BIOINFORMATICS ORIGINAL PAPER

Vol. 00 no. 00 2006, pages 1–8
doi:10.1093/bioinformatics/btl338

Systems biology

A lock-and-key model for protein–protein interactions

Julie L. Morrison[1,2,*], Rainer Breitling[3], Desmond J. Higham[2] and David R. Gilbert[1]

[1]Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, G12 8QQ, UK, [2]Department of Mathematics, University of Strathclyde, G1 1XH, UK and [3]Groningen Bioinformatics Centre, University of Groningen, Kerklaan 30, 9751 NN Haren, Netherlands

Received on February 20, 2006; received on April 26, 2006; accepted on June 15, 2006
Advance Access publication . . .
Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: Protein–protein interaction networks are one of the major post-genomic data sources available to molecular biologists. They provide a comprehensive view of the global interaction structure of an organism's proteome, as well as detailed information on specific interactions. Here we suggest a physical model of protein interactions that can be used to extract additional information at an intermediate level: It enables us to identify proteins which share biological interaction motifs, and also to identify potentially missing or spurious interactions.

Results: Our new graph model explains observed interactions between proteins by an underlying interaction of complementary binding domains (lock-and-key model). This leads to a novel graph-theoretical algorithm to identify bipartite subgraphs within protein–protein interaction networks where the underlying data are taken from yeast two-hybrid experimental results. By testing on synthetic data, we demonstrate that under certain modelling assumptions, the algorithm will return correct domain information about each protein in the network. Tests on data from various model organisms show that the local and global patterns predicted by the model are indeed found in experimental data. Using functional and protein structure annotations, we show that bipartite subnetworks can be identified that correspond to biologically relevant interaction motifs. Some of these are novel and we discuss an example involving SH3 domains from the *Saccharomyces cerevisiae* interactome.

Availability: The algorithm (in Matlab format) is available (see http://maths.strath.ac.uk/ãas96016/lock_key.html)

Contact: jmorriso@dcs.gla.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

AQ: Note that Supplementary Material should be available at Bioinformatics online. If not then please provide a URL for the same.

1 INTRODUCTION

The vast growth in availability of high-throughout protein–protein interaction datasets is widely documented (Bork *et al*., 2004) and has been accompanied by discussion emphasising the high error rates within such datasets. This combination necessitates the development of robust analytical techniques to gain knowledge about the resultant protein–protein interaction networks (Edwards *et al*., 2002). Graph-theoretic tools have proved successful, although they have largely focussed on global rather than local properties of the networks (Salwinski and Eisenberg, 2003). By modelling interactions based on local properties of the proteins we gain reasoning behind the occurrence of interactions, which may help in identifying false-positive and false-negative interactions. Also, understanding the interactions at a local level allows us in turn to make inferences about the global network topology.

The essence of our approach to modelling and thereby gaining further insight into the local and global structure of protein–protein interaction networks is the idea of lock-and-key domains. Physical interactions between protein domains are responsible for the interactions between proteins. Thus, modelling interaction networks in terms of the domains that each protein contains is biologically justified. The lock-and-key structure defines interactions to be observed, with some probability, between proteins which contain complementary domains (lock and key). This results in a network composed of near complete bipartite subgraphs. The algorithm designed in this lock-and-key framework is intended for application on networks derived from experiments where interactions are observed in a pairwise fashion, such as yeast two-hybrid data (Y2H). For the purpose of this paper we use the term 'domain' in the broadest possible sense. Common lock domains (as well as key domains) can be equivalent interaction surfaces, without being evolutionary homologues and even without a strict requirement for similarity and exact definition at the structural level.

This modelling approach has greater biological grounding than previous attempts that model protein–protein interaction networks with off-the-shelf classes of random graph. In particular, it had been widely believed that the degree distribution of protein–protein interaction networks followed a power-law, indicating a scale-free structure (Jeong *et al*., 2000). There is mounting evidence to suggest that this is not the case (Prulj *et al*., 2004; Khanin and Wit, 2006), so simply fitting a scale-free model to the data is not a valid approach.

The use of protein domains to validate protein–protein interactions is growing. For example, a statistical method developed by Riley *et al*. (2005), can be used to verify known domain–domain interactions, identify highly specific domain–domain interactions and find domain–domain interactions involving domains of unknown function. The novelty of our approach is that domain information is identified from interaction data alone.

Assuming the lock-and-key interaction structure, we define a mathematical model and a subsequent algorithm that allows us to extract domain information about each protein in the network. This approach is verified on synthetic data generated using the lock-and-key definition. We also demonstrate that the approach is robust to the introduction of false positive and false negative interactions. We then identify a number of interaction structures indicating a

*To whom correspondence should be addressed.

lock-and-key pattern in real interactomes across a wide range of species and provide biological interpretations for some of these structures.

2 METHODS

2.1 Data

The mathematical model on which we base our analysis describes pairwise interactions of proteins, rather than agglomerates or large complexes. This corresponds most closely to the experimental situation prevailing in Y2H experiments. Y2H interactions were obtained from BIND—the Biomolecular Interaction Network Database Version 3.8 (June 20, 2005) (Alfarano *et al*., 2005). In an attempt to cover as broad a range of species as possible, networks were constructed for all species for which >500 interactions had been reported. These were *Helicobacter pylori*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*. In the networks each node represents a protein and each protein–protein interaction is represented by an edge. For yeast (*S.cerevisiae*) we also examined networks corresponding to the classical Y2H studies by Ito *et al*. (2001) and Uetz *et al*. (2000). Further details of these networks are given in Supplementary Material.

The presence of noise in high-throughput protein–protein interaction datasets is widely known [it has been suggested that between 30–50% of high-throughput interactions are biologically relevant (Bader *et al*., 2004)] and, thus, we understand that the datasets are far from complete. Despite the presence of false positives and false negatives, our aim is to produce a robust algorithm which will identify any bipartite structures if they exist in the available data.

As a negative control, we also analysed the yeast dataset described in von Mering *et al*. (2002), which combines interactions observed for yeast using a number of experimental techniques, including mass spectrometry, that do not identify pairwise binding. As these type of data do not conform to our model, we do not expect to find strong bipartite patterns. Protein domain and function annotations were extracted from annotation files obtained from Affymetrix.

2.2 Interaction model and analytical algorithm

We propose a model for protein–protein interaction networks that reflects the manner in which proteins bind to each other in experiments such as Y2H assays. The model is based on a lock-and-key principle, where proteins interact only if one protein contains the 'lock' aspect of some interaction surface, and the other protein contains the matching 'key'. We also assume that an interaction will be observed between such a pair of proteins with some probability $0 \leq \theta \leq 1$. An immediate consequence of these assumptions is the prediction that there will exist nearly complete bipartite subgraphs within the protein–protein interaction networks, i.e. two groups of proteins with little or no intra-group connections but strong inter-group connections. This idea of 'complementary domains' was introduced in Thomas *et al*. (2003). In that work, domains were assigned at random in order to develop a random graph model that matched the degree distribution of experimental datasets. In our work, we impose further assumptions and develop a technique for identifying domains. Thus, unlike in Thomas *et al*. (2003), our aim is to extract information from the network. We note that there can be any number of lock-and-key pairs within a protein–protein interaction network and thus the network may consist of many bipartite subgraphs. In our analysis, we focus on identifying proteins associated with one specific lock-and-key pair at a time. We point out that the choice of which element to call a lock and which a key is arbitrary.

We introduce the following notation. Let $A \in \mathbb{R}^{N\times N}$ denote the adjacency graph, where $a_{ij} = a_{ji} = 1$ if proteins i and j interact and $a_{ij} = a_{ji} = 0$ otherwise. Focussing on a particular lock/key combination, we define an indicator vector $\mathbf{u} \in \mathbb{R}^N$ such that

$$u_i = \begin{cases} \alpha, & \text{if protein } i \text{ has lock} \\ \beta, & \text{if protein } i \text{ has key} \\ 0, & \text{otherwise} \end{cases}.$$

Here α and β are real numbers that will be specified later.

In order to get a clean mathematical analysis, we make the following simplifying assumptions (justification for these is given in the next paragraph).

(1) For this lock-and-key combination, any protein which contains the lock/key
   (a) will not also contain the key/lock, and
   (b) will only interact with a protein containing the complementary key/lock (it will not contain any other locks or keys).

(2) For each protein having this lock or key, owing to experimental constraints only a fixed proportion, θ, of its connections with the matching key/lock will be recorded.

The first assumption ensures that the bipartite subgraph under consideration is isolated from the rest of the network. Note that we are not placing restrictions on the interactions of proteins in the remainder of the network. The second assumption is a type of mean-field approximation—individual proteins in the subgraph connect with the average frequency of the ensemble. Although these are clearly idealizations, we demonstrate below that the main features from our analysis are robust to the presence of multidomain proteins and varying connectivity frequency.

If we let locksum and keysum denote the total number of proteins that contain the one particular lock or key under investigation, our assumptions imply that the i-th component of the matrix-vector product $A\mathbf{u}$ is given by

$$(A\mathbf{u})_i := \sum_{j=1}^{N} a_{ij}u_j = \begin{cases} \beta \times \theta \times \text{keysum}, & \text{if protein } i \text{ has lock} \\ \alpha \times \theta \times \text{locksum}, & \text{if protein } i \text{ has key} \\ 0, & \text{otherwise.} \end{cases}$$

If $\beta\ \theta$ keysum $= \lambda\alpha$ and $\alpha\theta$ locksum $= \lambda\beta$, for some value λ, then we have $(A\mathbf{u})_i = (\lambda\mathbf{u})_i$. In this case, $\mathbf{u}$ is an eigenvector of A with eigenvalue λ. These constraints give $\alpha^2/\beta^2 =$ keysum/locksum. Ignoring trivial re-scalings, this results in two distinct solutions, $\lambda = \pm\theta\sqrt{\text{keysum} \times \text{locksum}}$. Thus, we predict that the matrix A will have a pair of eigenvalues $\pm\theta\sqrt{\text{keysum} \times \text{locksum}}$ with corresponding eigenvectors whose non-zero components take only two possible values: one value $\pm\sqrt{\text{keysum}}$ and the other value $\pm\sqrt{\text{locksum}}$.

In other words, if we let $\mathbf{ind}^{[\text{lock}]}$ and $\mathbf{ind}^{[\text{key}]}$ be indicator vectors for the lock and key, so that

$$\mathbf{ind}_i^{[\text{lock}]} = \begin{cases} 1 & \text{if protein } i \text{ haslock} \\ 0 & \text{otherwise} \end{cases}$$

and similarly for $\mathbf{ind}^{[\text{key}]}$, then the two eigenvectors have the form

$$\mathbf{u}^{|a|} = \alpha\mathbf{ind}^{|\text{lock}|} + \beta\mathbf{ind}^{|\text{key}|}, \quad \mathbf{u}^{|b|} = -\alpha\mathbf{ind}^{|\text{lock}|} + \beta\mathbf{ind}^{|\text{key}|}.$$

Hence $\mathbf{u}^{|a|} + \mathbf{u}^{|b|} = 2\beta\mathbf{ind}^{|\text{key}|}$ and $\mathbf{u}^{|a|} - \mathbf{u}^{|b|} = 2\alpha\mathbf{ind}^{|\text{lock}|}$, so the sum and difference of the eigenvectors reveal which proteins have the lock and which have the key. We will refer to these vectors as the Sum and Difference vectors, and they form the basis of our algorithm to determine domain information.

Since the model involves a number of simplifying assumptions, we expect equalities to become approximations for real data. Fortunately, symmetric matrices have well-conditioned eigenvalues and eigenvectors (Golub and Van Loan, 1996), and hence the predictions from the model are likely to carry through when the idealised adjacency matrix undergoes perturbations. Supporting tests are carried out below.

3 RESULTS AND DISCUSSION

Using synthetic data generated under the lock-and-key principle, we first show that the eigenvalues and eigenvectors continue to hold

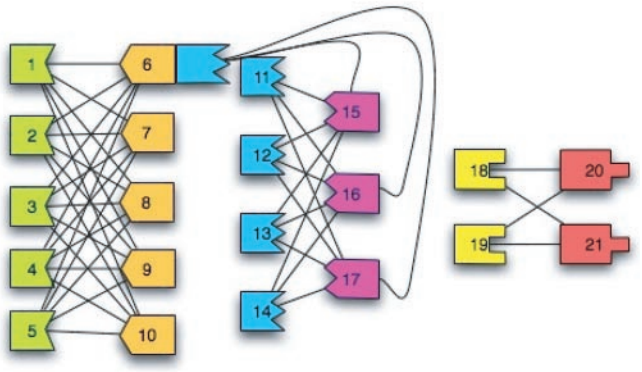

Fig. 1. Synthetic Network with $\theta = 1$.

useful information when the simplifying assumptions are relaxed. This leads to the development of an algorithm, which we then test on experimental datasets.

3.1 Synthetic data

For our first test case we consider the network shown in Figure 1. This network has three interaction types, with a total of six lock and key domains. The protein labelled 6 contains two interaction domains, violating one of our simplifying assumptions. Motivated by our analysis, we first calculate the eigenvalues of the corresponding adjacency matrix and look for pairs of the form $\pm\lambda$. We find that there exist two eigenvalues of ± 2. By taking the sum and the difference of the eigenvectors corresponding to these eigenvalues, we are able to identify the proteins in the network with the lock and the key of one particular interaction type. From Figure 2a, we can see that only two non-zero values exist in the sum and difference vectors. These correspond directly to the proteins labelled 18, 19, 20 and 21, which have the lock and key of the third interaction type. Despite the fact that protein number 6 contains two sites belonging to different domains, we still have two remaining eigenvalue pairs of ± 3.76 and ± 5.18. For the second pair, the Sum and Difference vectors provide domain classification for both remaining domains (Fig. 2b). The non-zero values in the Sum vector correspond to the proteins that contain the key of the first and second interaction types, where as the non-zero values in the Difference vector correspond to the proteins that have the lock of the first and second interaction types, and also to the single protein that contains two domains. (Of course, all protein and domain numbering is purely arbitrary and is only for reference purposes.)

We now test our algorithm's ability to recover the correct domain information when the network above is altered so that only 80% of the possible links are observed ($\theta = 0.8$). We find that we can still classify the eigenvalues into three pairs of ± 1.62, ± 3.17 and ± 4.34. We first examine the Sum and Difference vectors corresponding to the ± 1.62 pair (Fig. 3a). Although these vectors do not show equal non-zero components, extracting any non-zero components from both vectors leads to the two groups containing the key and lock aspects of the third interaction type. Determining domain information about the remaining two interaction types is less straightforward, but can be be done with either of the two remaining eigenvector pairs. Using the Sum and Difference vectors corresponding to the ± 4.34 eigenvalue pair, from Figure 3b we see that proteins may be assigned to the lock/key domain if their associated component is greater than some threshold. If 0.4 is chosen as the threshold, we find all proteins that contain the lock/key aspect of the first interaction surface, excluding the single protein that contains two interaction domains. All remaining non-zero components identify the second interaction type.

Based on the idea of assigning proteins to domains if their corresponding component in either vector is above a threshold value, the following pseudo code describes our algorithm.

```
Calculate eigenvalues/vectors of adjacency matrix
Group eigenvalues into pairs of the form ≈ ±λ
For each eigenvalue pair (with eigenvectors u_a and u_b)
    Construct Sum = u_a u_b and Diff = u_a − u_b
    Sort Sum and Diff by decreasing magnitude
    Identify a threshold for each vector
    Assign components of Sum above threshold to lock
    Assign components of Diff above threshold to key
end
```

As a measure of how well the algorithm performs, a bipartite graph with 15 locks and 20 keys was embedded within a random network with a total of 50 nodes. For both vectors, we measured the area under the receiver operating characteristic curve (AUC) (see Bamber, 1975; Gribskov and Robinson, 1996) when both vectors were ordered by decreasing order of magnitude. We predict that the proteins containing the lock/key should be ordered at the top of the Sum/Difference vectors. This analysis was conducted for values of $0.1 \leq \theta \leq 1$ and averaged over 200 runs for each θ. Decreasing θ is equivalent to increasing the false-negative rate of recording interactions in the network. We also varied the false-positive rate, defined as the percentage of interactions wrongly predicted. These were introduced randomly across the entire network. Figure 4 shows the AUC against θ for three false-positive rates.

We see from Figure 4 that in all cases where $\theta > 0.7$, the ordered vectors correctly predict the domain structure (AUC $= 1$). For values of $\theta > 0.4$, the sorted vectors still produce highly accurate information (AUC > 0.9). At a high false-positive rate and lower values of θ, the domain prediction should be treated with caution although performance is still much better than random and we could still expect to obtain useful information from the ordered vectors.

To further evaluate our algorithm, we used a list of domain–domain interactions observed in PDB structures obtained from the 3DID database (http://3did.embl.de/). The list of observed domain–domain interactions is accompanied by experimentally observed protein–protein interactions which support the known domain–domain interaction. Initially a network was constructed using these protein–protein interactions, however, the algorithm was unable to identify any bipartite structures within the data. This is not surprising, as the data are based on sparse observations scattered across a wide range of organisms and thus do not provide a sufficiently accurate sample of any complete protein–protein interaction network. Also, the data are biased towards intra-protein interactions, which our approach is not designed to detect. As an

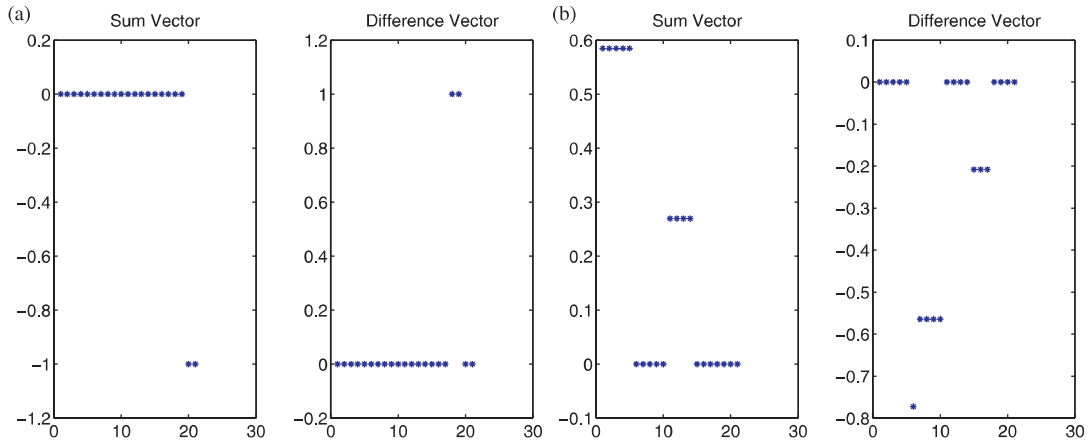

Fig. 2. Eigenvectors of the Synthetic Network. (a) Sum and Difference vectors for $\lambda = \pm 12$ (b) Sum and Difference vectors for $\lambda = \pm 5.18$

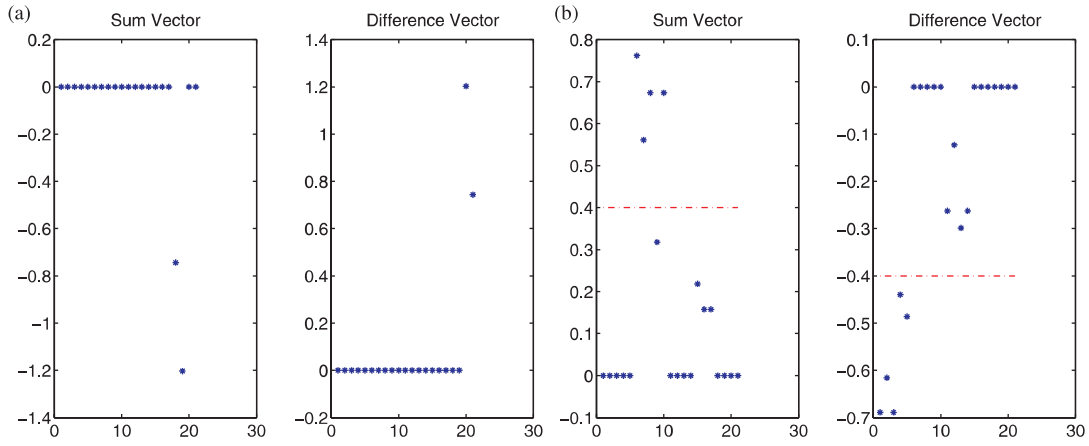

Fig. 3. Eigenvectors of the Synthetic Network with $\theta = 0.8$. (a) Sum and Difference vectors for $\lambda = \pm 1.62$. (b) Sum and Difference vectors for $\lambda = \pm 4.34$.

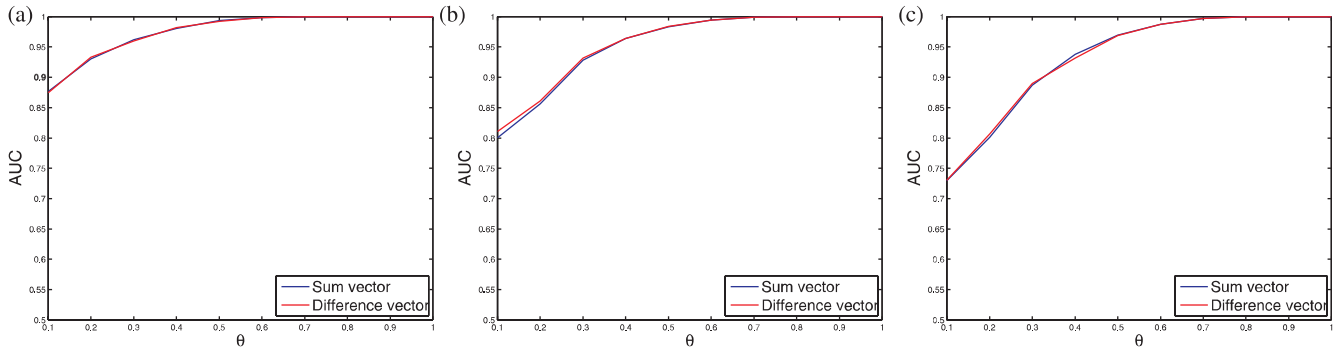

Fig. 4. ROC analysis of algorithm. (a) False positive = 0% (b) False positive = 20% (c) False positive = 40%.

alternative evaluation based on known domain information, we combined the domain–domain interaction data with from 3DID with Pfam domain assignments (Bateman *et al.*, 2004) for the yeast proteome. This was used to construct an interaction network of yeast proteins where interactions were included between any two proteins that contained a Pfam domain pair known to interact. To measure the ability of the algorithm to check for a large number of bipartite subgraphs, an automated approach to finding

Table 1. Measure of quality of bipartite subgraphs found in network constructed from domain–domain information [FP-rate (down) versus FN rate (across)]

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
0%	0.66	0.56	0.57	0.54	0.53	0.50	0.49	0.47	0.48	0.46
10%	0.50	0.47	0.45	0.44	0.41	0.40	0.37	0.39	0.36	0.34
20%	0.41	0.38	0.38	0.34	0.34	0.31	0.30	0.29	0.27	0.26
30%	0.35	0.33	0.31	0.28	0.28	0.26	0.27	0.25	0.22	0.22
40%	0.31	0.29	0.27	0.25	0.25	0.23	0.23	0.22	0.20	0.20
50%	0.28	0.26	0.27	0.24	0.22	0.23	0.20	0.20	0.19	0.19
60%	0.25	0.25	0.23	0.22	0.21	0.20	0.19	0.18	0.17	0.17
70%	0.23	0.21	0.21	0.21	0.20	0.19	0.18	0.17	0.16	0.17
80%	0.23	0.21	0.20	0.20	0.18	0.18	0.18	0.17	0.16	0.16
90%	0.20	0.18	0.19	0.18	0.18	0.16	0.16	0.16	0.16	0.15

lock-and-keys in the network was developed (further details can be found in the Supplementary Material). For each pair of Sum and Difference vectors found for this network, the 'quality' of the bipartite subgraph found was measured. To do this, all domains present in the lock and key proteins were found, and from all possible domain pairs, those which were used to construct the network were determined. For each domain pair, the total number of proteins containing each domain is known, and thus the proportion of those found in the bipartite subgraph can be found. The proportions for each domain pair were summed, and divided by the total number of domain pairs; this was the measure of quality of the subgraph. This measure was calculated for the exact network, and also perturbed networks where false-positive and false-negative interactions were randomly introduced. Results are given in Table 1. For comparison, we tested a random allocation of proteins to lock and key groups of equal size to those identified by the algorithm. The measure in this case produces a value of 0.01. We conclude that the algorithm is able to identify bipartite structure in data where they exist.

Having tested the robustness of the sorted vectors to predict domain information, we have confidence to apply our approach to experimental datasets where it is understood that false-negative and false-positive rates are high. The domain assignment predicted by the algorithm in each case can be verified by checking if a near bipartite structure exists between the assigned lock and key domain proteins.

3.2 Biological experimental data

From testing on synthetic networks, it is apparent that a heuristic approach is required to identify domain information and, thus, bipartite subgraphs in experimental datasets. In this case we can only hope to identify approximate eigenvalue pairs. This is mainly owing to the well-known noisiness of the datasets, which include a large number of false-positives and false-negatives, but also to the presence of multi-domain/multi-interaction proteins.

For all networks, except for the negative control (von Mering dataset), we are able to identify three approximate pairs of eigenvalues. For each pair in every dataset we attempt to delimit a bipartite subgraph using the method explained above. The threshold value for inclusion in the subgraph varies in each case, and is chosen by inspection of the Sum and Difference vectors. Figure 5 illustrates subgraphs that were identified. Here, for ease of visualization, we are showing the adjacency matrix, with a dot denoting a non-zero entry. Where more than one subgraph is shown for a particular species, these came from different eigenvector pairs.

We note that these structures may be used to infer protein–protein interactions by proposing that the lock and key pairs which have not been experimentally observed to interact, may in fact do so.

3.3 Biological interpretation of bipartite subgraphs

To validate the biological relevance of the observed bipartite structures we chose to focus on the yeast interaction network reported by Uetz *et al*. (2000), which has been widely analysed and comes from an organism with exceptionally well-understood biology. The subgraphs from Figure 5a and b are shown in Figure 6 with corresponding protein names.

We first focus on the smaller bipartite subgraph obtained using the second eigenvector pair (Fig. 6b). Members of this subgraph are discussed in the original paper (Uetz *et al*., 2000) as part of a larger LSM pathway. The entire group of LSM pathway proteins has 18 members, of which we have identified 8. Additional members are found if we look at the largest components in the Sum and Difference vectors: We find 17 of these proteins within the top 22 of the Difference vector, and the one protein which is not found there is ranked third in the Sum vector. This gives further evidence that these vectors represent biologically relevant information. For another validation of our results, we use the iterative Group Analysis method (iGA) (Breitling *et al*., 2004). In comparison with our technique, which uses an artificial threshold to identify bipartite subgraphs, the iGA method takes a ranked list of the entire dataset as input, along with annotations for each entity in the network, and identifies any enriched subgraphs that exist within the highly ranked proteins. We produce the ranked list by ordering the proteins in the network based on the ordering of the Sum and Difference vectors used to identify the bipartite subgraphs. The results for the second eigenvalue pair are given in Table 2. We can see that ranking the proteins on both vectors produces similar results and confirms that these proteins are involved in the LSM pathway since proteins annotated with the Pfam database term LSM are highly enriched in both lists. The results also identify the Sm domain as being highly enriched among the proteins in the bipartite set. This is again owing to the LSM proteins which are characterized by this domain. It is, however, unlikely that the Sm domain is the interaction domain in this case, since we find that the Sm domain is present in both the 'key' and 'lock' group, and both vectors produce similar rankings of proteins. This suggests that the bipartite structure identified may in fact be part of a fully connected cluster, and the connections which have been experimentally observed indicate a bipartite structure by chance. It is also important to note that the iGA analysis gives strong indications with respect to the biological function of this particular bipartite structure. It seems to be involved in spliceosomal rRNA processing, again in accordance with previous biological knowledge (Pillai *et al*., 2003).

Having validated our approach on a known subgraph, which was already discussed in the original publication, we now investigate the bipartite structure identified from the first eigenvector pair (Fig. 6a). To our knowledge this biologically very interesting group has so far escaped attention. As above, we use the iGA method to identify the enriched protein domains and functions present within this subgroup. The results are given in Table 3. Results are only included

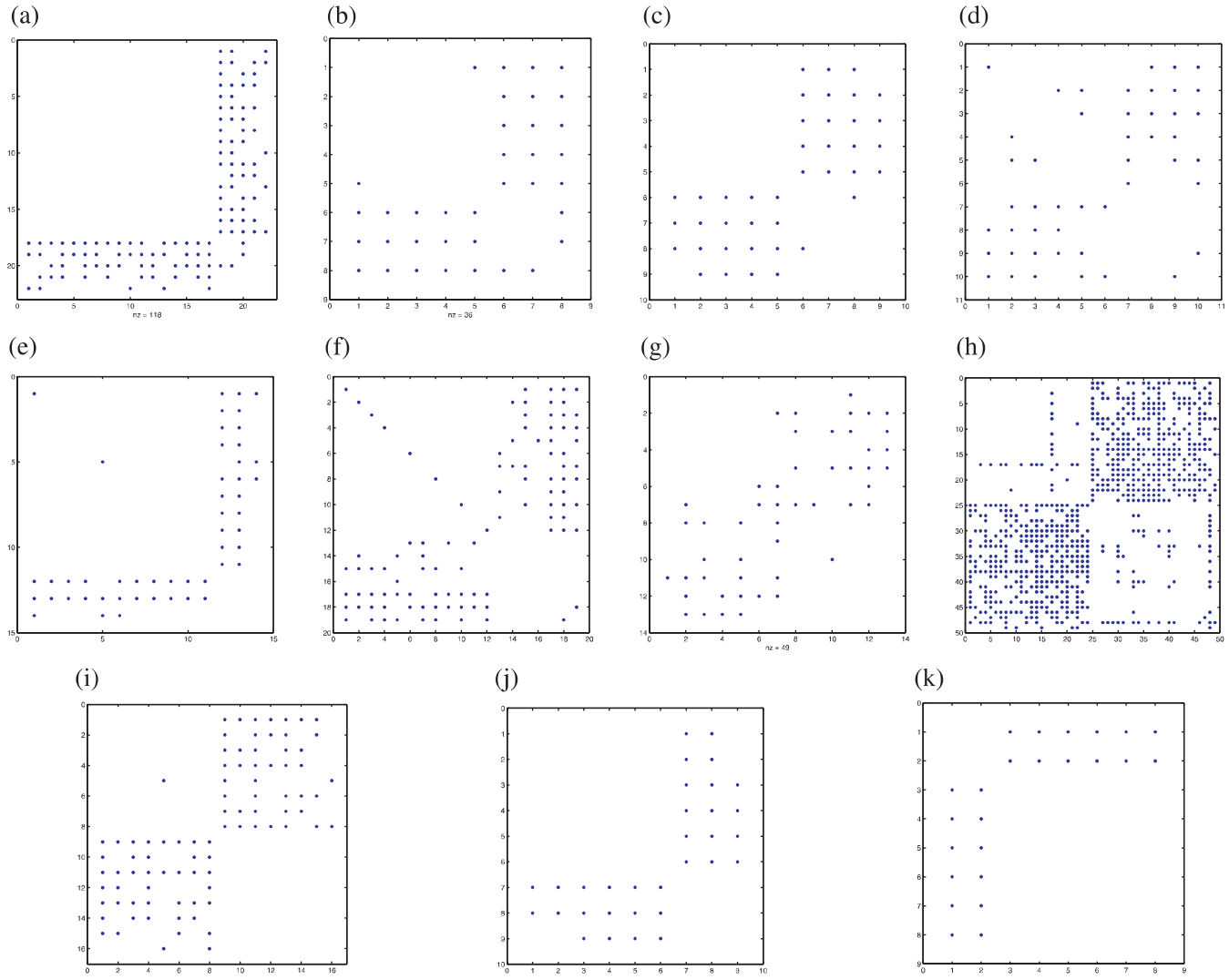

Fig. 5. Bipartite subgraphs in interactomes of different species: a bipartite structure is indicated by a two-by-two checkerboard pattern with the non-zero blocks away from the diagonal. (a) Uetz Network. 1st pair. (b) Uetz Network. 2nd Pair. (c) *A.thaliana*, 2nd Pair. (d) *H.sapiens,* 1st Pair. (e) *S.cerevisiae,* 2nd Pair. (f) *S.cerevisiae,* 3rd Pair. (g) H.*pylori,* 1st Pair. (h) *D.melanogaster,*1st Pair. (i) *D.melanogaster,* 2nd Pair. (j) *M.musculus*, 1st Pair. (k) *M.musculus*, 3rd Pair.

where the enriched subclass includes members from the bipartite subgraph.

In this case, the iGA method clearly shows that proteins with the SH3 domain are strikingly enriched within the 'key' group which is derived from the difference vector. We also obtain a first indication of the biological relevance of the interaction pattern: The GO terms 'actin cortical patch', 'actin filament organization', 'transmembrane' and 'integral to Golgi membrane' are overly abundant among the proteins of interest. These results are further strengthened when we examine the Gene Ontology annotations for the lock and key groups directly, rather than on the entire eigenvectors. The resulting *p*-values for these are listed in Table 4. Again, many proteins of the lock group are annotated with terms involving actin and Golgi, with even stronger support when these terms are combined ('all actin/Golgi combined').

The biological relevance of this interaction pattern is obvious, but was entirely unknown when the interaction dataset was first reported. The SH3 domain is one of the best characterized protein binding motifs (Mayer, 2001). It is present in all our 'key' proteins (Fig. 7) and is very likely to be the physical representative of the interaction motif. Where more than one SH3 domain is present within a protein, we are unable to determine which domain is interacting. On the other hand, the proteins of the 'lock' group are part of the actin cortical patchassembly mechanism of vesicle endocytosis (Drees *et al*., 2001). They were also identified as part of a larger group by a clustering method in Arnau *et al*. (2005), but missing the highly relevant interaction with SH3 domain proteins. The involvement of SH3 proteins in linking cytoskeletal dynamics and the trafficking of vesicles, particularly Golgi membranes, has only very recently been discovered in biological experiments (Friesen *et al*., 2005; Kessels and Qualmann, 2004). By linking vesicular membranes with actin polymerization, SH3 domain proteins contribute the crucial mechanistic connection between membrane trafficking

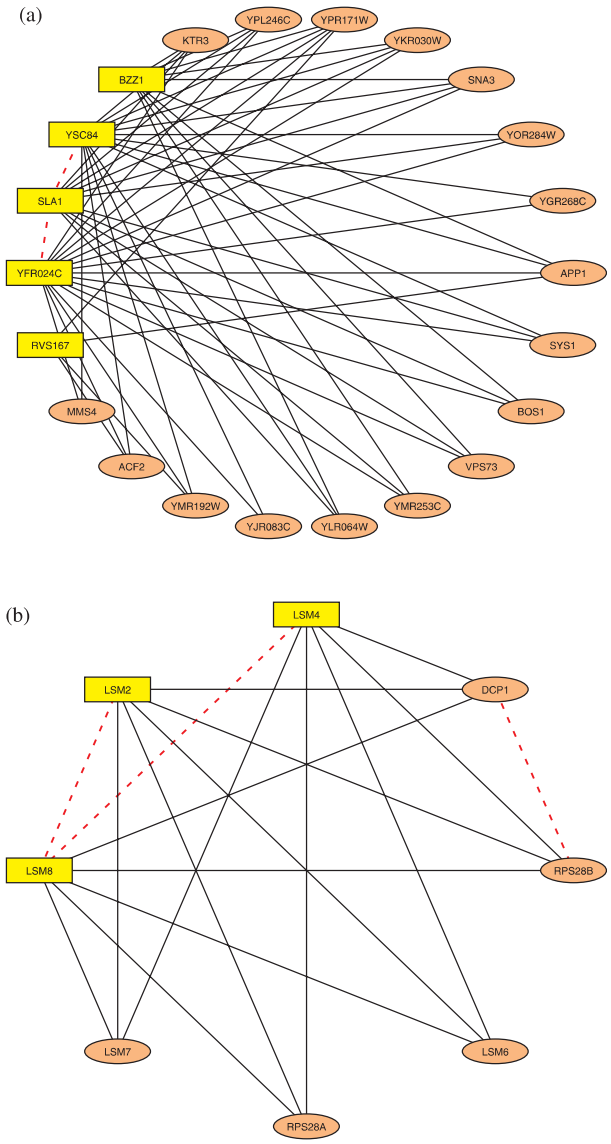

Fig. 6. Bipartite subgraphs in the Uetz network. Dotted lines indicate intra-group connections. (a) Extracted using first eigenvalue pair (b) Extracted using second eigenvalue pair.

and the cytoskeleton. The bipartite subgraph that we have identified extends on the previously reported interactions and may motivate important cell biological follow-up experiments.

For all other bipartite subgraphs identified by our algorithm, protein names and annotations are given where available in the Supplementary Material. To our knowledge these novel bipartite structures and most of the corresponding interactions have not previously been reported. These additional subgraphs include a number of other biologically very interesting gene groups, such as the ion-transporter module identified in the *Mus musculus* interactome, which further highlights the validity of our approach.

Although the number of bipartite subgraphs identified is reasonably small, our method is wholly reliant on the underlying data which is understood to be extremely noisy. At present we tend to detect only the striking examples of lock-and-key interactions. As data becomes more reliable and complete, we expect our approach to identify the lock-and-key interactions with greater coverage.

Table 2. iGA results for the second eigenvector pair ordering

Vector	Database	Class	*p*-value
Sum	Interpro	Sm_like_riboprot	6.56×10^{-9}
	Pfam	LSM	6.56×10^{-9}
	ProDom	SnRNP	6.56×10^{-9}
	InterPro	snRNP_Sm	6.56×10^{-9}
	SMART	Sm	6.41×10^{-9}
	GO	Nuclear mRNA splicing, via spliceosome	6.11×10^{-8}
	GO	Small nuclear ribonucleoprotein complex	2.18×10^{-8}
	InterPro	snRNP	1.15×10^{-8}
	GO	Pre-mRNA splicing factor activity	5.13×10^{-7}
Difference	Interpro	snRNP_Sm	8.17×10^{-9}
	ProDom	SnRNP	8.17×10^{-9}
	Pfam	LSM	8.17×10^{-9}
	Interpro	Sm_like_riboprot	8.17×10^{-9}
	SMART	Sm	7.18×10^{-9}
	GO	rRNA processing	4.65×10^{-8}
	GO	Nuclear mRNA splicing, via spliceosome	3.66×10^{-8}
	GO	Small nuclear ribonucleoprotein complex	1.15×10^{-8}
	GO	Pre-mRNA splicing factor activity	1.16×10^{-7}

Table 3. iGA results for the first eigenvector pair ordering

Vector	Database	Class	*p*-value
Add	GO	Integral to Golgi membrane	3.99×10^{-5}
Difference	ProDom	SH3	8.14×10^{-9}
	PRINTS	SH3DOMAIN	3.73×10^{-9}
	PROSITE	SH3	1.21×10^{-9}
	Interpro	SH3	1.21×10^{-9}
	Pfam	SH3	1.21×10^{-9}
	SMART	SH3	1.05×10^{-8}
	GO	Actin cortical patch	4.56×10^{-7}
	Pfam/Interpro	DUF500	2.88×10^{-6}
	GO	Actin filament organization	3.35×10^{-5}

Table 4. *p*-values for GO annotations found in Key and Lock Group

	GO term	*p*-value
Key Group	Actin filament organization	2.9916×10^{-7}
Lock Group	Actin cytoskeleton organization	4.3125×10^{-5}
	All actin combined	1.6337×10^{-11}
	Actin cortical patch assembly	1.8927×10^{-4}
	Integral to Golgi membrane	2.4789×10^{-4}
	All Golgi combined	4.6637×10^{-5}

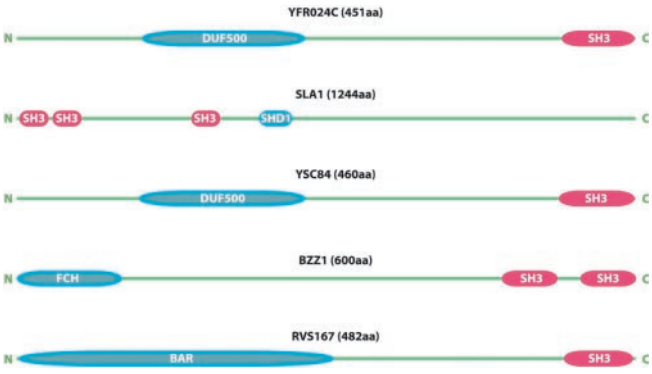

Fig. 7. SH3 domains in key group proteins.

4 CONCLUSION

From the initial lock-and-key model of protein–protein interaction networks, we have devised an algorithm that identifies proteins containing the lock and key aspects of a particular interaction surface. This is achieved through a search for bipartite subgraphs in protein–protein interaction networks derived from Y2H experiments using a spectral approach. Unlike traditional clustering techniques, we identify groups that are not internally highly similar, but have a large number of interactions with another group. We have demonstrated that under certain modelling assumptions our approach is guaranteed to identify the correct domain information about proteins in a network. As experimental interaction networks are only approximated by our model, we adopt a heuristic approach to identifying bipartite subgraphs. The main ingredients of the algorithm are Sum and Difference vectors, formed from the corresponding eigenvectors of eigenvalue pairs of (approximately) the form $\pm\lambda$. We demonstrated that this approach reveals bipartite subgraphs across a large variety of protein interaction networks from diverse species. For one of these subgraphs, from *S.cerevisiae*, we showed how our method discovers a novel and biologically exciting interacting group, including identification of the physiological function and the physical interaction motif, the SH3 domain. Used in this way, our approach has the potential to add considerable value to the experimentally observed interaction networks.

ACKNOWLEDGEMENTS

J.L.M. was supported by a Synergy scholarship (www.strath.gla.ac.uk/synergy). R.B. was supported by a Caledonian Research Foundation Personal Fellowship. D.J.H. was supported by EPSRC grant GR/S62383/01.

Conflict of Interest: none declared.

REFERENCES

Alfarano,C. *et al*. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*., **1**, D418–D424.

Arnau,V. *et al*. (2001) Iterative cluster analysis of protein interaction data. *Bioinformatics*, **21**, 364–378.

Bader,J.S. *et al*. (2004) Gaining confidence in high-throughput protein interactions. *Nat. Biotechnol*., **22**, 78–85.

Bamber,D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol*., **12**, 387–415.

Bateman,A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.

Breitling,R. *et al*. (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, **5**, 34.

Bork,P. *et al*. (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol*., **14**, 292–299.

Drees,B.L. *et al*. (2001) A protein interaction map for cell polarity development. *J. Cell boil*., **154**, 549–571.

Edwards,A.M. *et al*. (2002) Bridging structural biology and genomics: assessing protein–protein interaction datasets. *Trends Genet*., **18**, 529–536.

Friesen,H. *et al*. (2005) Interaction of *Saccharomyces cerevisiae* cortical actin patch protein Rvs167p with proteins involved in ER to Golgi vesicle trafficking. *Genetics*, **170**, 555–568.

Golub,G.H. and Van Loan,C.F. (1996) *Matrix Computations*. The John Hopkins University Press.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem*., **20**, 25–33.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interaction interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jeong,H. *et al*. (2000) The large-scale organization of metabolic networks. *Nature*, **507**, 651–654.

Kessels,M.M. and Qualmann,B. (2004) The syndapin protein family: linking membrane trafficking with the cytoskeleton. *J. Cell Sci*., **17**, 3077–3086.

Khanin,R. and Wit,E. (2006) How scale-free are gene-networks? *J. Comput. Biol*., (in press).

Mayer,B.J. (2001) SH3 domains: complexity in moderation. *J. Cell Sci*., **114**, 1253–1261.

Mrowka,R. *et al*. (2001) Is there a bias in proteome research? *Genome Res*., **11**, 1971–1973.

Pillai,R.S. *et al*. (2003) Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes Dev*., **17**, 2321–2333.

Prulj,N. *et al*. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3808–3515.

Riley,R. *et al*. (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*., **6**, R89.

Salwinski,L. and Eisenberg,D. (2003) Computational methods of analysis of protein–protein interactions. *Curr. Opin. Struct. Biol*., **13**, 377–382.

Thomas,A. *et al*. (2003) On the structure of protein–protein interaction networks. *Biochem. Soc. Trans*., **31**, 1491–1496.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale datasets of protein–protein interactions. *Nature*, **417**, 399–403.

AQ: Please update in-press reference Khanin,R. and Wit,E. (2006) citation if it has now been published.

AQ: Please provide citation for reference Mrowka,R. et al. (2001) in text.