consistent similar subgraphs / COG functional classes
0	Local	125	23	225 / 10
	Global	108	15	137 / 8
0.1	Local	161	22	226 / 10
	Global	155	20	202 / 9
0.5	Local	164	17	182 / 7
	Global	164	17	182 / 7
1	Local	115	11	125 / 6
	Global	117	13	150 / 6
All	Local	557	28	276 / 9
	Global	532	26	258 / 9

The number of significant similar subgraphs found and the function composition of those subgraphs varied with the relative values of external edge weight w_e to internal edge weight w_i. Similar subgraphs were sought in the *D. melanogaster* and *S. cerevisiae* protein interaction networks with a maximum of four shared and four unshared nodes with $w_i=1$ and $w_e=0$, 0.1, 0.5 and 1. The individual searches and the combined results (labelled All) are shown here. Corresponding nodes were determined using the Hungarian method. Listed here are the total number of functionally consistent significant similar subgraphs found and the number of non-redundant functionally consistent similar subgraphs. "Global" and "local" significance was determined as described in Materials and Methods. Subgraph pairs were deemed to be functionally consistent if >50% of node function classes (for example, RNA processing and modification) were common. Also listed are the number of different nodes and the number of COG functional classes covered by the non-redundant functionally consistent similar subgraphs.

weight $w_e=0.1$ than when external edges are not considered ($w_e=0$). However, this is not the case when considering only non-redundant, functionally consistent subgraphs. The number of significant non-redundant, functionally consistent subgraphs was similar for the global and local measures of significance overall. There was a small increase in the number of non-redundant, functionally consistent subgraphs for $w_e=0.1$ of global significance. The number of functionally consistent subgraph pairs decreases when $w_e=0.5$ and $w_e=1$, probably because external edges dominate the subgraph comparison (the subgraph pairs studied here typically had an order of magnitude more external than internal edges).

Combining the subgraph searches with different external edge weights (as described in Materials and Methods) resulted in a higher number of significant subgraph pairs than the subgraph search considering only internal edges ($w_e=0$). While subgraph searches conducted with $w_e>0$ did not each find larger numbers of significant subgraph pairs, they clearly found subgraphs different from the subgraph search considering only internal edges. Clearly, considering network context locates significant functionally consistent subgraphs that would otherwise be missed.

The above results show that including the network context of subgraphs locates more meaningful function units within interaction networks than comparing only internal interactions. This has clear implications for the prediction of protein function. Recent work has shown that a protein found in a pair of similar subgraphs, containing a large number of other proteins performing a given function, can be inferred to also perform that function.[13] The increased number of functionally consistent subgraphs found by considering $w_e>0$ suggests that considering the external network context in subgraph comparison may assist in predicting protein function from interaction networks.

The shared-edge ratios for the protein interaction subgraph pairs found here were much lower than those found for metabolic networks. The highest shared-edge ratio for subgraph pairs found with external edge weight $w_e=0.1$ was 0.34. The low number of shared edges in the pair of protein–protein interaction networks is perhaps due to relatively poor data.[23,24] This has been observed also in a previous network comparison study.[9] Despite noisy data, this study and previous studies[7–13] have demonstrated significant similarities between protein interaction networks. Figure 4 shows four of the most similar functionally consistent subgraph pairs in the *D. melanogaster* and *S. cerevisiae* protein–protein interaction networks (from four different function classes) found in this study. The function of each protein represented in Figure 4 is given in Table 3. The proteins in the most significant subgraph pair are largely involved in pre-mRNA processing[25,26] (Figure 4(a)). This is perhaps a surprising result, given that *S. cerevisiae* genes have relatively few introns compared to *D. melanogaster*. Hence, it might be expected that the protein interactions within the pre-mRNA splicing machinery would be quite different in these two species relative to other parts of their interaction networks. However, RNA processing subgraphs have also been found in a previous network comparison study.[13] A number of other processes were found associated with functionally consistent subgraph pairs. For example, Figure 4(b) shows a pair of subgraphs containing proteins involved in cell division. Specifically, these proteins are part of the mitotic checkpoint, which ensures that chromosomes segregate properly during mitosis. Importantly, a method that allowed for gaps was required to find this subgraph pair. Figure 4(c) and (d) show similar

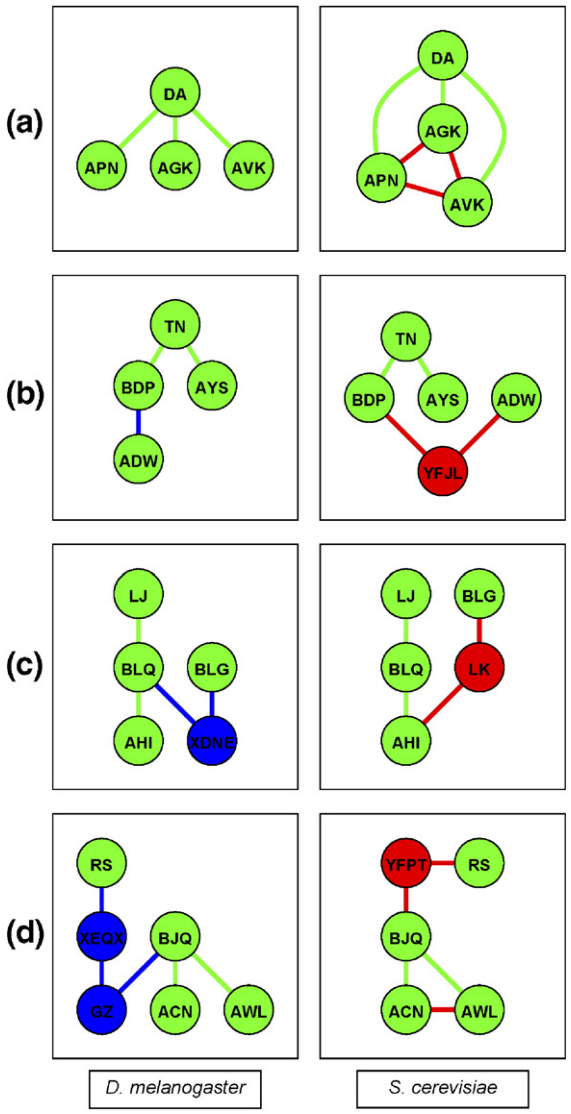


Figure 4. Four pairs of similar subgraphs from a comparison of *D. melanogaster* and *S. cerevisiae* protein–protein interaction networks. Nodes correspond to proteins, and edges correspond to physical interactions between proteins in each species. Only interactions between the subgraph members are shown. Green nodes and edges are common to subgraphs in both species. Blue features in the *D. melanogaster* network are not found in the *S. cerevisiae* subgraph. Red features in the *S. cerevisiae* network are not found in the *D. melanogaster* subgraph. Corresponding nodes have the same label. The function of each protein (according to COG) is listed in Table 3. Four of the most similar subgraph pairs found in the *D. melanogaster* and *S. cerevisiae* protein–protein interaction networks with corresponding nodes determined by the Hungarian method are shown. Subgraph pairs containing four shared nodes each were found using parameters $w_e = 0.1$ and $w_i = 1$. (a) A pair of subgraphs involved in RNA processing (with a shared-edge ratio of 0.34). (b) A pair of subgraphs involved in cell division (with a shared-edge ratio of 0.22). The pair of subgraphs in (c) represent chaperones (with a shared-edge ratio of 0.22) and (d) contains a pair of subgraphs involved in DNA repair (with a shared-edge ratio of 0.20). All subgraph pairs shown here had a p value $< 5.9 \times 10^{-5}$.

subgraphs involved with chaperone activity and DNA repair, respectively. Both of these systems are clearly important to these organisms, Hsp90 is an essential chaperone for eukaryotes and base excision repair is required to remove incorrect or damaged nucleotides from DNA. The subgraph pairs shown in Figure 4(a) and (b) each have 100% of nodes, and the pairs in Figure 4(c) and (d) each have 60% of nodes, within a single COG functional class. In all, functionally consistent subgraph pairs were found corresponding to sections of nine different COG functional classes.

Node matching using COG functions

One possible impediment to finding similar subgraphs in protein interaction networks is the incorrect assignment of corresponding nodes. In order to ascertain the impact of incorrect assignments, we determined corresponding nodes using the COG function matching method described previously. While fewer correspondences could be determined in this way, it ensured that all corresponding nodes had the same functions. To the authors' knowledge, this approach to evaluating node matching in network comparison is novel.

A total of 951 corresponding node pairs were found for the *D. melanogaster* and *S. cerevisiae* interaction networks using the COG function matching method. As before, searches were conducted for subgraph pairs with a maximum of four shared and four unshared nodes with internal edge weight $w_i = 1$ and external edge weights $w_e = 0$, 0.1, 0.5 and 1.0. The number of significant and functionally consistent subgraph pairs found showed a similar dependence on w_e to that found previously. That is, a value of $w_e = 0.1$ resulted in the largest number of significant and functionally consistent subgraph pairs according to the global measure of significance (data not shown). The number of non-redundant, functionally consistent subgraphs in this case was 16, covering a total of 168 different nodes in *D. melanogaster* and *S. cerevisiae*, fewer than that found with the Hungarian method of node matching (Table 2). The subgraph pairs found were often similar to those found with the Hungarian method. The highest scoring functionally consistent subgraph pair was identical with that found for the Hungarian approach (Figure 4(a)), involved with pre-mRNA splicing. Subgraph pairs representing Hsp90 chaperones and DNA repair proteins were similar to those found with the Hungarian method (Figure 4(c) and 4(d), respectively). However, there were some important differences. For example, a subgraph pair representing SNARE and SNAP proteins involved with vesicular transport was found to be significant using this approach. Also, no functionally consistent subgraph pair was found representing processes involved in cell division, unlike the Hungarian method (Figure 4(b)). This was probably because the subgraph pairs found with the Hungarian method had two protein pairs with the same function, of which only one could be

Table 3. Functions of proteins in Figure 4

	D. melanogaster		*S. cerevisiae*	
	Symbol	COG class	Symbol	COG class
(a)	DA	U6 snRNA-associated Sm-like protein	DA	U6 snRNA-associated Sm-like protein
	APN	Small nuclear ribonucleoprotein (snRNP) Sm core protein	APN	Small nuclear ribonucleoprotein (snRNP) Sm core protein
	AGK	Small nuclear ribonucleoprotein F	AGK	Small nuclear ribonucleoprotein F
	AVK	Small Nuclear ribonucleoprotein splicing factor	AVK	Small Nuclear ribonucleoprotein splicing factor
(b)	TN	Mitotic spindle checkpoint protein BUB3, WD repeat superfamily	TN	Mitotic spindle checkpoint protein BUB3, WD repeat superfamily
	BDP	Mitotic checkpoint serine/threonine protein kinase	BDP	Mitotic checkpoint serine/threonine protein kinase
	AYS	Mitotic checkpoint serine/threonine protein kinase	AYS	Mitotic checkpoint serine/threonine protein kinase
	ADW	Cyclin B and related kinase-activating proteins	ADW	Cyclin B and related kinase-activating proteins
			YFJL	Anaphase promoting complex, Cdc20, Cdh1, and Ama1 subunits
(c)	LJ	Hsp90 co-chaperone CNS1 (contains TPR repeats)	LJ	Hsp90 co-chaperone CNS1 (contains TPR repeats)
	BLQ	Molecular chaperone (HSP90 family)	BLQ	Molecular chaperone (HSP90 family)
	AHI	Molecular co-chaperone STI1	AHI	Molecular co-chaperone STI1
	BLG	Mannosyltransferase	BLG	Mannosyltransferase
	XDNE	–	LK	Ubiquitin activating enzyme UBA1
(d)	RS	–	RS	Uncharacterised conserved protein, contains WD40 repeats
	BJQ	Structure-specific endonuclease ERCC1-XPF, ERCC1 component	BJQ	Structure-specific endonuclease ERCC1-XPF, ERCC1 component
	CAN	DNA excision repair protein XPA/XPAC/RAD14	ACN	DNA excision repair protein XPA/XPAC/RAD14
	AWL	Structure-specific endonuclease ERCC1-XPF, catalytic component XPF/ERCC4	AWL	Structure-specific endonuclease ERCC1-XPF, catalytic component XPF/ERCC4
	XEQX	Uncharacterised conserved protein H4	YFPT	Nucleotide excision repair complex XPC-HR23B, subunit XPC/DPB11
	GZ	Protein kinase PCTAIRE and related kinases		

The function of each protein represented in Figure 4 is listed here. For (a)–(d), the symbols in Figure 4 in each species are listed next to their function according to the COG database. Dashes indicate that COG lists no function for that protein.

matched with the COG function-matching method. Overall, using existing function data produces a smaller number of functionally consistent similar subgraphs than using sequence similarity alone for the systems studied here. Also, this approach does not necessarily remove all ambiguities and, as function data are not available for all proteins, will ignore some network similarities in general. It is anticipated that more sophisticated node-matching techniques, incorporating information from the local network environment as well as the node pair of interest, will improve subgraph similarity matching. The techniques used and discussed in this study, however, represent an important first step.

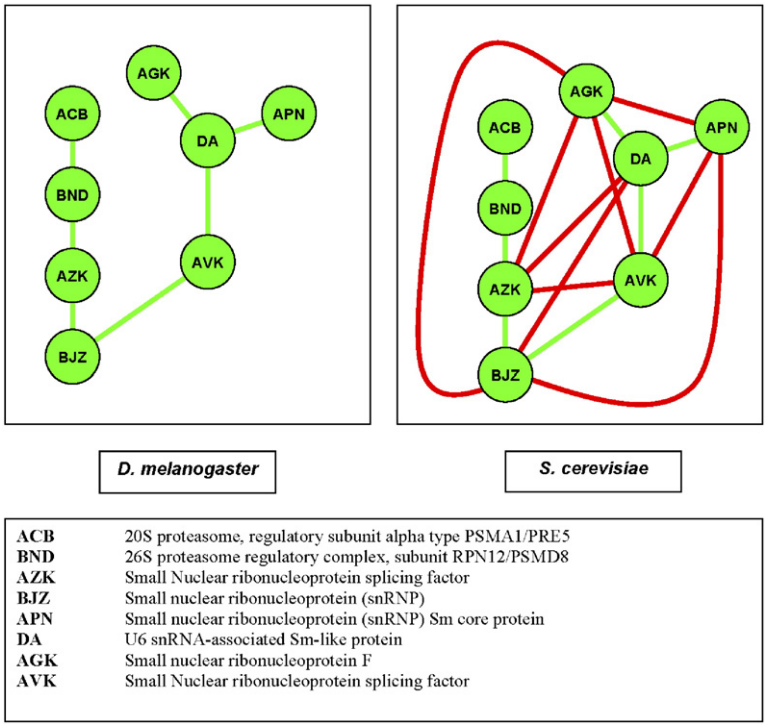
Larger subgraphs

While PHUNKEE generalises to searching for larger subgraphs in protein interaction networks, few functionally consistent subgraphs are found. Similar subgraph pairs of eight shared nodes describing part of the pre-mRNA splicing complex (Figure 5) were one of the few larger functional systems found consistently. As more accurate data become available, PHUNKEE could be used to find larger similar subgraphs within protein interaction networks.

Comparison to NetworkBlast

We compared the results of PHUNKEE to that of the publicly available NetworkBlast software.[13] NetworkBlast locates similar short strings (paths) of nodes in a pair of networks as well as similar densely interacting clusters. We applied NetworkBlast to the comparison of the *D. melanogaster* and *S. cerevisiae* protein interaction networks and identified significant functionally consistent paths and clusters. To assess the number of independent subgraphs found and network coverage, redundant paths and clusters were removed in a fashion similar to that described above for the PHUNKEE method.

NetworkBlast identified two significant functionally consistent clusters and 26 functionally consistent paths (four of these paths are shown in Figure 6 (Table 4)). Of these, both functionally consistent clusters and 13 of the functionally consistent paths were considered to be non-redundant. Overall, the non-redundant, functionally consistent paths and clusters belonged to seven different COG functional classes, two less than the four node subgraphs found by PHUNKEE. Unlike NetworkBlast, PHUNKEE identified similar subgraphs involved with transcription, the cytoskeleton and replication, recombination and repair. However, NetworkBlast identified paths involved with intracellular trafficking and PHUNKEE did not. Similar paths and clusters found by NetworkBlast often overlapped with subgraphs identified by PHUNKEE. NetworkBlast paths and clusters covered 134 nodes in total, which is fewer than those found by PHUNKEE (Table 2). However, it should be noted that NetworkBlast does not list nodes that form part of the gaps in identified paths and clusters,



ACB	20S proteasome, regulatory subunit alpha type PSMA1/PRE5
BND	26S proteasome regulatory complex, subunit RPN12/PSMD8
AZK	Small Nuclear ribonucleoprotein splicing factor
BJZ	Small nuclear ribonucleoprotein (snRNP)
APN	Small nuclear ribonucleoprotein (snRNP) Sm core protein
DA	U6 snRNA-associated Sm-like protein
AGK	Small nuclear ribonucleoprotein F
AVK	Small Nuclear ribonucleoprotein splicing factor

Figure 5. A pair of similar subgraphs from a comparison of *D. melanogaster* and *S. cerevisiae* protein–protein interaction networks. Subgraphs are represented as in Figure 4. The function of each protein (according to COG) is listed. One of the most similar subgraph pairs found in the *D. melanogaster* and *S. cerevisiae* protein–protein interaction networks with corresponding nodes determined by the COG function matching method is shown. Subgraph pairs containing eight shared nodes each were found using parameters $w_e = 0.1$ and $w_i = 1$. The pair of subgraphs is involved in RNA processing (with a shared-edge ratio of 0.30 and a p value of $< 5.9 \times 10^{-5}$).

and hence were not included in the calculated number of nodes covered. Also, the clusters identified by NetworkBlast were considerably bigger than the subgraphs found by PHUNKEE, the larger of the subgraphs being 15 nodes for both non-redundant, functionally consistent clusters. It is clearly more difficult to find larger functionally consistent subgraphs. The two functionally consistent clusters found were part of the pre-mRNA splicing complex and part of the proteasome regulatory complex. These overlap, to some extent, with two of the larger eight node subgraphs identified by PHUNKEE, but PHUNKEE did not find functionally consistent subgraphs with 15 shared nodes.

The reason that PHUNKEE finds more subgraphs of size four nodes than NetworkBlast finds paths of length four nodes is probably due to the nature of the respective searches. Any significant subgraph that is branched in its structure (such as those shown in Figure 4(a)–(c)) cannot be identified as a significant path by NetworkBlast. They may, however, be identified as part of a significant densely interacting cluster or by subsequent further clustering of paths. By also considering the similarity of a subgraphs external network context, PHUNKEE can find additional similar subgraphs that NetworkBlast may not. However, NetworkBlast finds similar subgraphs missed by PHUNKEE and is able to find larger subgraphs. This is possibly because NetworkBlast sifts through all possible matches between sequence-similar proteins as it locates similar paths and clusters. PHUNKEE, however, first uses the Hungarian algorithm to allocate one-to-one node matches that form the basis of the subsequent search for similar subgraphs. Correct node matches may be missed if they are not part of the initial sequence-based global node matching. Incorporating all possible node matches during the subgraph search in a way similar to NetworkBlast could improve subsequent versions of the PHUNKEE algorithm. Clearly, the NetworkBlast and PHUNKEE algorithms each have advantageous features relative to one another, and offer complementary approaches to biological network comparison. To demonstrate this point, the four ($w_e = 0$, 0.1, 0.5, 1) PHUNKEE four node similar subgraph searches conducted before were combined with the NetworkBlast similar four node path search described above. Using the same approach to determining significance as the combination of PHUNKEE subgraph searches discussed previously (see Materials and Methods), the combined PHUNKEE/NetworkBlast subgraph searches yielded a total of 30 non-redundant, functionally consistent subgraph pairs (with either global or local significance measures used for PHUNKEE). Thus, the combination of PHUNKEE and NetworkBlast (paths only) resulted in a larger number of similar subgraphs than either approach in isolation.

Conclusion

Given the increasing amount of biological network data being generated, automated methods of locating similarities in networks are going to become increasingly useful. Here, we have constructed a novel (and general) network comparison algorithm and introduced the new concept of network context. PHUNKEE is able to find significant matches between subgraphs in protein–protein interaction and metabolic networks. Their similar network context implies that these processes have been relatively well maintained during the course of

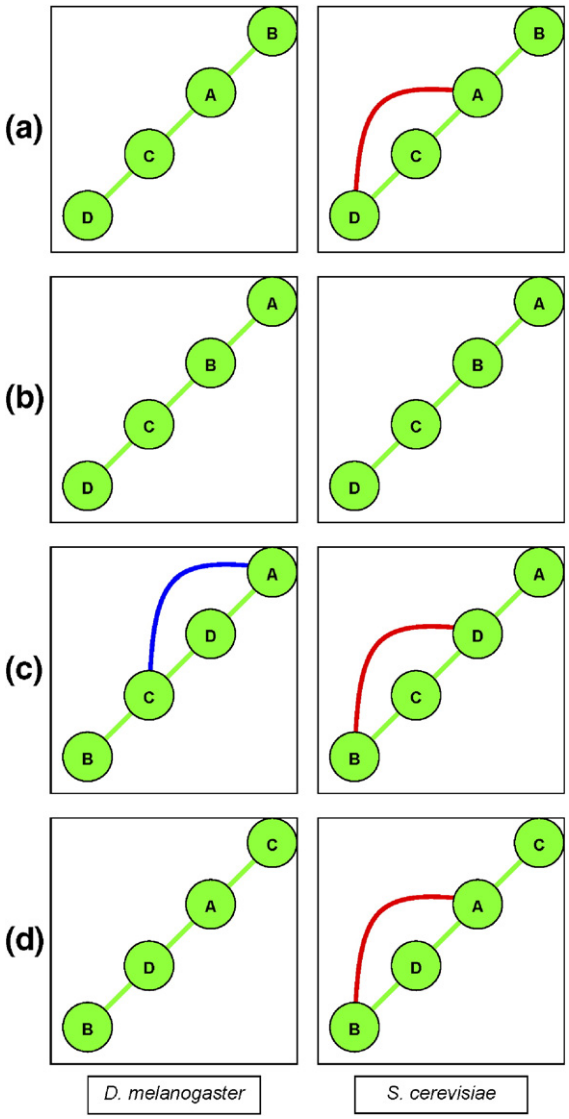


Figure 6. Four pairs of similar paths found by NetworkBlast from a comparison of *D. melanogaster* and *S. cerevisiae* protein–protein interaction networks. Subgraphs are represented as in Figure 4. The function of each protein (according to COG) is listed in Table 4. All paths presented were significantly similar ($p < 0.01$). Paths were found in such processes as signal transduction (a), cell division (b), RNA processing (c) and vesicular transport (d).

evolution and that the relationship of these subgraphs with the rest of the network have also remained relatively unaltered. The comparison of network context, rather than of internal edges only, aids the discrimination of significant similar subgraphs. By including information from external edges, PHUNKEE can locate more subgraph pairs with constituent nodes involved in the same functional processes. This has the potential to assist the prediction of protein function. In addition, we found that PHUNKEE could locate significant functionally consistent subgraphs that are not identified by NetworkBlast, a popular network comparison method.

Materials and Methods

Metabolic network data

Metabolic network data wee taken from a relational database, developed in a previous study,[27] which draws on information from the KEGG database.[28,29] Small-molecule metabolism of each species was modelled as a network of metabolites. Each node in the network was a metabolite and each edge was an enzyme function (defined by its EC number) of one or more enzymes in that species. An edge connecting two metabolites represents an enzyme that catalyses a reaction with one of the metabolites as a reactant and the other as a product. Molecules that participate in a very large number of reactions (for example, H_2O) were removed from the networks.

Protein–protein interaction data

For protein–protein interaction networks, the nodes represented proteins, and edges represented physical interactions between pairs of proteins, as determined by experiment. The correspondences between proteins in different networks were determined using sequence similarity relationships, calculated by the program BLAST,[30,31] or using sequence similarity in conjunction with function information taken from the COG database.[18,19] Protein–protein interaction data was taken from the DIP database.[20]

Matching nodes

The approach used to locate similar subgraphs in this study is as follows (Figure 2). First, the correspondence between nodes in a pair of networks was determined. Relationships between nodes in different networks can be represented as a weighted bipartite graph (Figure 2(a)). A bipartite graph is one in which the nodes can be divided into two sets such that no edge exists between any pair of nodes in the same set. The weighted edges between nodes in different networks indicate the degree of similarity between them. Nodes in a pair of biological networks may be similar to more than one node in the opposite network. For example, a protein in a protein interaction network may have an orthologue in another network but that orthologue may be difficult to distinguish from any paralogues that may exist. For protein–protein interaction networks, two methods were used to find corresponding nodes. A weighted bipartite graph-matching algorithm (the Hungarian method[17]) was used to determine an optimal one-to-one correspondence of nodes (Figure 2(b)). An alternative method determined corresponding nodes by finding the most sequence-similar pair of proteins with the same COG function (referred to here as the COG function-matching method).

For both of these methods, sequence similarity relationships between proteins in opposite networks were determined using BLAST sequence comparisons.[30,31] There were many BLAST hits between some proteins (Figure 2(a)) with differing sequence similarity. The weight assigned to each BLAST hit with an E-value of E between a pair of proteins was defined as -log(E) (or as 1000 if $E = 0.0$). The weight assigned to the correspondence between each pair of proteins was the sum of the weights corresponding to BLAST hits in each direction. Firstly, the one-to-one node correspondence was calculated using the

Table 4. Functions of proteins in Figure 6

	D. melanogaster		S. cerevisiae	
	Symbol	COG class	Symbol	COG class
(a)	A	–	A	–
	B	MAPKKK (MAP kinase kinase kinase) SSK2 and related serine/threonine protein kinases	B	–
	C	–	C	cAMP-dependent protein kinase catalytic subunit (PKA)
	D	Serine/threonine protein kinase	D	cAMP-dependent protein kinase catalytic subunit (PKA)
(b)	A	Mitotic checkpoint serine/threonine protein kinase	A	Mitotic checkpoint serine/threonine protein kinase
	B	Mitotic checkpoint serine/threonine protein kinase	B	Mitotic checkpoint serine/threonine protein kinase
	C	Mitotic checkpoint serine/threonine protein kinase	C	Mitotic checkpoint serine/threonine protein kinase
	D	Cyclin B and related kinase-activating proteins	D	Cyclin B and related kinase-activating proteins
(c)	A	–	A	Small nuclear ribonucleoprotein (snRNP) SMF
	B	Small nuclear ribonucleoprotein F	B	Small nuclear ribonucleoprotein F
	C	U6 snRNA-associated Sm-like protein	C	U6 snRNA-associated Sm-like protein
	D	Small nuclear ribonucleoprotein (snRNP) Sm core protein	D	Small nuclear ribonucleoprotein (snRNP) Sm core protein
(d)	A	Protein required for fusion of vesicles in vesicular transport, alpha-SNAP	A	Protein required for fusion of vesicles in vesicular transport, alpha-SNAP
	B	Vacuolar sorting protein VPS45/Stt10 (Sec1 family)	B	Vacuolar sorting protein VPS45/Stt10 (Sec1 family)
	C	SNARE protein Syntaxin 1 and related proteins	C	SNARE protein PEP12/VAM3/Syntaxin 7/Syntaxin 17
	D	SNARE protein TLG2/Syntaxin 16	D	SNARE protein TLG2/Syntaxin 16

The function of each protein represented in Figure 6 is listed here. For (a)–(d), the symbols in Figure 6 in each species are listed next to their function according to the COG database. Dashes indicate that COG lists no function for that protein.

Hungarian method.[17] The code used to perform the Hungarian algorithm was a slight modification of that found in Knuth's GraphBase.[32] Any final correspondence in which the E-value in either direction was >0.001 was removed from consideration. Secondly, the one-to-one node correspondence was calculated using the COG function-matching method. In this case, for each COG functional category, the protein pair with the highest weight was deemed to be a pair of corresponding nodes. In this way, all corresponding node pairs had the same function. However, as COG does not define functions for all sequence-similar proteins, and only one corresponding node pair was selected for each function, there were fewer corresponding nodes overall (951).

For the comparison of metabolic networks, the corresponding nodes were simply those that represented the same compound. Thus, it was not necessary to use a node-matching step in PHUNKEE to achieve a one-to-one correspondence between nodes in the pair of metabolic networks (Figure 2(a)).

Shared-edge ratio

The corresponding nodes determined by the above process were deemed to be equivalent (Figure 2(b)). When comparing subgraphs, a distinction was made between internal and external nodes. A node was defined to be internal if it belonged to one of the pair of subgraphs and external if it did not. Corresponding nodes that are both internal or both external are called shared nodes. Corresponding nodes consisting of one internal and one external node, and nodes with no corresponding node in the other species, are called unshared nodes. If a pair of shared nodes is connected by the same type of edge (for example, if the edges represented the same enzyme function) in both species then that edge is called a shared edge. All other edges are called unshared edges.

The measure of similarity of a pair of subgraphs is the shared-edge ratio (Figure 1), which is the weighted proportion of edges adjoining all nodes in a subgraph pair that are shared edges. Internal and external edges are weighted by the user-defined parameters w_i and w_e, respectively. An internal edge is any edge connecting two internal nodes. An external edge is any edge connecting an internal node and an external node. Thus, the shared-edge ratio (*SER*) is given by:

$$SER = \frac{\sum_{\text{edges}} \delta_{\text{shared}} w}{\sum_{\text{edges}} w}$$

where $\delta_{\text{shared}} = 1$ if the edge is shared and $\delta_{\text{shared}} = 0$ is the edge is unshared and where $w = w_i$ if the edge is internal and $w = w_e$ if the edge is external.

Searching for similar subgraphs

A grouping algorithm is then employed to find the most similar subgraphs in the two networks. A pair of corresponding nodes were selected to be the start of a trial pair of subgraphs (the nodes labelled A in Figure 2(b)). Nodes were then progressively grouped to this subgraph according to the shared-edge ratio that would occur in the new subgraph that would result (the nodes labelled B in Figure 2(c) and those labelled C in Figure 2(d)). Pairs of corresponding nodes considered for grouping are those within a user-defined maximum distance of a node within the current subgraph. Distances between pairs of nodes in each network were precalculated using the Floyd-Warshall algorithm.[33] Only distances from shared nodes within the current subgraph were considered, due to memory considerations. Any unshared nodes (and adjoining edges) that lie on the shortest path between the current subgraph and the corresponding node pair being considered would also form part of the new subgraph (for example, the node G in Figure 2(c)). For each potential new subgraph, the shared-edge ratio was calculated and the subgraph with the highest ratio was chosen to be the new current subgraph. In the event of more than one potential new subgraph having the highest shared-edge ratio, the first one found was selected. This process was repeated until the subgraph reached a user-defined maximum number of nodes or edges.

Similar pairs of subgraphs were searched for, using all possible corresponding node pairs as the initial subgraph pair.

Non-redundant subgraphs

To eliminate redundant subgraphs, a simple clustering procedure was employed. The list of subgraphs was first ordered by descending shared-edge ratio. Then, any subgraph pair with >25% nodes in common with the subgraph pair at the top of the list was eliminated. This procedure was repeated for the next subgraph pair remaining on the list and so on until no more subgraph pairs could be eliminated.

Statistical significance

Two different measures of statistical significance of subgraph matches were used in this study, the global and the local significance. Common to both measures was the random generation of a large number of subgraph pairs and their corresponding shared-edge ratios. This distribution of shared-edge ratios was used to measure the significance of the subgraph match. The significance p value was taken as the proportion of that distribution with a greater shared-edge ratio than that of the subgraph pair of interest. The difference between the global and the local significance measures was the manner in which the random subgraph pairs were generated.

Local significance

For this measure, a large number of random subgraph pair searches were conducted on the pair of networks of interest. To generate a random subgraph pair, the process described in the Searching for similar subgraphs section was followed with one exception: instead of selecting the new subgraph with the highest shared-edge ratio when extending the current subgraph, one of the new subgraphs was simply chosen at random. Any random pair of subgraphs that did not reach the user-defined subgraph size was not considered. For the comparison of protein-protein interaction networks from *D. melanogaster* and *S. cerevisiae*, ten random subgraph pairs were generated for each corresponding node pair. Where corresponding nodes were determined using the Hungarian method, a total of 13,350 random subgraphs resulted. For the COG function-matching method, 8220 random trials were generated. For metabolic network comparison, 25 random subgraph pairs were generated for each corresponding node pair (giving a total number between 4450 and 12,025 for the species compared in this study).

Global significance

For this measure, a large number of subgraph pair searches were conducted on random networks. The random networks were generated by randomly reassigning node and edge labels for the networks of interest. The random subgraph pairs were then generated by applying the search algorithm described in the Searching for similar subgraphs section to pairs of random networks. Once again, any random pair of subgraphs that did not reach the user-defined subgraph size was not considered. For the comparison of protein–protein interaction networks from *D. melanogaster* and *S. cerevisiae*, random subgraph pairs were generated for each corresponding node pair in ten pairs of random networks. Where corresponding nodes were determined using the Hungarian method, a total of 16,610 random subgraphs resulted. For the COG function-matching method, 8991 random trials were generated. For metabolic network comparison, random subgraph pairs were generated for each corresponding node pair in 25 pairs of random networks (giving a total number between 5160 and 12,433 for the species compared in this study).

NetworkBLAST

The *D. melanogaster* and *S. cerevisiae* protein interaction networks were also compared using the NetworkBLAST program‡.[13] Similar pairs of clusters and paths (four nodes long) were identified. The statistical significance of each resulting cluster and path was determined using clusters and paths generated from 100 random simulations. In each case, the p value was given by the proportion of scores from the random simulations that were more favourable than the score of the cluster/path of interest.

Combining subgraph searches

Subgraph searches were combined to determine whether this would result in a larger number of significant subgraph pairs than an isolated subgraph search. A modified level of statistical significance was required to account for the fact that finding subgraph pairs of p value = 0.01 simply by chance is more likely in multiple searches than in a single search. Therefore, we adjusted the p value required for significance in each of the N combined runs to $p = 0.01/N$. For example, when combining four PHUNKEE runs with different external edge weights ($w_e = 0$, 0.1, 0.5, 1) we selected subgraph pairs from each search with $p < 0.0025$. We applied a similar approach when comparing single PHUNKEE and single NetworkBlast searches to a combination of four PHUNKEE searches and one NetworkBlast search, selecting subgraph pairs with a significance p value < 0.002 in each search.

Acknowledgements

The authors thank R. Chaleil, K. Fleming, R. Wienzierl, K. O'Hare, N. Angelopolous, M. Sergot and H. Watanabe for useful discussions. We thank T. Ideker and M. Smoot for help with the NetworkBlast software. This work was supported by BBSRC grant number 28/BEP17011.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.03.013

‡ Available at the website chianti.ucsd.edu/NetworkBlast/

References

1. Alm, E. & Arkin, A. P. (2003). Biological networks. *Curr. Opin. Struct. Biol.* **13**, 193–202.
2. Tucker, C. L., Gera, J. F. & Uetz, P. (2001). Towards an understanding of complex protein networks. *Trends Cell. Biol.* **11**, 102–106.
3. Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A. & Palsson, B. O. (2003). Metabolic pathways in the post-genome era. *Trends Biochem. Sci.* **28**, 250–258.
4. Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. & Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14**, 283–291.
5. Barabasi, A. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.
6. Albert, R., Jeong, H. & Barabasi, A. (2000). Error and attack tolerance of complex networks. *Nature*, **406**, 378–382.
7. Yu, H., Zhu, X., Greenbaum, D., Karro, J. & Gerstein, M. (2004). TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucl. Acids Res.* **32**, 328–337.
8. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I. *et al.* (2004). Superfamilies of evolved and designed networks. *Science*, **303**, 1538–1542.
9. Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R. & Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA*, **100**, 11394–11399.
10. Ogata, H., Fujibuchi, W., Guto, S. & Kanehisa, M. (2000). A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucl. Acids Res.* **28**, 4021–4028.
11. Koyutürk, M., Grama, A. & Szpankowski, W. (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, **20**, i200–i207.
12. Berg, J. & Lässig, M. (2004). Local graph alignment and motif search in biological networks. *Proc. Natl Acad. Sci. USA*, **101**, 14689–14694.
13. Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P. *et al.* (2005). Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.
14. Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H. & Batzoglou, S. (2006). Græmlin: general and robust alignment of multiple large interaction networks. *Genome Res.* **16**, 1169–1181.
15. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnol.* **21**, 697–700.
16. Milo, R., Shen-Orr, S., Itzkowitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.
17. Kuhn, H. W. (1995). The Hungarian method for the assignment problem. *Naval Res. Logistics Quart.* **2**, 88–97.
18. Tatsuov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V. *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
19. Tatsuov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631–637.
20. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucl. Acids Res.* **28**, 289–291.
21. Grumbling, G., Strelets, V. & Consortium, F. (2006). FlyBase: anatomical data, images and queries. *Nucl. Acids Res.* **34**, D484–D488.
22. Cherry, J. M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C. *et al.* (1997). Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387**, 67–73.
23. Goldberg, D. S. & Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA*, **100**, 4372–4376.
24. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
25. Jurica, M. S. & Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell*, **12**, 5–14.
26. Hastings, M. L. & Krainer, A. R. (2001). Pre-mRNA splicing in the new milliennium. *Curr. Opin. Cell Biol.* **13**, 302–309.
27. Alves, R., Chaleil, R. & Sternberg, M. J. (2002). Evolution of enzymes in metabolism: a network perspective. *J. Mol. Biol.* **320**, 751–770.
28. Kanehisa, M. (1997). A database for post-genome analysis. *Trends Genet.* **13**, 375–376.
29. Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30.
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
31. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
32. Knuth, D. E. (1993). *The Stanford GraphBase: A Platform for Combinatorial Computing*. ACM Press, New York.
33. Sedgewick, R. (2002). *Algorithms in C*, 3rd edit. Addison-Wesley, Boston.

Edited by B. Honig

(Received 22 September 2006; received in revised form 9 February 2007; accepted 2 March 2007)
Available online 14 March 2007