Pacific Symposium on Biocomputing 11:231-242(2006)

IDENTIFYING PROTEIN COMPLEXES IN HIGH-THROUGHPUT PROTEIN INTERACTION SCREENS USING AN INFINITE LATENT FEATURE MODEL*

WEI CHU & ZOUBIN GHAHRAMANI

Gatsby Computational Neuroscience Unit,
University College London,
London, WC1N 3AR, UK
E-mail: chuwei,zoubin@gatsby.ucl.ac.uk

ROLAND KRAUSE

Max-Planck-Institute for Molecular Genetics
D-10117 Berlin, Germany
E-mail: roland.krause@molgen.mpg.de

DAVID L. WILD

Keck Graduate Institute of Applied Life Sciences
Claremont, CA 91171, USA
E-mail: wild@kgi.edu

We propose a Bayesian approach to identify protein complexes and their constituents from high-throughput protein-protein interaction screens. An infinite latent feature model that allows for multi-complex membership by individual proteins is coupled with a graph diffusion kernel that evaluates the likelihood of two proteins belonging to the same complex. Gibbs sampling is then used to infer a catalog of protein complexes from the interaction screen data. An advantage of this model is that it places no prior constraints on the number of complexes and automatically infers the number of significant complexes from the data. Validation results using affinity purification/mass spectrometry experimental data from yeast RNA-processing complexes indicate that our method is capable of partitioning the data in a biologically meaningful way.
A supplementary web site containing larger versions of the figures is available at http://public.kgi.edu/∼wild/PSB06/index.html.

*This work was supported by the National Institutes of Health Grant Number 1 P01 GM63208. A part of this work was carried out at Institute for Pure and Applied Mathematics (IPAM) of UCLA. We are grateful to Thomas L. Griffiths for the Matlab script of the Gibbs sampler.

Pacific Symposium on Biocomputing 11:231-242(2006)

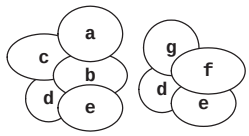
1. Introduction

The analysis of protein-protein interactions forms an essential part of the "systems biology" enterprise. Many cellular functions are performed by multi-protein complexes and the identification and analysis of protein complex membership reveals insights into both the topological properties and functional organization of protein networks. Recently, high-throughput techniques have been developed to investigate physical binding between the constituents of protein complexes on a proteome-wide scale. The yeast two-hybrid assay (Y2H), a means of assessing whether two single proteins interact, has been adapted to systematically test pairwise protein interactions on a large scale[1,2], whereas affinity purification techniques using mass spectrometry (APMS)[3] provide a particularly effective approach to identifying protein complexes that contain more than two components. These techniques have been used to perform large scale protein-protein interaction screens in the yeast *Saccharomyces cerevisiae*[4,5,6] and the bacterium *Escherichia coli*[7].

In the APMS techniques, as described by Kumar and Snyder[3], individual proteins are tagged and used as "baits" to form physiological complexes with other proteins in the cells. Then, using the tag, each bait protein is purified, retrieving the proteins to which it binds, which may *sometimes* constitute the entire complex. The proteins extracted with the bait protein are identified using standard mass spectrometry methods. The raw results of these experiments are often referred to as "purifications" and may differ substantially from what is thought to exist in the cell and what is annotated in databases of protein complexes. Identification of actual protein complexes from these "purifications" often involves manual post-processing based on the existence of overlaps between the purifications[4]. Attempts to automate complex identification have involved the use of binary protein-protein interaction graphs[8,9,10], unsupervised clustering based on special similarity measures[13,6] and graph-theoretic approaches[14]. However, these approaches are bedeviled by a number of problems, such as fact that the exact number of complexes is initially unknown; the presence of potential contaminant proteins (which may themselves form complexes); the fact that the experiments do not always retrieve whole complexes, but only sub-complexes; and the presence of shared components, which need to be assigned to more than one complex.

In this paper, we propose a probabilistic algorithm to identify protein complex membership using the data from affinity purification/mass spec-

Pacific Symposium on Biocomputing 11:231-242(2006)



(i) Native Protein Complexes

Protein	a	b	c	d	e	f	g
Bait a	1	1	1	1	*0*	0	0
Bait d	*0*	1	1	1	1	1	1
Bait f	0	0	0	1	1	1	1

(ii) Purification Matrix, B

	a	b	c	d	e	f	g
a	1	1	1	1	0	0	0
b	1	2	2	2	1	1	1
c	1	2	2	2	1	1	1
d	1	2	2	3	2	2	2
e	0	1	1	2	2	2	2
f	0	1	1	2	2	2	2
g	0	1	1	2	2	2	2

(iii) Adjacency Matrix, A

Complex	a	b	c	d	e	f	g
1	1	1	1	1	1	0	0
2	0	0	0	1	1	1	1
3	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(iv) Complex Membership, Z

Figure 1. Representations of APMS results. (i) shows the native composition of two protein complexes. (ii) represents the purification results for bait proteins, in which 1's denote positive and 0's denote negative, respectively. (iii) presents the corresponding adjacency matrix. (iv) specifies the complex membership of these proteins, where each entry is a binary random variable which indicates the members of the corresponding protein complex.

trometry (APMS) experiments. The membership between pairs of proteins is represented by a graph diffusion kernel, and the complexes and their constituents are identified by an infinite latent feature model. This model allows for multi-complex membership by individual proteins, which is fundamental for procuring an accurate protein complex catalog. The approach is novel as it identifies the complexes directly from experimental purifications. Our method provides insights into the organization of protein complexes into a core and peripherally located, possibly transiently binding components[11,12].

2. Data Representation

An example of the APMS method is shown in Figure 1. The purification results are usually recorded in the form of a binary matrix as shown in Figure 1(ii). Note that the APMS technology is neither perfectly sensitive nor specific, resulting in the failure to detect certain components (false negatives – FN) and the identification of proteins which are not members of a complex (false positives – FP). FN observations are represented by italic 0's in the example shown in Figure 1(ii). Let $\mathbf{B}$ denote the purification matrix with size $S \times N$, where S is the number of bait proteins and N is the number of proteins found in purifications. The corresponding adjacency matrix $\mathbf{A}$ is defined by $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, which is a symmetric $N \times N$ matrix. The

Pacific Symposium on Biocomputing 11:231-242(2006)

ij-th element, $\mathbf{A}_{ij}$, is the number of purifications in which both protein i and protein j appear. The similarity between protein pairs can be measured by a graph diffusion kernel[15] based on $\mathbf{A}$. In this work, we focus on the von Neumann diffusion kernel[16] as the closeness measure. Kernel methods have also been applied to the inference of biological networks from other data sources[17,18].

3. The von Neumann Diffusion Kernel

The element $\mathbf{A}_{ij}$ in the adjacency matrix can be thought of as the number of distinct "paths" between protein i and protein j discovered by the APMS experiment. For example in Figure 1, there are two paths between protein d and protein e and no path directly connecting protein a and protein e. However, we could also reach protein e indirectly from protein a via the paths through the neighbors of protein a, e.g. a-b-e a-c-e and a-d-e. The number of distinct paths with length 2 between a pair of proteins can be directly counted by the matrix product $\mathbf{AA}$. More generally, the number of paths from protein i to protein j of length ℓ on the graph can be directly counted as the ij-th element of the matrix $\mathbf{A}^\ell$. The closeness between a pair of proteins can be measured by the number of distinct paths with different length. The von Neumann diffusion kernel[16] is the limit of the sum of the geometric series, defined as

$$\mathbf{K} := \sum_{\ell=1}^{\infty} \gamma^{\ell-1} \mathbf{A}^\ell = \mathbf{A}\,(1-\gamma\mathbf{A})^{-1}. \tag{1}$$

where γ is the diffusion factor to ensure the longer range connections decay exponentially.[a] The normalized kernel is an appropriate measure of similarity, which is defined as

$$\mathbf{D}_{ij} = \frac{\mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}\mathbf{K}_{jj}}}. \tag{2}$$

Note that the matrix elements are between 0 and 1. $\mathbf{D}_{ij} = 0$ implies protein i is isolated from protein j. On the contrary, $\mathbf{D}_{ij}$ approaches 1 if the protein pair is tightly connected. The elements of the normalized von Neumann kernel (2) provide a probabilistic measure on the pairwise membership of two proteins in the same complex. This interpretation makes $\mathbf{D}_{ij}$ suitable for use as a likelihood in a probabilistic model, which we now describe.

[a]The von Neumann kernel (1) is positive definite only if $0 < \gamma < \rho^{-1}$ where ρ is the spectral radius of $\mathbf{A}$. γ could be learnt from the data. In this work, we set $\gamma = (1+\kappa)^{-1}\rho^{-1}$ where κ is the proportion of non-zero elements in the adjacency matrix $\mathbf{A}$.

Pacific Symposium on Biocomputing 11:231-242(2006)

4. Protein Complex Membership

Protein complex membership can be represented as a binary matrix, denoted as $\mathbf{Z}$ (see Figure 1(iv)). Each column of the matrix $\mathbf{Z}$ is denoted by $\mathbf{z}_i$, known as the feature vector of the protein. The length of $\mathbf{z}_i$ is variable, as the number of protein complexes is actually unknown. The membership of the i-th protein in complex c is indicated by a binary random variable z_{ci}. Note that each protein may belong to multiple complexes. The learning task is to infer a catalog of protein complexes and their constituents from the APMS experimental data.

4.1. *An Infinite Latent Feature Model*

Griffiths and Ghahramani[19] have proposed a probability distribution over binary matrices with a fixed number of columns (proteins) and an infinite number of rows (complexes), which is particularly suitable for use as a prior in probabilistic models that represent proteins with multiple complex membership. In the following we describe this infinite latent feature model[19] in the context of protein complex membership identification. Since the exact number of complexes is initially unknown, we start with a finite model that assumes C complexes, and then take the limit as $C \to \infty$ to obtain the prior distribution over the binary matrix $\mathbf{Z}$. As in other non-parametric models, taking this limit ensures that the model is flexible enough to capture any number of complexes.

We assume that each protein belongs to a complex c with probability π_c, and then given the set $\pi = \{\pi_1, \pi_2, \ldots, \pi_C\}$ the probability of matrix $\mathbf{Z}$ is a product of binomial distributions

$$\mathcal{P}(\mathbf{Z}|\pi) = \prod_{c=1}^{C} \prod_{i=1}^{N} \mathcal{P}(z_{ci}|\pi_c) = \prod_{c=1}^{C} \pi_c^{n_c}(1-\pi_c)^{N-n_c}, \tag{3}$$

where $n_c = \sum_{i=1}^{N} z_{ci}$ is the number of constituent proteins belonging to the complex c. As suggested by Griffiths and Ghahramani[19], the beta distribution is chosen to be beta$(\frac{\alpha}{C}, 1)$ where α is a model parameter.[b] In the probabilistic model we have defined, each z_{ci} is independent of all other memberships and the π_c's are also independent of each other. Given this prior on π, we can simplify this model by integrating over all possible

[b]We set $\alpha = 1$ in this work which represents our prior belief that each protein is expected to belong to one complex but probably not many more.

Pacific Symposium on Biocomputing 11:231-242(2006)

settings for π, and then compute the conditional distribution for any z_{ci} as follows

$$\mathcal{P}(z_{ci}|\mathbf{Z}_{-i,c}) = \frac{n_{-i,c} + \frac{\alpha}{C}}{N + \frac{\alpha}{C}}, \tag{4}$$

where $\mathbf{Z}_{-i,c}$ denotes the entries of $\mathbf{Z}$ except z_{ci}, and $n_{-i,c}$ is the number of proteins belonging to the complex c, not including the protein i.

The infinite model can be obtained from the finite model by taking the limit of (4) as $C \to \infty$. The conditional distribution of z_{ci} in the infinite model is then

$$\mathcal{P}(z_{ci}|\mathbf{Z}_{-i,c}) = \frac{n_{-i,c}}{N}, \tag{5}$$

for any c such that $n_{-i,c} > 0$. As for the c's with $n_{-i,c} = 0$, it can be shown that the number of new complexes associated with this protein, denoted as ν_i, has a Poisson distribution with the parameter $\frac{\alpha}{N}$ as follows,

$$\mathcal{P}(\nu_i|\mathbf{Z}_{-i,c}) = \left(\frac{\alpha}{N}\right)^{\nu_i} \frac{\exp(-\frac{\alpha}{N})}{\nu_i!}. \tag{6}$$

Details of all these properties of the infinite latent feature model can be found in Ref. 19.

4.2. *Likelihood Evaluation*

Given a particular protein complex membership matrix $\mathbf{Z}$, the pairwise membership can be determined by examining whether $\mathbf{z}_i^T\mathbf{z}_j > 0$ or $\mathbf{z}_i^T\mathbf{z}_j = 0$, which categorizes the protein pairs respectively into two classes, members of the same complex or not. The likelihood can be evaluated by the von Neumann diffusion kernel (2) directly, as $\mathbf{D}_{ij}$ exactly measures the probability of the protein pair being members of a protein complex. Therefore, the likelihood can be evaluated as follows,

$$\mathcal{P}(\mathbf{D}|\mathbf{Z}) = \prod_{\{ij:\mathbf{z}_i^T\mathbf{z}_j>0\}} (\mathbf{D}_{ij})^{\mathbf{z}_i^T\mathbf{z}_j} \prod_{\{ij:\mathbf{z}_i^T\mathbf{z}_j=0\}} (1-\mathbf{D}_{ij}), \tag{7}$$

where $\{ij\}$ denotes any distinct pair and $\mathbf{D}$ denotes the normalized von Neumann kernel matrix obtained from the APMS experiments.

4.3. *Membership Inference*

Based on Bayes' theorem, the posterior distribution of the protein complex membership $\mathbf{Z}$ can be given by $\mathcal{P}(\mathbf{Z}|\mathbf{D}) \propto \mathcal{P}(\mathbf{D}|\mathbf{Z})\mathcal{P}(\mathbf{Z})$, where $\mathcal{P}(\mathbf{D}|\mathbf{Z})$ is defined as in (7), and $\mathcal{P}(\mathbf{Z})$ is defined by the infinite latent feature model.

Pacific Symposium on Biocomputing 11:231-242(2006)

We have defined a posterior distribution for the protein complex membership that does not assume a fixed number of protein complexes and allows for multiple membership. In the following, we describe a Gibbs sampler to carry out inference in the infinite latent feature model. The critical quantity required in the Gibbs sampling is the conditional distribution

$$\mathcal{P}(z_{ci}|\mathbf{Z}_{-i,c}, \mathbf{D}) \propto \mathcal{P}(\mathbf{D}|\mathbf{Z})\mathcal{P}(z_{ci}|\mathbf{Z}_{-i,c}), \tag{8}$$

where the likelihood $\mathcal{P}(\mathbf{D}|\mathbf{Z})$ is defined as in (7), and $\mathcal{P}(z_{ci}|\mathbf{Z}_{-i,c})$ is defined as in (5) for any c with $n_{-i,c} > 0$. While for the complexes with $n_{-i,c} = 0$, the conditional distribution over the number of new complexes taken by the protein can be computed as follows

$$\mathcal{P}(\nu_i|\mathbf{Z}_{-i,c}, \mathbf{D}) \propto \mathcal{P}(\mathbf{D}|\mathbf{Z})\mathcal{P}(\nu_i|\mathbf{Z}_{-i,c}), \tag{9}$$

where $\mathcal{P}(\nu_i|\mathbf{Z}_{-i,c})$ is a Poisson distribution defined as in (6). Note that the membership of new complexes does not change the pairwise membership at all. So the likelihood $\mathcal{P}(\mathbf{D}|\mathbf{Z})$ stays equal for any value of ν_i in our model. The overall algorithm can be summarized as follows,

(1) *Initialize* $\mathbf{Z}$ *randomly*, usually start with one complex.
(2) *For* $t = 1$ *to* T
    (a) For each i and each c with $n_{-i,c} > 0$, sample z_{ci} in the distribution (8).
    (b) For each i, sample the number of new complexes in the Poisson distribution (6).
    (c) Save the sample $\mathbf{Z}$.
(3) *Exit*

In this work, we collected 1000 samples after burning in the first 1000 samples as an approximate estimate of the posterior distribution of $\mathbf{Z}$. The computational overhead of this algorithm is approximately $\mathcal{O}(TCN^2)$, where T denotes the number of samples we collect, C denotes the number of significant complexes in the data and N denotes the number of hit proteins. The algorithm has been implemented in Matlab. On a Linux Athlon 1800 desktop, it took about 89.7 seconds to process the RNA Polymerase complex data set described below.

5. Results

To validate our approach, we have applied the above algorithm to two experimental data sets: the purifications corresponding to the proteins contained

Pacific Symposium on Biocomputing 11:231-242(2006)

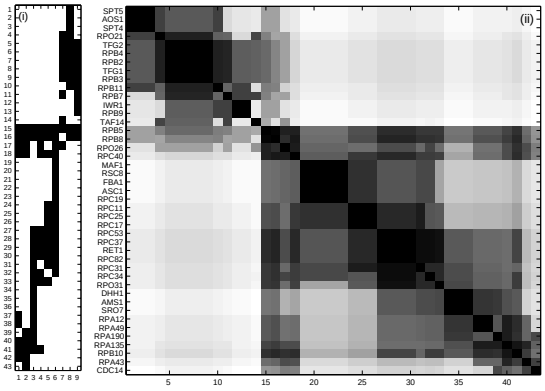


Figure 2. The RNA Polymerase complexes. (i) presents the purification results using 9 bait proteins. (ii) presents the corresponding normalized von Neumann kernel matrix, where the gray scale indicates the probability of pairwise membership defined as in (2). The proteins are sorted according to the inferred complex membership

in the RNA Polymerase complexes from the whole proteome screen of Gavin et. al[4], and the yeast RNA-Processing Complexes data of Krogan et. al[6]. The RNA Polymerase complexes are a particularly suitable and tractable test case, since they share five components and their three-dimensional structure has been determined by X-ray crystallography[20], providing a "gold standard". The data set used comprised 9 purifications (baits) and 43 proteins (hits) extracted from the data of Gavin et al.[4] as shown in Figure 2(i). Figure 2(ii) shows the normalized von Neumann diffusion kernel (2) for this data. The final protein complex assignments are shown in Figure 3. Proteins correctly assigned to the three RNA Polymerase complexes according to the MIPS protein complex database[21] are marked with crosses in Figure 3. TFG1/2 and SPT5 appear as members of the RNAP II complex under the APMS conditions used and thus are included in the prediction. RPB5 and 8 are clearly seen to be shared amongst all 3 complexes, whilst RPO26 and RPC40 are shared between RNAP I and III. Due to the nature of the experimental data, one cannot assume a perfect match, particularly as only a few baits of the whole complex were actually purified.

A subset of the more extensive yeast RNA-processing complex data of Krogan et al. comprising the "reliably identified" complexes[6] consisted of a data set of 71 purifications (baits) and 240 proteins (hits). We restricted the analysis to the data sample used for hierarchical clustering in the study of Krogan et al. (MALDI data). Inspection of the normalized von Neumann diffusion kernel for this data (Figure 4, bottom right) indicated that a subset of this data (24 baits and 49 proteins) formed clear and unam-

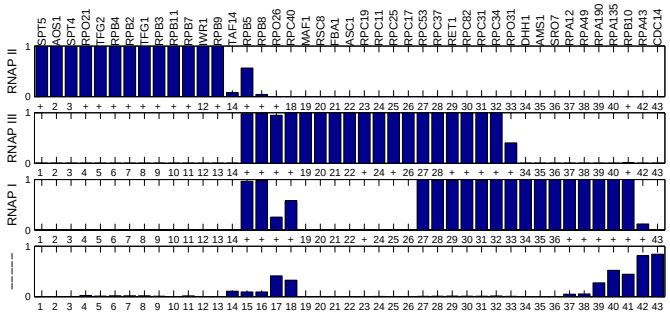


Figure 3. The four largest complexes identified by our algorithm. The bars indicate the probability of membership of the proteins. The top 3 complexes correspond to RNAP II, RNAP III and RNAP I respectively. The cross indices indicate the members of the three RNA polymerase complexes according to the MIPS protein complex database.

biguous clusters amongst themselves, and so were not processed further. This subset of the data comprised 9 clusters with membership ≥ 2, the largest of which was the 19S regulatory subunit of the proteasome comprising RPN1,2,3,5,6,7,9,10,11,12 and RPT1,3,6. Some known proteins of the proteasome are missing because they were not part of the experimental data set. The remaining data set, comprising 47 baits and 191 proteins were analyzed using the method described above, resulting in 20 clusters being identified. The assignments into the 20 identified clusters are shown in Figure 5. Most of the complexes correspond to those identified by Krogan et al.[6] such as the RNA polymerase complexes, the SSU processome, the exosome and U6 specific snRNP core. In the case of the SSU processome, we also identify a distinct UTP-A complex which includes UTP4,5,8,9,15, NAN1 and POL5, and a distinct UTP22/RRP7 complex. Interestingly, this appears as distinct from the casein kinase II complex although it co-purifies with casein kinase II subunits. Our method was also capable of identifying the TFG1/2 complex as a separate entity to the RNAP II complex even though some of its elements were purified using TFG1 as a bait. These, and other examples indicate that our method is capable of partitioning the data in a biologically meaningful way.

6. Discussion and Conclusions

We have demonstrated that the algorithm produces biological meaningful results and yields insights into how the clusters were generated, which is important for the interpretation of the results. In particular, the assignment of a protein to more than one complex and the choice of the number

Pacific Symposium on Biocomputing 11:231-242(2006)

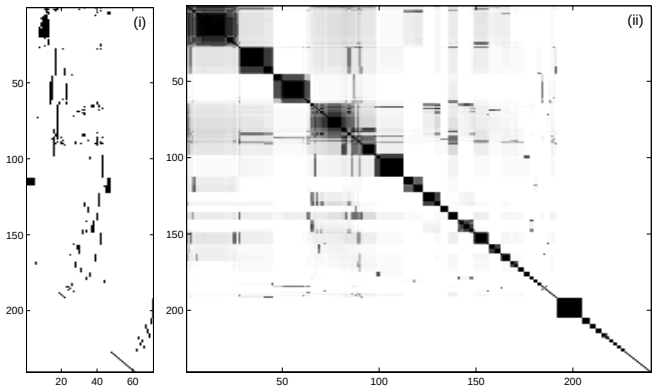


Figure 4. Yeast RNA-processing complex data of Krogan et al. (i) presents the purification results using 71 bait proteins. (ii) presents the corresponding normalized von Neumann kernel matrix of 240 hit proteins where the gray scale indicates the probability of pairwise membership defined as in (2). The proteins are sorted according to the inferred complex membership. A larger version of the figure can be found on the supplementary web site. The names of the first 191 proteins are also indexed in Figure 5.

of complexes can be performed without heuristic assumptions, a major improvement over previous methods.

Obviously, the method relies on experimental data and assignment of some artifacts such as ribosomal contaminants to complexes cannot be avoided. It should also be noted that there is no consensus amongst experts on how to identify artifacts. The interpretation must be applied in a similar fashion across the whole network of protein-protein interactions.

It would be necessary to introduce reference data sets for more comprehensive comparisons with standardized methods for identifying protein complexes. The current data sets are sparse and there is little independent confirmation for most complexes. Our method works well for the data sets used here and further improvements should be obtained when the individual complexes are better sampled.

References

1. P. Uetz, L. Giot, G. Cagney, et al., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, **403**, 623-627, (2000).
2. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci USA*, **98(8)**, 4569-4574 (2001).
3. A. Kumar and M. Snyder, Protein complexes take the bait, *Nature*, **415**, 123-124, (2002).

Pacific Symposium on Biocomputing 11:231-242(2006)

4. A.C. Gavin, M. Bosche, R. Krause, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, **415**, 141-147, (2002).
5. Y. Ho, A. Gruhler, A. Heilbut, et al., Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, *Nature*, **415**, 180-183, (2002).
6. N.J. Krogan, W.T. Peng, G. Cagney, et al., High-definition macromolecular composition of yeast RNA-processing complexes, *Molecular Cell*, **13**, 225-239, (2004).
7. G. Butland, J. M. Peregrin-Alvarez, J. Li, et al., Interaction network containing conserved and essential protein complexes in *Escherichia Coli*, *Nature*, **433**, 531-537, (2005).
8. G.D. Bader and C.W. Hogue, Analyzing yeast protein-protein interaction data obtained from different sources, *Nat Biotechnol*, **20(10)**, 991-997, (2002).
9. G.D. Bader and C.W. Hogue, An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics*, **4(1)**, (2003).
10. V. Spirin and L.A. Mirny, Protein complexes and functional modules in molecular networks, *Proc Natl Acad Sci USA*, **100(21)**, 12123-12128, (2003).
11. J. Hollunder, A. Beyer and T. Wilhelm, Identification and characterization of protein subcomplexes in yeast, *Proteomics*, **5(8)**, 2082-9, (2005).
12. Z. Dezso, Z. Oltvai and A.L. Barabasi, Bioinformatics analysis of experimentally determined protein complexes in the Yeast *Saccharomyces cerevisiae*, *Genome Res*, **13**, 2450-2454, (2003).
13. R. Krause, C. von Mering and P. Bork, A comprehensive set of protein complexes in yeast: Mining large scale protein-protein interaction screens, *Bioinformatics*, **19(15)**, 1901-1908, (2003).
14. D. Scholtens and R. Gentleman, Making Sense of High-throughput Protein-protein Interaction Data, *Statistical Applications in Genetics and Molecular Biology* **3**, Article 39, (2004).
15. R. Kondor and J. Lafferty, Diffusion kernels on graphs and other discrete structures, *ICML*, 315-322, (2002).
16. J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, (2004).
17. A. Ben-Hur and W.S. Noble, Kernel methods for predicting protein-protein interactions, *Bioinformatics*, **21**, Suppl. 1, i38-i46, (2005).
18. K. Tsuda and W.S. Noble, Learning kernels from biological networks by maximizing entropy, *Bioinformatics*, **20**, Suppl. 1, i326-i333, (2004).
19. T. L. Griffiths and Z. Ghahramani, Infinite latent feature models and the Indian buffet process *Technical Report*, **GCNU TR 2005-001**, University College London, (2005).
20. P. Cramer, D.A. Bushnell, J. Fu, et al., Architecture of RNA polymerase II and implications for the transcription mechnism, *Science*, **288**, 640-649, (2000).
21. H.W. Mewes, D. Frishman, U. Guldener, et al., MIPS: a database for genomes and protein sequences, *Nucleic Acids Res*, **30**, 31-34, (2002).

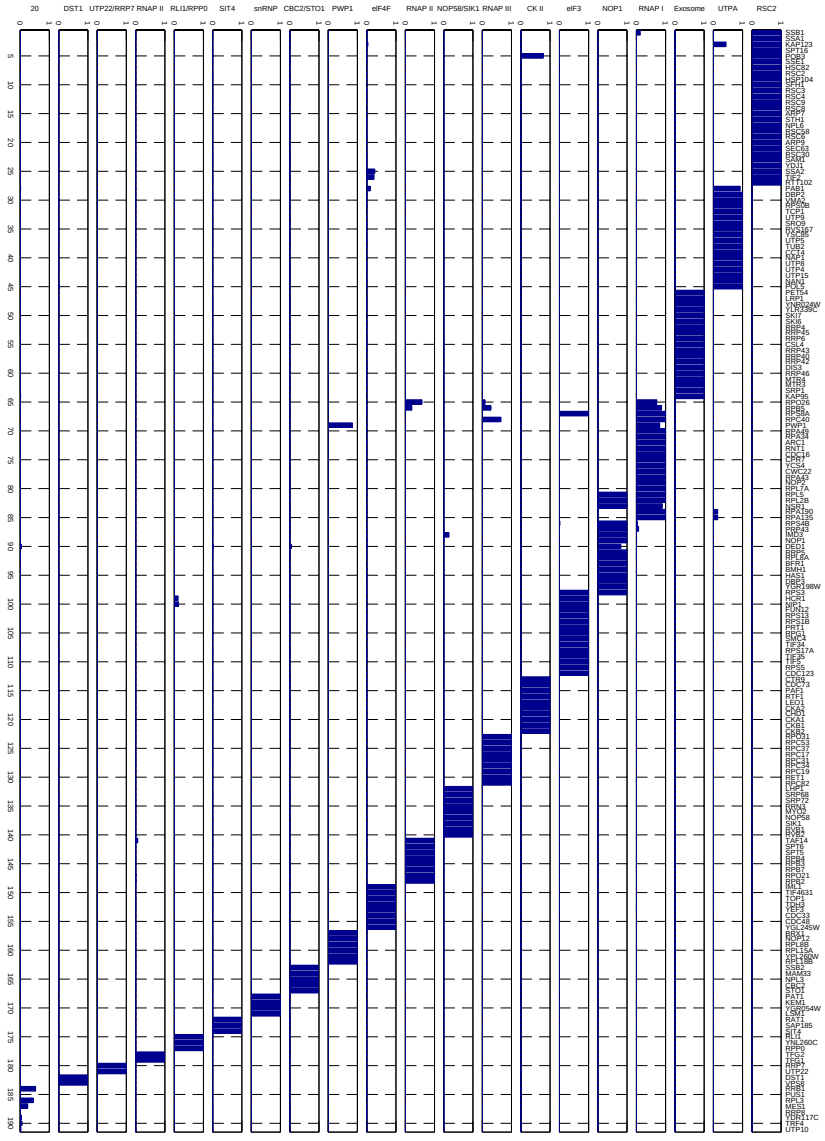


Figure 5. The assignments of the 191 proteins into 20 inferred complexes. The bars indicate the probability of membership of these proteins. Complexes identified (left to right) are: DST1; UTP22/RRP7; RNA Polymerase II; RLI1/RPP0; SIT4; U6 specific snRNP core; CBC2/STO1; PWP1/BRX1/NOP12; mRNA cap-binding/EIF4F; RNA Polymerase II; SRP3/NOP58/SIK1; RNA Polymerase III; Casein Kinase II; eIF3; NOP1; RNA Polymerase I; Exosome; SSU processome (UTPA); RSC2.