

A Complex-Based Reconstruction of the *S. cerevisiae* Interactome

Haidong Wang¹, Boyko Kakaradov^{1,*}, Sean R. Collins^{2,3,4,*}, Lena Karotki⁵, Dorothea Fiedler^{2,3,4}, Michael Shales^{2,3}, Kevan M. Shokat^{2,3,4}, Tobias Walther⁵, Nevan J. Krogan^{2,3}, Daphne Koller¹

¹Computer Science Department, Stanford University, Stanford, CA 94305, USA

²Department of Cellular and Molecular Pharmacology, University of California-San Francisco, San Francisco, CA 94158. USA

³The California Institute for Quantitative Biomedical Research

⁴Howard Hughes Medical Institute

⁵Max-Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

* These authors contributed equally

Corresponding authors:

Daphne Koller

Nevan Krogan

Email: koller@cs.stanford.edu

Email: krogan@cmp.ucsf.edu

Tel: (650) 723-6598

Tel: (415) 476-2980

Fax: (650) 725-1449

Fax: (415) 514-9736

Abbreviations:

EMAP: Epistatic Mini-Array Profile

EM: Expectation Maximization

GO: Gene Ontology

HAC: Hierarchical Agglomerative Clustering

HACO: Hierarchical Agglomerative Clustering with Overlap

JC: Jaccard Coefficient

PCC: Pearson Correlation Coefficient

PE: Purification Enrichment

TAP-MS: Tandem Affinity-Purification – Mass-Spec

TF: Transcription Factor

Y2H: Yeast 2-Hybrid

Summary

Most cellular processes are performed by proteomic units that interact with each other. These units are often stoichiometrically stable complexes comprised of several proteins. To obtain a faithful view of the protein interactome we must view it in terms of these basic units (complexes and proteins) and the interactions between them. This paper makes two contributions towards this goal. First, it provides a new algorithm for reconstruction of stable complexes from a variety of heterogeneous biological assays; our approach combines state-of-the-art machine learning methods with a novel hierarchical clustering algorithm that allows clusters to overlap. We demonstrate that our approach constructs over 40% more known complexes than other recent methods, and that the complexes it produces are more biologically coherent, even compared to the reference set. We provide experimental support for some of our novel predictions, identifying both a new complex involved in nutrient starvation and a new component of the eisosome complex. Second, we provide a high-accuracy algorithm for the novel problem of predicting transient interactions involving complexes. We show that our complex level network, which we call ComplexNet, provides novel insights regarding the protein-protein interaction network. In particular, we reinterpret the finding that “hubs” in the network are enriched for being essential, showing instead that essential proteins tend to be clustered together in essential complexes, and that these essential complexes tend to be large.

Introduction

Biological processes exhibit a hierarchical structure in which the basic working units, proteins, physically associate to form stoichiometrically stable complexes. Complexes interact with individual proteins or other complexes to form functional modules and pathways that carry out most cellular processes. Such higher level interactions are more transient than those within complexes and are highly dependent on temporal and spatial context. The function of each protein or complex depends on its interaction partners. Therefore, a faithful reconstruction of the entire set of complexes in the cell is essential to identifying the function of individual proteins and complexes, as well as serving as a building block for understanding the higher level organization of the cell, such as the interactions of complexes and proteins within cellular pathways. In this paper, we describe a novel method for reconstruction of complexes from a variety of biological assays, and a method for predicting the network of interactions relating these core cellular units – complexes and proteins.

Our reconstruction effort focuses on the yeast *Saccharomyces cerevisiae*. Yeast serves as the prototypical case study for the reconstruction of protein-protein interaction networks. Moreover,

the yeast complexes often have conserved orthologs in other organisms, including human, and are of interest in their own right. Several studies [1-4], using a variety of assays, have generated high-throughput data that directly measure protein-protein interactions. Most notably, two high-quality data sets [3, 4] used tandem affinity purification followed by mass-spectrometry (TAP-MS) to provide a proteome-wide measurement of protein complexes. These data provide the basis for attempting a comprehensive reconstruction of a large fraction of the protein complexes in this organism. Indeed, a number of works [5, 6] have attempted such a reconstruction. Generally speaking, all use the same general procedure: one or more data sources are used to estimate a set of affinities between pairs of proteins, essentially measuring the likelihood of that pair to participate together in a complex; these affinities induce a weighted graph, whose nodes are proteins and whose edges encode the affinities; a clustering algorithm is then used to construct complexes – sets of proteins that have high affinity in the graph. Although similar at a high level, the different methods differ significantly on the design choices made for the key steps in the process.

Recent works (since 2006) all focus on processing the proteome-wide TAP-MS data and using the results to define complexes. Gavin et al. [3], Collins et al. [7] and Hart et al. [5] all use probabilistic models that compare the number of interactions observed between proteins in the data versus the number expected in some null model. Collins et al. [7] and Hart et al. [5] both used all three of the available high-throughput data sets [2-4], in an attempt to provide a unified interaction network. The two unified networks resulting from these studies were shown to have large overlap and to achieve comparable agreement with the set of co-complex interactions in the MIPS data set [8], which are collated from previous small-scale studies. The interaction graphs resulting from the computed affinity scores are then clustered to produce a set of identified complexes. Gavin et al [3], Hart et al. [5], and Pu et al. [6] all use a Markov clustering [9] (MCL) procedure; Collins et al. [7] use a hierarchical agglomerative clustering (HAC) procedure, but do not suggest a computational procedure for using the resulting dendrogram to produce specific complex predictions.

Despite the fairly high quality of these networks and the agreement between them, they still contain many false positives and negatives. False negatives can arise, for example, from the difficulty in detecting interactions involving low-abundance proteins or membrane proteins; or from cases where the tag added to the bait protein during TAP-MS prevents binding of the bait to its interacting partners. False positives can arise, for example, from complexes that share components; or from the contaminants that bind to the bait non-specifically after cell lysis. Therefore, the set of complexes derived from the protein-protein interaction network alone has limited accuracy. Less than 20% of the MIPS complexes [8], which are derived from reliable

small-scale experiments, are exactly captured by the predictions of Pu et al. [6] or by those of Hart et al. [5].

In this study, we construct a method that generates a set of complexes with higher sensitivity and coverage by integrating multiple sources of data, including mRNA gene expression data, cellular localization, and yeast 2-hybrid data. The data integration approach was used in some early works on predicting protein-protein interactions [10, 11] and more recently by Qiu et al. [12], but these studies focus only on predicting pairs of proteins in the same complex, and not on reconstructing entire complexes. Many recent studies [13-21] have successfully integrated multiple types of data to predict functional linkage between proteins, constructing a graph whose pairwise affinity score summarizes the information from different sources of data. However, since the data integration is not trained toward predicting complexes, the high-affinity pairs contain transient binding partners, and even protein pairs that never interact directly but merely function in the same pathways. When these graphs are clustered, the clusters correspond to a variety of cellular entities, including pathways, functional modules, or co-expression clusters. We develop a data integration approach that is aimed directly at the problem of predicting stoichiometrically stable complexes.

We use a two phase automated procedure that we train on a new high-quality reference set that we generated from annotations in MIPS and SGD, and from manual curation of the literature. In the first phase, we use *boosting* [22], a state-of-the-art machine learning method, to train an affinity function that is specifically aimed at predicting whether two proteins are co-complexed. Unlike most other learning methods, boosting is capable of inducing useful features by combining different aspects of the raw data, making it particularly well-suited to a data-integration setting. Once we generate the learned affinity graph over pairs of proteins, we predict complexes by using a novel clustering algorithm called HACO (Hierarchical Agglomerative Clustering with Overlap). The HACO algorithm is a simple and elegant extension of HAC that addresses many of its limitations, such as the irreversible commitment to a possibly incorrect clustering decision. HACO can be applied to any setting where HAC is applied; given the enormous usefulness of HAC for the analysis of biological data sets of many different types (e.g., [7, 23, 24]), we believe that HACO may be applicable in a broad range of other tasks.

To validate our approach, we test the ability of our methods and other methods to predict reference complexes that were not used in training. By integrating multiple sources of data, we recover more reference complexes than other state-of-the-art methods [5, 6] when applied to the same set of yeast proteins. We also validated our predicted set of complexes against external data sources that are not used in the training. In all cases, our predictions are shown to be more coherent than other methods, and in many cases, more coherent even than the set of reference complexes.

A detailed examination of our predicted complexes suggests that many of them were previously known but not included in our (comprehensive) reference set, suggesting that our complexes form a valuable new set of reference complexes. In several cases, our predicted complexes were not previously characterized. We experimentally validated two of these predictions: a new component in the recently characterized eisosome complex [25], which marks the site of endocytosis in eukaryotes; and a newly characterized six-protein complex, including 4 phosphatases, which appears to be involved in the response to nutrient starvation, and which we name the NSC (Nutrient Starvation Complex).

The complex-based view provides a new perspective on the analysis and reconstruction of the protein interaction network. In the past, Jeong et al. [26] have suggested that the degree of a protein in an interaction network is positively correlated with its essentiality, and have argued that “hubs” in the network are more likely to be essential because they are involved in more interactions. Our analysis presents a complex-based alternative view: essential proteins tend to cluster together in essential complexes [5], and essential complexes tend to be large; thus, the essential “hubs” in the network are often members in large complexes comprised mostly of essential proteins. We also reformulate the task of reconstructing the protein interaction network. Rather than considering interactions between individual proteins [27-29] – a somewhat confusing network that confounds interactions within complexes and interactions between complexes – we tackle the novel task of predicting a comprehensive protein interaction network that involves both individual proteins and larger complexes. We argue that these entities are the right building blocks in reconstructing cellular processes, providing a view of cellular interaction networks that is both easier to interpret than the complex network of interactions between individual proteins, and more faithful to biological reality. Moreover, a complex, which is a stable collection of many proteins that act together, provides a more robust basis for predicting interactions, as we can combine signals for all its constituent proteins, reducing sensitivity to noise.

To accomplish this goal, we construct a reference set of complex-complex interactions, considering two complexes to interact if they are significantly enriched for reliable interactions between their components. We further augmented this set with a hand-curated list of established complex-complex interactions. We then use a machine learning approach to detect the “signature” of such interactions from a large set of assays that are likely to be indicative. We explore different machine learning methods, and show that a partially supervised naïve Bayes model, where we learn the model from both labeled and unlabeled interactions, provides the best performance. This model is applied both to our predicted complexes and to individual proteins, providing a new, comprehensive reconstruction of the *S. cerevisiae* interaction network, which can be downloaded from our project webpage [30]. We show that entities that are predicted to interact are more likely to share the same functional categories. A detailed investigation of our new predicted interactions

presents many that are established in the literature, as well as some that are novel but consistent, presenting plausible hypotheses for further investigation.

Experimental Procedures

Complex prediction

Constructing a set of reference complexes

We compiled a reference set of complexes by combining literature-derived results from small-scale experiments in MIPS [8] and SGD [31] with a hand-curated list (see our supporting website [32]) that we generated. The MIPS, SGD, and hand-curated set contain 225, 195, and 164 complexes respectively (Supplementary Figure S1(a)). Below we describe our method for establishing correspondence between the three lists and combining them into a high-confidence reference set suitable for training our method and for evaluating the accuracy of its predictions.

Our approach consisted of five processing steps. First, we merged similar complexes from the original lists (see below), resulting in a list of 543 complexes. Second, we removed 112 redundant complexes which were proper subsets of other complexes. Third, we removed the five largest complexes: the four ribosomal subunits and the small nucleolar ribonucleoprotein complex; these complexes are so large that they greatly overwhelm the signal, both in training the method and in evaluating the results. Fourth, we restricted the complexes to the set of 2195 proteins that have adequate amount of experimental evidence (see below). Finally, we removed single-protein complexes, arriving at the final list of 340 complexes. With at least 2 and on average 4.9 proteins per complex, this set of complexes contained 1100 unique proteins and a total of 1661 protein members, showing that the reference complexes contain notable overlap (proteins that are shared by multiple complexes).

In the first step of this merging process, we define each candidate complex from the three curated lists as a node in an undirected graph (or network). Two complexes are connected by an edge if they overlap significantly, i.e., their Jaccard similarity coefficient is greater than 0.7 (see JC metric below), with an edge weight equal to the JC value. We found 422 isolated nodes in the graph, corresponding to unique complexes that do not overlap significantly with any other complexes in the list. The task of merging similar complexes is equivalent to that of finding several types of connected components in this graph. A complete subgraph with average edge weight of 1 is equivalent to a group of complexes with identical protein content that appear under multiple names in at least two of the curated lists. We found 66 such groups, which correspond to complexes that we regard as very high-confidence because of multiple corroborating evidence. A complete subgraph in the rest of the network with average edge weight less than 1 (but greater than 0.7) is equivalent to a group of complexes whose protein contents are reported differently by the different

curated lists. We found 45 such groups and produced a consensus complex for each, resolving conflicts by a majority vote: a protein was included in the resulting complex only if it was found in more than half of the candidate complexes from the conflicted group. The remaining 18 nodes formed 4 connected components but no complete subgraphs, each component indicating non-transitive overlaps between three or more candidate complexes (e.g. A overlaps with B, and B overlaps with C, but A does not overlap significantly with C). Manual inspection and consultation with experts resulted in 10 unique complexes being added to the reference list. The distribution of complex sizes in our reference set is shown in Figures S2.

Constructing positive and negative co-complex protein pairs

The set of positive co-complexed protein pairs consists of all protein pairs that appear in the same complex in the reference set. For the negative set, we first consider all protein pairs (P_1, P_2) such that P_1 is in a reference complex and P_2 outside any version of that complex, in any of the three hand-curated set; we then exclude any pair that is within some other reference complex. The result of this process is 5065 positive pairs and about 1 million negative pairs.

Features for predicting co-complexed relations

We constructed features for our protein-protein interaction network using five different data sources: the purification enrichment (PE) score from the consolidated network of Collins et al. [7], a cellular component from a truncated version of the Gene Ontology (GO) [33], trans-membrane proteins [31], co-expression [34], and yeast two-hybrid (Y2H) interactions [35, 36].

Our highest-coverage source regarding protein-protein interaction comes from high-throughput TAP-MS data of the Gavin [3] and Krogan [4] data sets. The recent work of Collins et al. [7] provides a coherent and systematic way of integrating the data from these separate assays into a high-quality score that measures the probability of a protein pair to be co-complexed. The recent work of Hart et al. [5] provides a different integration method, but the results are quite similar, providing support for both of these procedures. We derived five features from the PE analysis: the direct score is computed based only on bait-prey information in the purifications; the indirect score is computed based on prey-prey information; the actual PE score is the sum of direct and indirect scores; the scaled score maps the PE score to a value between 0 and 1 to approximate the confidence value that the pair represents a true interaction; finally each protein is represented by a vector of its scaled PE scores with all the other proteins (where we assign its interaction with itself a score of 1), and we define our PE-distance feature as the cosine distance between the vectors of two proteins.

As the PE score provides most of the signals in predicting complexes (See Results section), we only kept the 2390 proteins that have at least one scaled PE score above 0.2 with some other

protein. Although this set only covers about 40% of the approximately 6000 yeast genes, it covers 81% of all protein members in the lists of high-quality complexes that comprised our reference set. As noted earlier, we exclude proteins that appear exclusively in the four ribosomal subunits and the small nucleolar ribonucleoprotein complex. This resulted in the final list of 2195 proteins, on which we performed our complex prediction.

Yeast two-hybrid assays also provide a direct measurement of protein-protein interactions. We derived these data from the assays of Ito et al. [35] and Uetz et al. [36]. Interacting pairs are assigned feature value 1. Pairs of proteins that appeared in the assay but not observed to interact are assigned feature value -1. All other pairs have 0 as their feature values.

The Gene Ontology (GO) cellular component hierarchy [33] was downloaded on June 25, 2007. An examination of the hierarchy showed that many of the smaller categories (lower in the hierarchy) refer to particular complexes whose information is derived from the same small-scale experiment that inform our reference set. Thus, in order to achieve a fair evaluation using the reference set, we remove categories of size less than 120 that can potentially contain the answer. The remaining 44 out of 564 categories represent high-level cellular localization information, much of which is obtained through high-throughput experiments [37]. Some sample categories include "endoplasmic reticulum part", "nuclear chromosome part", "mitochondrial membrane", and "cytoplasm".

We derived two pairwise localization features from the GO cellular component. One is the semantic distance measure [38], which is the log size of the smallest category that contains both proteins. However, this feature is a pessimistic assessment regarding the co-localization of the two proteins, as lack of annotation of a protein in some category, particularly one that is a subset of its most specific category, does not necessarily mean that it cannot belong to this category. Therefore, we construct a second feature, which is the log size of the smallest possible group that could contain both proteins (given the current evidence). It is computed in the following way between protein A and protein B, whose most specific categories are X and Y respectively. If X is a sub-category of Y, then the two proteins might belong together to any group if they were to be annotated with enough detail. Therefore, we use log of 120, the size of the smallest category, as our second feature. On the other hand, if X and Y are not sub-categories of each other, we denote Z to be the smallest common super-category of X and Y. We then denote X' (resp. Y') to be the category one level down the path from Z to X (resp. Y). Thus, assuming that A and B belong to the two different categories at X' and Y', the smallest semantic category that we can form that may contain them both is $X' \cup Y'$. Thus, our second feature is $\log(|X' \cup Y'|)$.

A list of membrane proteins are obtained by parsing the trans-membrane annotations in SGD [31]. A pair of proteins is considered to be membrane if at least one of the proteins is found in the

membrane. The first membrane feature is 1 if the pair is membrane and 0 otherwise. The second and third features are the product of the first feature with the direct and indirect PE score of the two proteins, respectively. This allows our boosting model to take into account the known fact that TAP-MS purifications work differently on membrane proteins from non-membrane proteins.

Microarray data were downloaded from Stanford Microarray Database (SMD) [34] on Dec. 5th, 2006, which contains a total of 902 experiments for Yeast divided into 19 categories. The data were normalized to mean 0 and standard deviation 1. We construct a feature by computing the mean-centered Pearson correlation coefficient between the expression profiles of two proteins.

A final feature is obtained from small-scale physical interactions. We downloaded protein-protein interactions from MIPS [8] and DIP [39] on 21 March 2006. We extracted from MIPS those physical interactions that are non-high-throughput yeast two-hybrid or affinity chromatography. For DIP, we picked non-genetic interactions that are derived from small-scale experiments or verified by multiple experiments. This feature has value 1 for observed interactions, and feature value 0 for all other pairs. Importantly, there is a risk of cyclicity between these small-scale interactions and the reference complexes. Therefore, to avoid a positive bias in our results, we omitted this feature in the cross-validation runs, which are evaluated against the reference complexes. For those runs that are trained on the entire set of reference complexes, this cyclicity is not a concern, so this feature was included.

There are a total of 12 features for cross-validation runs and 13 features for runs that are trained on the entire reference set.

Integrating multiple features using the logit-boost

Boosting [22] is a class of algorithms that iteratively combines weak learners to give a representative ensemble. Each weak learner is a simple classifier, such as a decision stump, that may only weakly correlate with the labels. After a weak learner is trained, we add it to the ensemble with appropriate weight. In the next iteration, the algorithm puts more weights on the data points that are classified incorrectly by the current ensemble, which the next weak learner will focus on. Boosting is able to perform automatic feature selection and has better or comparable accuracy with other state-of-the-art classifiers such as support vector machines (SVMs) [40] in many domains. We implemented a version of boosting algorithms called LogitBoost [22] that uses decision stumps as weak learners and the logit function as the loss function. This variant is shown to be more robust to outliers and overfitting than the standard AdaBoost variant [41]. Our experiments (data not shown) showed that this method performs well on our data, compared to other versions of Boosting and other classification algorithms such as logistic regression and SVMs. The prediction of the learned ensemble classifier on a given protein pair is taken to be the pair's

affinity in the clustering algorithm below.

The Hierarchical Agglomerative Clustering with Overlap (HACO) Algorithm

The standard HAC algorithm with average linkage [42], maintains a pool of merging candidate sets where the distance between two non-overlapping sets is:

$$d(A, B) = \frac{1}{|A| |B|} \sum_{P \in A, Q \in B} d(P, Q)$$

In our setting, we take $d(P, Q)$ as the negative of the affinity between protein P and protein Q . Note that $d(A, B)$ is the average of the edge distance between proteins in A and proteins in B .

In HAC, at each step, we pick the two non-overlapping sets with the closest distance, A and B , and merge them to create a new set, M . M is added to the pool, while the sets A and B are removed. Therefore in later steps, we could only consider the superset M , and would never be able to use A or B again to merge with some other set. Assume that there is another set C whose distance to A is only slightly larger than $d(A, B)$. In this case, the decision to merge A with B rather than with C is arbitrary and unstable. When the actual clusters overlap, a more appropriate solution would be to have two *overlapping* merged candidates: $M = A \cup B$ and $N = A \cup C$. We adapt HAC to accommodate this intuition. We define the *divergence* between A and M as a measure of the cohesiveness of the set M is outside of A (Figure S3):

$$\text{divergence}(A, M) = \frac{1}{|E|} \sum_{(P, Q) \in E} d(P, Q)$$

where E is the set of pairs in M , but not in A : $E = \{(P, Q) \mid (P, Q) \in M \times M - A \times A, P < Q\}$. (Here, ' $<$ ' can be any ordering among the proteins, such as alphabetical, to avoid a pair appearing twice in the set E).

If M is not overlapping with C , we have the choice of whether to use A or M to merge with C . If $\text{divergence}(A, N) - \text{divergence}(A, M)$ is small, it makes sense to merge A and C to create a new set N that is almost as coherent as M . On the other hand, if the difference is large, we would prefer to replace A with its superset M as the merging candidate to C .

In practice, we use $d(A, C)$ to approximate $\text{divergence}(A, N)$: we check whether $\Delta = d(A, C) - \text{divergence}(A, M)$ is small. $\text{divergence}(A, N)$ is the weighted average of $d(A, C)$ and $d(C)$, the distance within C . $d(C)$ tends to be smaller than $d(A, C)$ because pairs within C , which is formed earlier by some merging, are more coherent than pairs between A and C . Therefore, $d(A, C)$ tends

to be smaller than $\text{divergence}(A, N)$, so keeping Δ small is generally a more stringent requirement for ensuring that N is almost as coherent as M . Moreover, by forcing $d(A, C)$ to be small, we make sure the set N is coherent not just because the distance within C is small. With this consideration, we defined the modified distance between A and C to be (Figure S3):

$$d'(A, C) = \begin{cases} d(A, C) & \text{if } \Delta < \rho \\ \infty & \text{if } \Delta \geq \rho \end{cases}$$

The modified distance d' is used to pick the two closest sets to merge in the next iteration. If Δ is smaller than a margin, we make d' equal to d and thus allow A and C to merge. On the other hand, if Δ is large, we make d' infinity and thus prohibit A and C from merging, in favor of merging their supersets. ρ is the *margin* parameter: the larger the margin ρ , the more likely a set A is to be re-used, resulting in more overlapping subsets constructed by the algorithm. If the margin is 0, it reduces to the standard HAC. Therefore, our HACO algorithm is a generalization of the HAC. Note that we can eliminate a set from the merging candidate pool when its modified distances to all other sets are ∞ . Of course we can define other modified distance as long as it is larger when Δ is large and close to $d(A, C)$ when Δ is small.

In practice, A might have multiple supersets in the pool. Therefore, we look at all of A 's supersets in the pool that are not overlapping with C and use the set $M_{A,C}$ with smallest divergence from A , i.e., the one that provides the best replacement for A in terms of the proposed merger with C :

$$M_{A,C} = \arg \min_{M \text{ s.t. } A \subseteq M, C \cap M = \emptyset} \text{divergence}(A, M)$$

We do the same thing with C for its proposed merger with A :

$$M_{C,A} = \arg \min_{M \text{ s.t. } C \subseteq M, A \cap M = \emptyset} \text{divergence}(C, M)$$

The smaller of $\text{divergence}(A, M_{A,C})$ and $\text{divergence}(C, M_{C,A})$ is used to compute the modified distance.

The algorithm terminates when there are no more non-overlapping sets to merge. The output is a cluster-lattice, where the same cluster can be a child of multiple parents in the lattice. The lattice is cut at a certain threshold to generate a set of overlapping clusters. These predicted clusters are the sets that are still in the candidate pool when the distance in the merging process reaches the threshold.

Training and Test Regime

To evaluate our prediction accuracy against the reference set, we divide the 340 reference complexes into five disjoint subsets, or folds. As there are about a million negative pairs, for computational expediency, we randomly sample one tenth of the negative pairs to be used in training while setting each negative pair to have ten times the weight of the positive pairs.

For each fold in the five-fold cross validation, we hide one set and use the remaining four sets to train the affinity function for the protein pairs, the margin ρ for the HACO, and the cutoff threshold for the resulting cluster-lattice. We use the same training set in all steps of our pipeline and evaluate the final predictions on complexes in a separate test set which is hidden during all steps of the training process. We select the cutoff threshold by maximizing the coverage (see below for the definition) on the training set. To pick the margin ρ , we cannot use coverage alone because our model would always prefer a bigger margin that keeps more sets in the pool. Therefore, we choose ρ by maximizing the product of coverage and sensitivity (see below for the definition) on the training set. This approach trades off between the match with the reference set and the number of predicted complexes.

To evaluate our predictions against external data sources, such as biological coherence and essentiality, we augment our model with a feature constructed from small-scale physical interactions and train it on the entire set of 340 reference complexes. To avoid circularity between features and evaluation, we do not evaluate the predictions from such runs against the reference complexes.

Evaluation metrics for matching between predictions and reference complexes

The overlap between a reference complex R and a predicted complex C can be quantified in several ways [43] (figure S4):

Jaccard coefficient (JC): $|R \cap C| / |R \cup C|$

Hamming distance: $|R \cup C| - |R \cap C|$

We use both measures because of the size effect. For example, a Hamming distance of 2 between two large complexes, say both of size 5, is a good match. In this case $JC = 4 / 6 = 0.67$. On the other hand, a Hamming distance of 2 between two small complexes of size 2 implies an overlap of only one protein, which could arise simply by chance. In this case $JC = 1 / 3 = 0.33$.

We define the coverage and sensitivity of a set of predictions so we can systematically evaluate genome-wide predictions. For each reference complex, we find the prediction that has the highest Jaccard coefficient. We define the scaled Jaccard coefficient: $SJC(R, C) = \max\{0, 2JC(R, C) - 1\}$. We truncate the value at 0 because it may represent random overlap. In the above examples, the

matching of the two large complexes of size 5 and Hamming distance 2 would have $SJC = 0.33$ while the small ones of size 2 and Hamming distance 2 would have $SJC = 0$. We define the coverage as the average Jaccard coefficient per reference complex:

$$\frac{1}{m} \sum_{i=1}^m \max_{j=1}^n SJC (R_i, C_j)$$

where m is the number of reference complexes and n is the number of predicted complexes.

For sensitivity, we sum the Jaccard coefficients of all the overlapping (reference, prediction) complex pairs, and normalize by the total number of predicted complexes:

$$\frac{1}{n} \sum_{i=1}^m \sum_{j=1}^n SJC (R_i, C_j)$$

Biological coherence of predicted complexes

We evaluate biological coherence of the predicted complexes using several metrics. The first is average semantic distance in the GO Biological Process hierarchy. GO data were downloaded on 25 June 2007. We compute the distance between two proteins as the log size of their smallest common category [38], as for the cellular component hierarchy above.

We downloaded the protein expression data from [44]. We use log of measured protein levels in terms of molecules per cell as the protein abundance value.

The growth phenotype data was obtained from [45]. For each gene, its homozygous deletion strain is grown in 418 experiments with different drug treatments. The log ratio of the deletion strain's growth in no-drug control to its growth with the drug treatment is used to define the growth phenotype in that particular condition. For each pair of genes, we compute the Pearson correlation of the growth phenotypes across all 418 conditions, which is the measure used in the original paper.

We downloaded the transcriptional regulation data from [46, 47]. We used p-value cutoff at 0.001 and required conservation across species to define the transcription factors for each protein. We computed how many transcription factors are shared by any two proteins.

Complex-complex and complex-protein interaction prediction

Constructing a reference list of positive and negative complex-complex interactions

We derived a reliable set of *S. cerevisiae* protein-protein interactions from MIPS [8] and DIP [39], downloaded on March 21st, 2006. We extracted from MIPS those physical interactions that are non-high-throughput yeast two-hybrid or affinity chromatography. For DIP, we picked

non-genetic interactions that are derived from small-scale experiments or verified by multiple experiments. We compute the number of reliable interactions between proteins of two complexes, and compared it to what we expect if the reliable interactions are distributed randomly. We define the two complexes to be interacting if the enrichment of reliable interactions is more than 20 standard deviations above the mean. Such strong enrichment is needed because the reliable interactions are very sparse and the presence of even a very small number would result in a large deviation from the mean (e.g.: for two complexes of size 2 and 5 respectively, we only need 1 reliable interaction out of the total 10 pairs to get an enrichment of 10 standard deviations above the mean). We ended up with a list of 82 interactions between the set of 383 complexes we just predicted. To augment this list, we generated a list of 59 additional known interactions between 81 named complexes. To avoid the redundancy between those 81 named complexes and our 383 predicted complexes, we replace a predicted complex by a named complex if they overlap with $JC > 0.5$. This process gives us a total of 421 complexes with 133 unique interactions between them, which is used as our positive reference set. We create a negative reference set of 3173 non-interactions by using all pairs of named complexes that are not in our positive set. The interaction status of all the remaining pairs of complexes, named or predicted, is treated as unknown.

For protein-complex interactions, in addition to the above negative set between complexes, we randomly sampled 6560 protein-complex pairs that are not in the positive set and added them to our negative reference set. The number 6560 is chosen so the ratio of positive to negative pairs for protein-complex interactions is the same as the ratio for complex-complex interactions.

All our reference lists are available from our supporting website [32].

Features for predicting interactions

Since there is no direct measurement of complex-complex or complex-protein interactions, we try to use as much indirect evidence as possible. Besides all data sources used for identifying complexes, we added four additional data sources based on correlation of growth fitness, correlation of transcription factor profile, protein-protein interaction prediction, and condition specific expression correlation.

The correlation of growth fitness profile [45] is computed as described in the previous section regarding ‘Biological coherence of predicted complexes’

For each protein, we create a transcription factor (TF) profile vector, where each position in the vector represents a TF and its value is 1 if the TF is found to regulate the protein [46] and 0 if it is not. We used the same transcription regulation data as described in the previous section regarding

‘Biological coherence of predicted complexes’. For any pair of proteins, we compute the mutual information between the profile vectors of the two proteins using the method described in Date et al. [48].

There are many works in integrating multiple sources of data to predict protein-protein interactions. In particular, the InSite method [49] integrates protein sequence motifs, evidence for protein-protein interactions, and evidence for motif-motif interactions in a principled probabilistic framework to make high-quality predictions of protein-protein interactions. Here, we use the InSite method, but trained without the reliable interactions between complexes in our positive reference set. We use the predicted probabilities that two proteins interact as one more data source.

Here, we processed the expression data in accordance with our intuition that transient interactions occur under specific conditions, and we should only expect expression profiles of interacting proteins to be correlated only when at least one of the pair is active. Specifically, we divided our expression data into 76 conditions [50-58], each of which represents a particular time course. In accordance with convention, we quantify a protein’s activity under certain condition according to its maximum deviation from norm, or in other words the maximum absolute expression (assuming norm to be 0). For each condition, we define a protein to be differentially expressed, or active, if its maximum absolute expression is above a cutoff, which we specify to be 1.0. For each pair of proteins, we compute Pearson Correlation Coefficient (PCC) separately in each condition. If a protein in the pair is inactive under a condition, the PCC value for the condition is assumed to be 0. We use the PCC value, averaged across all conditions under which at least one protein out of the pair is active, as our last feature type. Initial investigation showed that this feature is better correlated with the reference complex-complex interactions than the overall PCC across all conditions. We note that, for the task of predicting when two proteins are co-complexed, the simple correlation performed better (data not shown), consistent with the fact that the activity of two members of a stable complex is likely to be similar across a wide range of conditions.

Aggregating signals between proteins into features between complexes

All forms of evidence in our analysis involve a pair of proteins. To predict interactions between two complexes, C and D, we aggregate the signals for all protein pairs between C and D and produce the following features:

$$f_{ij} = A_i(\{S_j(P, Q) \mid P \in C, Q \in D\})$$

where $A_i()$ is some aggregating function, such as: sum, max, mean, min, decayed max, decayed min, etc. (See Supplementary Table 1 for a complete list of aggregating functions and their definitions.)

S_j represents the j^{th} feature type between a pair of proteins. We also use four global features, independent of the data sources: size of the first complex, size of the second complex, number of protein pairs between the two complexes, and number of overlapping proteins between the two complexes. The features for interactions between a protein P and a complex C are identical, except that we only need to aggregate the signals over all pairs (P,Q) for Q in C.

The Naïve Bayes model that we use assumes all features to be conditionally independent of each other given the status of whether two complexes interact or not. Therefore for each data source, we pick only the best aggregating function in order to reduce the conditional dependencies between the features. To do this, we define r_{ij} to be the area under the ROC curve if we use the feature f_{ij} alone to predict complex-complex interactions: the greater r_{ij} , the stronger the correlation between the feature and the presence of a complex-complex interaction. Therefore, for Naïve Bayes, we use, for each feature type j , the feature $f_j = f_{ij}$ where i gives rise to the maximum value r_{ij} . Supplementary Table 2 lists the aggregating function chosen for each feature type.

Learning and predictions

We experimented with different machine learning algorithms for making our predictions: (1) a simple Naïve Bayes model, where the effects of different feature types are assumed to be independent; (2) a discriminative boosting algorithm, as we used in predicting co-complexed affinities between protein pairs above; (3) a Naïve Bayes model where the unlabeled complex-complex interactions are taken to be unobserved variables, and the model is trained via the Expectation Maximization (EM) algorithm. This last approach is based on the fact that the amount of labeled training data is quite limited in this task, but the unlabeled data also provides us with useful information about the behavior of different features in interacting and non-interacting pairs. A variant of this same approach was used with success in the InSite model [49].

More formally, for each pair of complexes, we construct an “interaction variable”, whose value is 1 if the two complexes are in the positive reference set of interacting complexes, 0 if they are in the negative reference set, and unobserved otherwise. Each feature of the complex pair is associated with two conditional distributions: one for the case of an interacting and the other for the case of a non-interacting pair. These distributions are defined via some parametric class (see Supplementary Table 3). The distributions for the different features are taken to be independent of each other within each of the two cases. The model is trained via the following EM procedure. We initialize the model parameters to those that would be obtained from maximum likelihood estimation (MLE) using the pairs in our reference set alone. We then iteratively repeat the following two steps until convergence. In the E-step, we use our current model to compute the marginal probability of each unobserved interaction variable given the features associated with the pair. We use the computed probability as a soft assignment to the interaction variable. In the

M-step, we learn the parameters for the distributions using MLE based on the inferred soft assignment to all interaction variables; the variables in the reference set are always fixed to their known value. We use the model obtained at convergence to predict, for each pair of complexes not in our reference set, the probability with which the pair interacts.

We used the same naïve Bayes + EM procedure when making predictions using only one of the features (PE score or Insite probability), which we used as a comparison baseline. In these comparisons, we used the same aggregator selected for the model using all the features.

When training using the LogitBoost model, we are not making independence assumptions between the different features. Hence, there we include all features f_{ij} , instead of just picking the best aggregating function for each feature type.

We used the same naïve Bayes + EM procedure for the protein-complex interaction predictions, although the best aggregating functions picked and the set of parametric classes used for the feature distributions was a little different. (See Supplementary Tables 2 and 4.)

Functional coherence of complexes predicted to interact

We evaluate whether two interacting complexes are more likely to share the same functional category. We used functional categories from MIPS [8], which has 18 functional categories with average 684 proteins per category. A complex is assigned to a particular functional category if more than half of its components belong to the functional category. We only perform our evaluation on complex pairs where both complexes are assigned to some MIPS functional category.

Experimental Validation

TAP Purification

2l of yeast culture expressing Pil1-TAP was grown to $OD_{600} = 0.8$ and subsequently harvested. The resulting pellet was resuspended in 8ml buffer A (150 mM K Ac, 20mM HEPES pH =7.4, 2mM Mg Ac, 5% glycerol) and frozen in liquid nitrogen. Total proteins were extracted by bead milling of the frozen pellet, followed by addition Triton-X100 to 1% (w/v) final concentration. Solubilized extracts were cleared by two centrifugations of 4min, 4000g and incubated with IgG sepharose for 2hrs. Beads were washed 6 times with 50 ml of buffer B (150 mM K Ac, 20mM HEPES pH=7.4, 2mM Mg Ac, 5% glycerol, 1% Triton-X100). Proteins were eluted by TEV protease cleavage in 200 μ l buffer for 2hrs and analyzed on a SDS-Page gel by Coomassie staining. Bands were cut, digested with trypsin, peptides were extracted and analyzed by LC-MS/MS as previously described [59].

EMAP Experiments

EMAP experiments and subsequent data analysis were done as previously described [60, 61]. Data from these experiments are presented on our supporting website [32].

Results

Method Overview

We compiled a reference set of complexes by combining literature-derived results from small-scale experiments in MIPS [8] and SGD [31] with a hand-curated list (see our supporting website [32]) that we generated. The MIPS, SGD, and hand-curated set contain 225, 195, and 164 complexes respectively (Supplementary Figure S1(a)). We established correspondence between the three lists and combined them into a high-confidence reference set suitable for training our method and for evaluating the accuracy of its predictions (see Methods). This curated set was compiled prior to the development of our method, and was not subsequently revised.

We then formulated the task of predicting whether two proteins were members of the same complex as a machine learning task. We used our reference set to construct a high-quality set of positive and negative examples. We constructed features that are useful for predicting this relationship from five different data sources: the purification enrichment (PE) score from the consolidated network of Collins et al. [7], a cellular component from a truncated version of the Gene Ontology (GO) [33], trans-membrane proteins [31], co-expression [34], and yeast two-hybrid (Y2H) interactions [35, 36] (see Methods). We then applied the Boosting algorithm [22] for training the predictor. Boosting was selected due to its high accuracy, robustness to outliers, and ability to perform automatic feature selection. The prediction of the boosting classifier on a given protein pair is taken to be the pair's affinity in the clustering algorithm below.

Our initial experiments showed that hierarchical agglomerative clustering (HAC), which progressively merges sets of proteins with strongest affinity, produces the best results for complex reconstruction if trained to optimize for that task. However, HAC has several significant limitations. First, it does not allow clusters to overlap, whereas actual complexes do share subunits. Second, it uses a single cutoff to decide the granularity of the complexes constructed. A cluster near the cutoff in the dendrogram can be formed even if it is the result of merging two relatively weakly connected sub-clusters A and B. Such a cluster, although of lower confidence, still excludes both A and B from being predicted as a complex; this occurs even if A and B are strong candidates for being a complex. Finally, once a set of proteins is merged with another set, it cannot merge with anything else even if the affinity is only slightly lower. Therefore an incorrect decision cannot be fixed later in the process.

To address these limitations, we constructed a novel clustering algorithm called HACO (HAC with Overlap) that allows a set of proteins to be merged with multiple other sets with which it has comparably strong affinity (see Methods). HACO addresses all of the limitations above: It produces clusters that can overlap. Second, when merging A and B into a single cluster C, it also has the option of leaving A and/or B as candidate complexes, avoiding a wrong decision because of an arbitrary cutoff. Finally, as it allows the same cluster to be used in multiple places, it avoids many mistakes that arise from an almost-arbitrary breaking of near-ties. Both our boosting algorithm and the HACO code are made freely available on our project webpage [32], allowing them to be used for predicting complexes with other forms of data.

Complex Predictions

Coverage and sensitivity of predicted complexes

We compiled a reference set of complexes from MIPS [8], SGD [31], and hand-curation (see our supporting website [32]), which is more comprehensive than previous studies [5, 6]. Although it still contains noise and bias, it provides us with the ultimate evaluation of our predictions. There are 340 complexes in our reference set with an average of 4.9 proteins per complex (Supplementary Figure S1(b)).

To predict complexes, we first trained our model to predict pairwise co-complex interactions, and then used our HACO algorithm to cluster the resulting pairwise affinity network into complexes. We constructed features for our protein-protein interaction network using five different data sources: the purification enrichment (PE) score from the consolidated network of Collins et al. [7], a cellular component from a truncated version of the Gene Ontology (GO) [33], trans-membrane proteins [31], co-expression [34], and yeast two-hybrid (Y2H) interactions [35, 36]. We tested our approach using a standard 5-fold cross-validation regime, training on 80% of the complexes and testing on the remaining 20%; the test set was not used in any aspect of the training of the model. For each fold in the five-fold cross-validation, we apply HACO to the affinity measure learned using the boosting model on the training data. We evaluate the resulting clusters on the hidden test set. We predicted 417.8 complexes per fold with at least two proteins for each complex. Each complex contains 4.30 proteins on average (Supplementary Fig S1).

We define a complex to be well-predicted if it is within Hamming distance (see Methods) of 2 to some predicted complex. However two small complexes can be quite different even if their Hamming distance is 2. Therefore we also require the Jaccard coefficient (see Methods), which takes into account the size of the complexes, to be above 0.5. We also measure the coverage and sensitivity of the set of predictions (see Methods): coverage measures how well the reference set is covered by our predictions and sensitivity measures how well each predicted complex overlaps with the reference set, a measure that takes into consideration the number of predicted complexes.

We compared our results to those of Bader [62], Gavin [3], Krogan [4], Hart [5], and Pu [6]. As we discussed, each method made different decisions for defining the affinity function and for clustering it. Bader et al. used a novel clustering algorithm called Molecular Complex Detection (MCODE) to detect densely connected regions in the protein-protein interaction network. Gavin et al. computed a socio-affinity score between each pair of proteins that compares the number of times the two proteins are observed together in some purifications relative to what is expected by chance. The pairwise network of socio-affinity scores are then subjected to a procedure that produces overlapping clusters. Complexes are composed of a “core” that appears in most runs of the clustering algorithm, and “attachments” that appear only in some. Most of the recent methods appear to have converged on using the MCL algorithm [9], albeit on different affinity functions: Krogan et al. used a machine learning approach, trained on MIPS reference complexes, to predict the confidence that a pair of proteins is in the same complex. Hart et al. defined a p-value by comparing observed relative to expected number of interactions, applied to three sets of purifications [2-4]. Pu et al. applied MCL directly to the purification enrichment (PE) score of Collins et al. [7]. All of these MCL-based methods produce non-overlapping clusters, although the method of Pu et al. used a post-processing phase to identify proteins that are likely to be recruited by multiple complexes.

Figure 1a shows the accuracy of our method in reconstructing the reference complexes, as compared to the best of these other approaches. As we can see, our affinity score achieves significantly better results than any of these methods; the results are better even when we use simple HAC for the clustering, and improve further when we use HACO. We note that Hart et al. and Pu et al. are the state-of-the-art in complex predictions and have been extensively compared with other complex prediction methods. In particular, Pu et al. applied MCL to the same set of PE scores [7] as we used. HACO was able to perfectly recover 42% and 46% more reference complexes compared with the Hart et al. and Pu et al. respectively (p-values < 0.01). The corresponding increase in sensitivity is 6% and 29% respectively and increase in coverage is 28% and 33% respectively. The results suggest that these improvements are a consequence of our use of data integration with state-of-the-art machine learning. In particular, the Pu method and the Hart method, both of which used MCL applied to different affinities obtained from the TAP-MS data, performed very similarly. Interestingly, HAC applied to the PE score performed slightly better than MCL applied to the PE score (HAC-PE vs. Pu). These three methods performed better than those of Bader et al., Gavin et al., and Krogan et al., likely due to the fact that these earlier methods used only a single set of purifications. These results demonstrate the importance of combining data from multiple data sources, integrated appropriately. We note that MIPS complexes are used, albeit in a very limited way, in generating the PE score. To avoid any risk of circular reasoning, we run the same experiments using the SGD complexes alone as an independent reference set; the

results (Supplementary Figure S2a) show that the improvement of our method over others remains consistent in this reference set as well.

The HACO algorithm helps address several of the limitations of the HAC approach. First, it reduces the sensitivity of the complex definitions to a single universal threshold in the hierarchy. One such example involves the 15-protein SAGA complex. Here, HAC predicts a 24-protein superset of the SAGA complex. This cluster is a much weaker cluster than SAGA itself: the average affinity between the SAGA proteins is 0.35, as compared to the average affinity, -1.19, for pairs within the 23 proteins excluding pairs of SAGA proteins. By comparison, HACO, by keeping multiple hypotheses relative to the cutoff, predicted both a 23 protein cluster (similar to the HAC prediction), but also predicted the subcluster that corresponds perfectly to the SAGA complex. The second limitation addressed by HACO is that it avoids an early commitment to incorrect outcomes. For example, the affinity between Rad23 and Png1 is slightly higher than that between Rad23 and Rad4. HAC incorrectly merges Rad23 and Png1, and now cannot reuse Rad23 in any other complex. HACO can reuse Rad23, merging it with Rad4 to create a complex that perfectly matches the NEF2 (nucleotide-excision repair factor 2) complex in the reference set.

Contribution of each data source for predicting complexes

Given the importance of data integration, it is useful to see which data sources play the most important role in our results. We first considered the contribution of each feature to our learned affinity function. Our approach uses LogitBoost [22], which defines the affinity function as the weighted sum of many weak learners, each of which is a decision stump on one of the features. The top weak learners involve features that are deemed to be most predictive. The top features in the order of their importance are: correlation of PE score (weight 3.84); semantic distance in the truncated GO cellular component categories (-2.2); direct PE score, which is based only on direct bait-prey interactions (0.58); small-scale physical interactions (0.55); and co-expression (0.16). It is interesting to note that the correlation of the PE score is deemed more informative than the PE score itself. One explanation is that the pairwise PE score between proteins P and Q is still a noisy measure for co-complexness, but if P and Q are truly co-complexed, they are likely to have similar interactions with other proteins.

As another metric for assessing the importance of each data source to the quality of our predictions, we successively applied our pipeline with HAC to the data source alone and to all data sources except that data source (Figures 2, S2). The PE score plays the dominant role and by itself predicts most of the complexes. Importantly, our method here combines different variations of PE score (direct, indirect, scaled, total, and correlation) using boosting, generating an affinity score that is quite a bit better at predicting complexes than the original scaled PE score (73/54/16 perfect matches/1-away/2-aways for HAC-PE in Figure 1a versus 81/50/19 for the PE-based features alone

in Figure 2). This result demonstrates the value of applying machine learning methods specifically optimized for the problem of complex identification. Nevertheless, we still get a significant improvement by integrating in other data sources.

Localization and expression have a similar effect. By itself, neither predicts any complexes at all; this is not surprising, as both are features with low precision. However, removing each of them decreases the accuracy, suggesting that they provide a signal that is independent of the PE score, and can help resolve some of its ambiguities and errors. The yeast two-hybrid feature has the opposite behavior: In isolation, it predicts a reasonable number of complexes; however, removing it does not decrease accuracy at all. This behavior can be explained by the hypothesis that yeast two-hybrid data largely correlates with PE score; thus, although it is predictive, it does not add much given the PE score data. This last hypothesis is further verified by the fact that localization and expression features appear within the top 5 weak learners whereas yeast two-hybrid feature does not.

Biological coherence of predicted complexes

Having tested the ability of our method to reconstruct reference complexes, we produced a final set of predictions from our method. Here, we train on all reference complexes, and introduce an additional feature relating to interaction in small-scale experiments; this feature was not used in the comparison to reference complexes to avoid potential circularity between this feature and the definition of the reference complexes. Overall, this process resulted in 383 predicted complexes, which can be found on our supporting website [32]. We evaluated the validity of these complexes by comparing to external data sources not used in the training and not directly related to reference complexes. For all biological coherence validations, we compute the coherence for each complex as the average of the coherence measure for all pairs in the complex. Then, we take the average across all complexes predicted. We compare to the methods of Hart et al. [5], and Pu et al. [6], which consistently out-performed all previous methods. As a different benchmark, we also compare to the coherence for the highest-affinity protein pairs (those that are most likely to belong to the same complex).

We validate our predictions by looking at various measures of biological coherence (Figure 3): similarity of GO biological process; similarity in the level of protein abundance for different complex components; correlation of growth defect profiles across a broad range of conditions; and co-regulation, as measured by sharing of transcription factors. For all measures, HACO with our affinity function considerably outperformed all other approaches, with the method of Hart being the closest competitor. Most striking were the improvements in correlation of growth phenotypes across multiple conditions and in coherence of the transcriptional regulation program. To specifically test our complex formation process, we also compared pairs of co-complexed proteins

with pairs that have high affinity (as computed by our boosting algorithm). The results were largely comparable, with the notable exception of protein abundance, where our complexes are 12% less coherent than the top affinity pairs; this suggests that proteins with lower affinity scores can be members of the complex, but also play other roles in the cell, reducing their correlation with other proteins in the same complex. The comparison to the reference complexes is also interesting: Our complexes are considerably more coherent than the reference complexes on regulator overlap, and perform similarly on correlation of abundance and growth phenotype. Conversely, our complexes are significantly less coherent than the reference complexes on GO biological process annotations; this is not surprising, as the reference complexes and GO annotations are derived (at least partly) from similar data sources, such as literature and small scale experiments. Overall, when comparing to data sources that were not used in constructing the reference complexes, our predictions seem to perform as well or better than the reference set, suggesting that our predictions provide a strong set of complexes that can be used as a new reference.

In-depth study of predicted complexes

We also did a systematic, manual evaluation of many of our predicted complexes. We first considered the complexes that were 1-away from the reference set, that is, a protein P and a complex A where P was either added to A or removed from A, in contradiction to the reference set. Most of these cases represented situations where it is unclear whether P really did belong in A or not, and different biologists often have different opinions. For example, the Torpedo complex, which is involved in transcriptional termination by RNA Polymerase II [63], was reported to be comprised of three subunits: the exonuclease Rat1, Rai1, and Rtt103. We predicted that Rtt103 was not a component of this complex, consistent with the weaker stoichiometric association of Rtt103 with the two other tightly-associated members of the complex [63]. In another example, we predicted that Csn12 was not a component of the COP9 signalosome, which is involved in deneddylation [27]. Consistent with this, Maytal-Kivity and colleagues demonstrated that Csn12 is the only component of this complex that is not required for the deneddylation activity [64]. Furthermore, we found that Csn12 but not other signalosome subunits is required for efficient mRNA splicing at a number of genes in budding yeast (Wilmes et al., submitted), suggesting that Csn12 plays multiple cellular roles and may not be an integral member of the complex. Finally, we predicted that Ski7 is part of the exosome complex, which harbors 3'-to-5' exonuclease activity, and acts on many different types of RNA. Evidence suggests that Ski7 acts as an adaptor to target the exosome to mRNAs lacking stop codons [65].

In other cases, however, the predictions made by our algorithm were interesting and worthy of further investigation. One such example is the eisosome, previously described to be primarily comprised of two subunits (Pil1 and Lsp1) [25]; however, we predicted that the complex contains

another, previously undescribed component, Mrp8. Consistent with this prediction, TAP purification of Mrp8 reveals that it is indeed a stoichiometrically stable member of the Eisosome complex (see Figure 4a). Further work will be required to determine the role that this novel subunit plays in eisosome function.

We also studied the novel complex predictions – those that did not match any of the reference complexes above our match threshold. A number of these turned out to be well-characterized complexes that, for some reason, had not (yet) been included into any of our three reference sets. For example, we identified: the Sit4/Sap185 heterodimer phosphatase complex [66]; a complex comprised of Yos9, Hrd3, Usa1, and Hrd1, which is involved in ER-associated degradation [67, 68]; and the U3-processome complex (complex 1129), involved in the generation and regulation of the small ribosomal particle [69]. Many others comprised plausible complexes that, to our knowledge, have not yet been characterized, and are worthy candidates for further investigation.

One such example is a complex (complex 1014) comprised of five components (YNL099C/OCA1, YNL056W/OCA2, YNL032W/SIW14/OCA3, YCR095C/OCA4, YHL029C/OCA5), four of which are putative phosphatases. One of the proteins (Oca1) has been previously shown to be required for cell cycle arrest in response to exposure to a lipid peroxide [70]. We note that the individual pairwise connections between these proteins were observed before, and that various forms of evidence support their shared function [14], including a shared phenotype of oxidant-induced cell-cycle arrest, which underlies the current name of many of these genes in SGD. However, this group was not previously identified as a complex, nor was its function characterized. Further supporting our prediction of this group as a complex is the fact that the chemical-genetic interaction profiles of the five genes were tightly clustered in a recent high-throughput study [45] (Figure 4b). Mutations in the components of the complex resulted in significant sensitivity to a number of conditions, including several that are related to nutrient starvation, including exposure to rapamycin, lysine dropout, and synthetic complete medium. To further characterize the functions of these factors, we subjected the mutants to quantitative genetic interaction profiling using an EMAP [60, 61, 71] focused on genes implicated in signaling, including protein and small-molecule kinases and phosphatases (supporting website [32]). Again, we found that the components of the complex had strong positive genetic interactions between them, and clustered tightly together within the set of ~500 genes included in the EMAP, both factors that indicate a strong functional connection [60]. Specifically we found that all components strong negative genetic interactions with Sch9, the yeast homolog of S6 kinase and a central node in nutrient signaling [72, 73]. Conversely, we found strong positive genetic interactions with GCN2, a protein kinase that phosphorylates the alpha subunit of translation initiation factor eIF2 (Sui2) in response to nutrient starvation [74]. Collectively, these data suggest an involvement of these proteins in response to nutrient starvation. Interestingly, both the chemogenomic profiling and the genetic interactions suggested a functional

connection to another uncharacterized gene, YDR067C, which may form a sixth member of the complex. Based on the coherence of this complex and its strong links to nutrient starvation, we propose to name this six-protein complex NSC (Nutrient Starvation Complex).

The predictions made by our algorithm also contained a number of mistakes, which fell into two main categories. The first comprised subsets of known complexes, such as subsets of the pre-60S ribosomal particle (complexes 1088 and 1106). These may represent functionally distinct submodules within larger complexes, and therefore may provide useful insight about complex structure. Consistent with this notion, we identified the de-ubiquitination unit of SAGA (Ubp8/Sgf11) [75-77]. The other category of error involved pairs of complexes that either interact or share subunits, and were merged by the HACO procedure into a single complex. For example, complex 1125 is comprised of two chromatin remodeling complexes, INO80-C and SWR-C, which have shared components, including Rvb1 and Rvb2, members of the RuvB family of helicases [78]. These two error modes illustrate the difficulty in selecting the appropriate granularity for making complex predictions, where some complexes occur fairly low in the clustergram, so that they have very high affinity with components outside the complex), whereas others occur very high in the clustergram, so that they contain components that have low affinity among themselves. This difficulty is perhaps one of the biggest challenges in accurately determining complexes. We note, however, that in some cases (such as the SAGA complex described above), the correct complexes themselves (or a slight variant) were sometimes also members in our set of predictions, a situation possible because of HACO's ability to make predictions at multiple levels of granularity. Thus, HACO is occasionally able to circumvent this challenge by trading off coverage for precision.

A Comprehensive Interaction Network

Complexes, together with individual proteins, comprise the basic units of the cell's interaction network. So far, most of the work [27-29] has focused on predictions of interactions between pairs of individual proteins. However, the view of the network in terms pairwise interactions loses much of its structure: Many interactions arise from co-complexness, so that a single large complex can give rise to a very dense (almost complete) subgraph in the network. Other pairwise interactions are representatives of interactions between larger complexes. We therefore set out to construct a comprehensive network of interactions between all basic units in the proteome, both complexes and proteins.

We compiled a reference set of complex-complex interactions (CCIs) and protein-complex interactions (PCIs) from reliable protein-protein interactions and hand-curation (see Methods). Importantly, to avoid circular reasoning, any interactions that we used in the construction of the gold-standard CCIs and PCIs were not given as features to the prediction algorithm. We used

ten-fold cross-validation to evaluate the ability of our model in accurately predicting CCIs. We randomly divide our reference interactions into ten sets. In each fold, we hide one set and train on the remaining nine sets. We then make predictions on the held-out set using the learned model. We compare three methods (see Methods): simple Naïve Bayes, a discriminative Boosting method, and Naïve Bayes with EM (NB+EM) that also makes use of the data for pairs that are not in our reference set. As we can see in Figure 5a, NB+EM performs better than both other methods, achieving very high performance: 44 of the top 50 predictions (88%) are in the positive reference set. We also compared these results to two state-of-the-art methods for predicting protein-protein interactions: the PE score and the InSite probabilities. As we can see, by integrating multiple sources of data, we are able to improve the accuracy to 0.88 (area under the ROC curve) from 0.85 and 0.79 for PE score and InSite probabilities respectively.

The PE score provides the strongest signal, and provides, by itself, accuracy on our reference set that is only somewhat lower than that of our integrated model. However, when evaluated on other metrics, our data integration provides more significant benefits. We expect interacting complexes to be involved in similar biological processes. Therefore, we measured the coherence of our CCI predictions relative to GO biological process annotations and MIPS functional categories (neither of which were used in training). Here, our predictions are considerably better than those obtained from the PE score alone (Figure 5b), suggesting that our set of reference complexes is perhaps somewhat biased towards areas that are well-covered by the TAP-MS assays.

We also apply our model to predict a unified network involving both proteins and complexes, a network that we call ComplexNet. In ComplexNet, we have both the interactions between two complexes and the interactions between a protein and a complex. As we can see from Figure 5c, by integrating multiple data sources, our Naïve Bayes model with EM is able to achieve higher accuracy than using either PE score or InSite probability alone. We generated predictions for all protein-complex pairs and complex-complex pairs by training on the entire reference set (see our supporting website [32] for the complete list of the predictions). Overall our predictions provide a comprehensive network of all of the interactions involving complexes. It can be combined with a set of high-quality protein-protein interactions (such as [49]), to provide a complete set of predictions for the *S. cerevisiae* protein interaction network. Figure 6 presents a fragment of the network.

We identified many CCIs that were expected and well-characterized, but not in the reference set, such as interactions between histones and several chromatin modifying complexes including the ISW1 complex, the HAT1 complex, and RSC. ComplexNet also suggests novel hypotheses, several of which have support in the literature. For example, we predict an interaction between complex 1035, which consists of the poorly characterized proteins Yer071c and Yir003w/Aim21,

with the yeast actin-capping protein (a Cap1-Cap2 heterodimer). Consistent with this prediction, high-throughput fluorescence microscopy found that Yir003w colocalizes with components of the actin cytoskeleton [37], and two-hybrid data has connected Yir003w to the actin-binding protein Abp1 [79]. Additionally, like deletion of CAP1 or CAP2, deletion of YER071C or YIR003W results in strong sensitivity to the actin-depolymerizing agent latrunculin [45]. Our observation suggests a more specific placement of this complex among the actin regulatory machinery. We also find several interesting interactions involving the centromere-localized kinetochore complex (Figure 6), some of which have independent support. Our prediction of an interaction between the kinetochore and the proteasome are supported by a recent report that levels of Cse4, a centromere-localized histone, are regulated by ubiquitin-proteasome-mediated proteolysis [80]. Our predicted link between the kinetochore and the spliceosome is consistent with evidence of a functional connection between these two factors [81]. The remaining connections we observe with the kinetochore (pre-60S ribosomal particle and RNA polymerase III) are intriguing but more work will be required to determine the validity and functional significance of these predicted relationships. We can also learn from the false positive predictions of CCIs. Our algorithm does make some apparently false positive predictions, and many of them fall into two main categories. Pairs of complexes that share a substantial number of common components, such as the SWR Complex and NuA4, are sometimes identified as interacting. Additionally, pairs of complexes which do not interact directly, but are one link away in the interaction network are sometimes identified. Along these lines, we identify an interaction between the NuA4 histone acetylase complex and the opposing RPD3(L) deacetylase complex. Both complexes have subunits with specificity for binding K4-trimethylated histone H3 [82] and have been found to be regulated by binding to 14-3-3 proteins [83]. Thus, even such a false positive may still provide interesting biological insights.

Essentiality and complex size

Much discussion has occurred regarding the relationship between essentiality and the structure of the protein-protein interaction network. Early work of Jeong et al. [26, 84] found that hub proteins in a protein-protein interaction network are more likely to be encoded by essential genes. More recent work [85] suggests that highly-connected proteins are simply more likely to participate in essential protein-protein interactions, and are therefore more likely to be essential. However, a deeper insight on the relationship between the protein network and essentiality can be obtained by considering the network at the level of complexes rather than pairwise interactions. Such an analysis was recently performed by Hart et al. [5], who showed that essential proteins are concentrated in certain complexes, resulting in a dichotomy of essential and non-essential complexes. This phenomenon was also found in our predicted complexes (Figure 7a). However, that finding does not explain why “hubs” in the network are more likely to be essential. We therefore looked into the distribution of essential proteins in complexes of different sizes, and found that the fraction of essential components in a complex tends to increase with complex size (Figure

7b). Moreover, when we aggregate over all complexes of a given size, larger complexes tend to have a far greater *proportion* of essential proteins among their components (Figure 7b). Components in a large complex are naturally highly connected in the protein interaction network, and therefore often form “hubs”. Thus, the finding regarding the essentiality of hubs very likely arises from the fact that large complexes are more likely to have a much higher ratio of essential genes. Our finding is consistent with the recent work of Zotenko et al. [86], who argue that essential hubs are often members of a densely connected set of proteins performing an essential cellular function. However, this analysis is still performed on the pairwise protein network, and hence is unable to identify the strong dependency between the size of a complex and its essentiality.

To test whether our finding truly explains the phenomenon of essential hubs, we tested whether essentiality is better explained by complex size or by hubness. We rank every protein based on the size of the largest complex to which it belongs, and for the K top-ranked proteins (for different values of K), plot the number of essential vs. non-essential proteins (Figure 7c). We plotted a similar curve by using the hubness of the protein – the degree in the yeast two-hybrid protein-protein interaction network [35, 36]. As we can see, complex size is a much better predictor for essentiality than hubness. We note that if we use the scaled PE score (at threshold > 0.5) to define a protein-interaction network, the hubness becomes a strong predictor of protein essentiality. However, PE score is more related to co-complexness than interaction, and thus this metric of hubness is directly related to complex size. Nevertheless, using complex size directly is still better than using scaled PE score. Interestingly, if we use the size of the largest enclosing reference complex to rank each protein, the result is slightly less predictive than using our predicted complexes, or even the scaled PE score directly.

Discussion

Identifying a comprehensive set of protein complexes in yeast is an important but challenging task. The high-quality and high-throughput TAP-MS data, which directly measures co-complexness, provides a starting point for accurately reconstructing these complexes. Indeed, two recent studies [5, 6] used the TAP-MS data to produce a set of complexes with the state-of-the-art performances. Both methods applied a simple clustering algorithm to a score derived directly from the TAP-MS data. In this paper, we are able to significantly improve the accuracy of the complex reconstruction in three ways. First, we carefully constructed a large set of reference complexes and trained our model so it specifically predicts co-membership in stoichiometrically stable complexes. Second, we integrated multiple sources of heterogeneous data so our predictions are more robust to noise and incomplete coverage in the TAP-MS data. Finally, we extended the highly effective hierarchical agglomerative clustering (HAC) algorithm to allow reconstruction of

clusters with overlap, a flexibility that allows it to circumvent many of the limitations of the standard HAC algorithm. We show that the resulting set of predicted complexes (available from our website [32]) has significantly higher accuracy and is more biologically coherent than that of other recent methods. In many cases, it is even more coherent than the reference set, indicating it is of high quality and can be used as a new reference set. When combined with our comprehensive, hand-curated reference set (also available from our website [32]), our work provides a significant new resource to the research community.

With our high-quality set of complexes, we are able to take a higher-level perspective on the protein-protein interaction network, viewing it in terms of interactions between atomic units – whether individual proteins or stable complexes. There has been much work on predicting protein-protein interactions. However, these pairwise interactions are often induced by higher-level relationships: those within a complex and those between complexes. Interactions within a complex give rise to densely connected subgraphs in the interaction network; interactions between complexes can give rise to a network of interconnections involving different members of the two complexes. Viewing the network in terms of its atomic units can help clarify its structure and its basic properties. We therefore defined the novel problem of predicting interactions between complexes and other complexes or proteins, and constructed a new, high-accuracy method for making such predictions. The result of our analysis is ComplexNet – a unified interaction network involving both proteins and complexes. We can now analyze the properties of this network, which better captures the true interactions underlying cellular processes. In particular, this network provides a new perspective on the previously-observed relationship between the “hubness” of a protein in the network and its essentiality, demonstrating that larger complexes are more likely to be essential, and comprise a large fraction of essential proteins. It would also be of interest to study other properties of this network, such as its connectivity or hierarchical structure.

To find a coherent set of proteins that form a complex, we have the choices of many different clustering algorithms. Brohee et al. [87] showed that MCL works well on protein-protein interaction network by comparing it with three other clustering algorithms in literature. So not surprisingly, Pu et al. [6] and Hart et al. [5] applied MCL to the TAP-MS network, which is confirmed by our results to be better than other existing methods in terms of reconstructing reference complexes and biological coherence. On the other hand, we found that HAC achieves about the same accuracy as MCL. Therefore, we focus on the best-proven method and try to further improve it by addressing some of its limitations. One of the significant advantages of our HACO algorithm, which extends the HAC, is its ability to create overlapping complexes. Indeed, the inability of traditional HAC to generate overlapping clusters is one of its major deficiencies in other types of data as well. Interestingly, in our results, there were relatively few cases where two

“correct” complexes shared subcomponents. Most of the benefit of HACO arose from avoiding mistakes arising from HAC’s greedy decisions, and from allowing predictions at different levels of granularity (e.g., a complex and one of its subunits). Nevertheless, the lack of extensive sharing of components between complexes was surprising, given that such sharing is present in the reference set. To some extent, this phenomenon is due to the tradeoff in HACO parameters between increasing the amount of component sharing and errors arising from merging of distinct complexes. However, HACO applied to other data sets (data not shown) did give rise to much more extensive sharing among different clusters. Thus, a complementary hypothesis is that some of the sharing of components between complexes arises when a protein plays roles in different complexes in different conditions. Our data, having been acquired almost entirely in YPD, would not reveal this condition-specific pleiotropy. It would be of great interest to acquire TAP-MS data in different conditions, and study the extent to which complex structure is condition specific.

We note that there are other clustering algorithms [88, 89] that also generate overlapping complexes. However, both of them are applicable only to a binary interaction network, so an application to our task would require that we discretize the continuous affinities between protein pairs into two values (interacting and non-interacting), using some fixed threshold. Our analysis of the affinities for reference complexes suggested strongly that proteins that are co-complexed often exhibit affinities over a very broad range, so that such a discretization would result in an unacceptable loss of useful information. On the other hand, HACO uses the continuous-valued affinities directly, which allows the finer resolution of the computed affinities to be used by the algorithm. We also note that we devised several other novel methods that attempt to construct overlapping clusters. For example, one method directly learns an affinity function to predict the likelihood that a set of proteins forms a complex, aiming to take advantage of features involving more than two proteins. HACO significantly outperformed all of our other proposed methods, and so we omit details.

There are still many reference complexes that are not matched by our predicted complexes. Many of them fall into roughly two categories. In the first category, proteins in the reference complex have high affinities with each other, and are grouped as a set during the HACO procedure. However, they are not selected in our predictions because they are not at the granularity where we cut our HACO cluster-lattice. They then become subsets or supersets of some predicted complexes. In fact, if we use all the sets generated during our HACO procedure as predicted complexes, 136 reference complexes would be perfectly predicted and 243 would be well matched by some predicted complexes, in comparison to 95 perfect matches and 189 good matches in our current predictions. However, this approach would result in far too many predictions (3478), greatly reducing sensitivity. This fact highlights the limitations in defining a universal level of affinity at which one determines that a group of proteins form a stable complex, and suggests that a more flexible technique may be a useful direction for future work. In the second category, the

proteins in the reference complex do not have high affinities with each other. This situation arises when the signal in the data is not sufficiently strong to indicate that two proteins are likely to interact. As most of our signal comes from the TAP-MS data, such “blind spots” can arise from limitations of this assay, such as complexes of low abundance or that are membrane-bound. In particular, we note that the TAP-MS data was all acquired in a single condition (rich media), and some complexes may simply not be present in the cell in that condition. Our inability to recover such complexes arises not from computational limitations, but from limitations in the data. New experimental assays are needed before these complexes can be reconstructed.

Like other previous approaches, our method was developed in the context of *S. cerevisiae*, where we have the most data relevant to protein-protein interactions. Having a high-quality set of predicted complexes is of significant value even in yeast, as yeast provides an excellent model for many core biological processes. Moreover, many key complexes are conserved from yeast to human, making our complex predictions valuable also to analysis of higher-level organisms. Finally, our method is general-purpose, and can easily be applied more broadly. Its ability to integrate multiple sets of diverse data makes it suitable for other organisms where we may not have the same type of data available as in yeast. With the increasing amount of high-throughput protein-protein interaction data, both TAP-MS [90] and other assays [91, 92], we should soon be able to provide a high-quality reconstruction of protein complexes in other organisms, including human.

Our work takes a step towards a more hierarchical view of the protein-protein interaction network, moving up from individual proteins to complexes as the basic interacting units. The next level of the hierarchy is the pathways that comprise cellular pathways. Although the notion of a “pathway” is not as well-defined, it would nevertheless be very useful to reconstruct pathways that are comprised of interacting complexes and proteins. This type of analysis will give us a unified perspective on the underlying hierarchical organization of the cell, and provide significant insight.

Acknowledgments: The work was supported by the National Science Foundation and by DARPA under the CALO program (H.W., B.K., and D.K.), the Howard Hughes Medical Institute (S.R.C. and K.M.S.), an Ernst Schering Postdoctoral Fellowship (D.F.), and from Sandler Family Funding (N.J.K.). We thank Maureen Hillenmeyer for useful discussions, and for early access to the chemical genomic data.

Author information: Correspondence and requests for materials should be addressed to Daphne Koller (koller@cs.stanford.edu) or Nevan J. Krogan (krogan@cmp.ucsf.edu).

References

1. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al.* **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415**(6868):141-147.
2. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al.* **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415**(6868):180-183.
3. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al.* **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440**(7084):631-636.
4. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP *et al.* **Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*.** *Nature* 2006, **440**(7084):637-643.
5. Hart GT, Lee I, Marcotte ER: **A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality.** *BMC Bioinformatics* 2007, **8**:236.
6. Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ: **Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*.** *Proteomics* 2007, **7**(6):944-960.

7. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: **Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae***. *Mol Cell Proteomics* 2007, **6**(3):439-450.
8. Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V: **MIPS: analysis and annotation of proteins from whole genomes in 2005**. *Nucleic Acids Res* 2006, **34**(Database issue):D169-172.
9. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res* 2002, **30**(7):1575-1584.
10. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**(5644):449-453.
11. Zhang LV, Wong SL, King OD, Roth FP: **Predicting co-complexed protein pairs using genomic and proteomic data integration**. *BMC Bioinformatics* 2004, **5**:38.
12. Qiu J, Noble WS: **Predicting co-complexed protein pairs from heterogeneous data**. *PLoS Comput Biol* 2008, **4**(4):e1000054.
13. Chen J, Yuan B: **Detecting functional modules in the yeast protein-protein interaction network**. *Bioinformatics* 2006, **22**(18):2283-2290.
14. Lee I, Li Z, Marcotte EM: **An Improved, Bias-Reduced Probabilistic Functional Gene**

Network of Baker's Yeast, *Saccharomyces cerevisiae*. *PLoS ONE* 2007, 2(10):e988.

15. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, 402(6757):83-86.
16. Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A: **From gene networks to gene function.** *Genome Res* 2003, 13(12):2568-2576.
17. Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D: **Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach.** *Genome Biol* 2003, 4(9):R59.
18. von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P: **Genome evolution reveals biochemical networks and functional modules.** *Proc Natl Acad Sci U S A* 2003, 100(26):15428-15433.
19. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins.** *Nucleic Acids Res* 2003, 31(1):258-261.
20. Yanai I, DeLisi C: **The society of genes: networks of functional links between genes from comparative genomics.** *Genome Biol* 2002, 3(11):research0064.

21. Yellaboina S, Goyal K, Mande SC: **Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data.** *Genome Res* 2007, **17**(4):527-535.
22. Collins M, Schapire R, Singer Y: **Logistic Regression, AdaBoost and Bregman Distances.** *Machine Learning* 2002, **48**:253-285.
23. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.
24. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M *et al*: **Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map.** *Nature* 2007, **446**(7137):806-810.
25. Walther TC, Brickner JH, Aguilar PS, Bernales S, Pantoja C, Walter P: **Eisosomes mark static sites of endocytosis.** *Nature* 2006, **439**(7079):998-1003.
26. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**(6833):41-42.
27. Wei N, Deng XW: **The COP9 signalosome.** *Annu Rev Cell Dev Biol* 2003, **19**:261-286.
28. Liu Y, Liu N, Zhao H: **Inferring protein-protein interactions through high-throughput**

- interaction data from diverse organisms. *Bioinformatics* 2005, **21**(15):3279-3285.
29. Bock JR, Gough DA: **Predicting protein--protein interactions from primary structure.** *Bioinformatics* 2001, **17**(5):455-460.
30. **Supporting website, CCI** [<http://dags.stanford.edu/CCI/>]
31. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M *et al.* **SGD: Saccharomyces Genome Database.** *Nucleic Acids Res* 1998, **26**(1):73-79.
32. **Supporting website, Complex** [<http://dags.stanford.edu/Complex/>]
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.* **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25-29.
34. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC *et al.* **The Stanford Microarray Database: data access and quality assessment tools.** *Nucleic Acids Res* 2003, **31**(1):94-96.
35. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci U S A* 2001, **98**(8):4569-4574.
36. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D,

- Narayan V, Srinivasan M, Pochart P *et al*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae***. *Nature* 2000, **403**(6770):623-627.
37. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast**. *Nature* 2003, **425**(6959):686-691.
38. Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation**. *Bioinformatics* 2003, **19**(10):1275-1283.
39. Xenarios I, Salwinski L, Duan XQJ, Higney P, Kim SM, Eisenberg D: **DIP ; the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Research* 2002, **30**(1):303-305.
40. Vapnik V: **The Nature of Statistical Learning Theory**, 2nd edn: Springer Verlag; 1999.
41. Freund Y, Schapire R: **A decision-theoretic generalization of on-line learning and an application to boosting**. *Journal of Computer and System Sciences* 1997, **55**(1):119-139.
42. Sokal R, Michener C: **A statistical method for evaluating systematic relationships**. *Univ Kans Sci Bull* 1958, **38**:1409-1438.

43. Krause R, von Mering C, Bork P: **A comprehensive set of protein complexes in yeast: mining large scale protein-protein interaction screens.** *Bioinformatics* 2003, **19**(15):1901-1908.
44. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast.** *Nature* 2003, **425**(6959):737-741.
45. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D *et al.* **The chemical genomic portrait of yeast: uncovering a phenotype for all genes.** *Science* 2008, **320**(5874):362-365.
46. Maclsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E: **An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*.** *BMC Bioinformatics* 2006, **7**:113.
47. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J *et al.* **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**(7004):99-104.
48. Date SV, Marcotte EM: **Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages.** *Nat Biotechnol* 2003, **21**(9):1055-1062.
49. Wang H, Segal E, Ben-Hur A, Li QR, Vidal M, Koller D: **InSite: a computational**

method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol* 2007, **8**(9):R192.

50. Zakrzewska A, Boorsma A, Brul S, Hellingwerf KJ, Klis FM: **Transcriptional response of *Saccharomyces cerevisiae* to the plasma membrane-perturbing compound chitosan.** *Eukaryot Cell* 2005, **4**(4):703-715.
51. Mercier G, Berthault N, Touleimat N, Kepes F, Fourel G, Gilson E, Dutreix M: **A haploid-specific transcriptional response to irradiation in *Saccharomyces cerevisiae*.** *Nucleic Acids Res* 2005, **33**(20):6635-6643.
52. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell* 2001, **12**(2):323-337.
53. Lai LC, Kosorukoff AL, Burke PV, Kwast KE: **Dynamical remodeling of the transcriptome during short-term anaerobiosis in *Saccharomyces cerevisiae*: differential response and role of Msn2 and/or Msn4 and other factors in galactose and glucose media.** *Mol Cell Biol* 2005, **25**(10):4075-4091.
54. O'Rourke SM, Herskowitz I: **A third osmosensing branch in *Saccharomyces cerevisiae* requires the Msb2 protein and functions in parallel with the Sho1 branch.** *Mol Cell Biol* 2002, **22**(13):4739-4749.
55. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D,

- Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11**(12):4241-4257.
56. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p.** *Mol Biol Cell* 2001, **12**(10):2987-3003.
57. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**(5338):680-686.
58. Kitagawa E, Akama K, Iwahashi H: **Effects of iodine on global gene expression in *Saccharomyces cerevisiae*.** *Biosci Biotechnol Biochem* 2005, **69**(12):2285-2293.
59. Shevchenko A, Wilm M, Vorm O, Mann M: **Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels.** *Anal Chem* 1996, **68**(5):850-858.
60. Collins SR, Schuldiner M, Krogan NJ, Weissman JS: **A strategy for extracting and analyzing large-scale quantitative epistatic interaction data.** *Genome Biol* 2006, **7**(7):R63.
61. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF *et al.* **Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile.** *Cell* 2005, **123**(3):507-519.
62. Bader GD, Hogue CW: **An automated method for finding molecular complexes in**

- large protein interaction networks. *BMC Bioinformatics* 2003, 4:2.
63. Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedeja E, Greenblatt JF, Buratowski S: **The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II.** *Nature* 2004, 432(7016):517-522.
64. Maytal-Kivity V, Piran R, Pick E, Hofmann K, Glickman MH: **COP9 signalosome components play a role in the mating pheromone response of *S. cerevisiae*.** *EMBO Rep* 2002, 3(12):1215-1221.
65. van Hoof A, Staples RR, Baker RE, Parker R: **Function of the ski4p (Csl4p) and Ski7p proteins in 3'-to-5' degradation of mRNA.** *Mol Cell Biol* 2000, 20(21):8230-8243.
66. Luke MM, Della Seta F, Di Como CJ, Sugimoto H, Kobayashi R, Arndt KT: **The SAP, a new family of proteins, associate and function positively with the SIT4 phosphatase.** *Mol Cell Biol* 1996, 16(6):2744-2755.
67. Denic V, Quan EM, Weissman JS: **A luminal surveillance complex that selects misfolded glycoproteins for ER-associated degradation.** *Cell* 2006, 126(2):349-359.
68. Carvalho P, Goder V, Rapoport TA: **Distinct ubiquitin-ligase complexes define convergent pathways for the degradation of ER proteins.** *Cell* 2006, 126(2):361-373.
69. Dragon F, Gallagher JE, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y *et al*: **A large**

nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* 2002, **417**(6892):967-970.

70. Alic N, Higgins VJ, Dawes IW: **Identification of a *Saccharomyces cerevisiae* gene that is required for G1 arrest in response to the lipid oxidation product linoleic acid hydroperoxide.** *Mol Biol Cell* 2001, **12**(6):1801-1810.
71. Fiedler D, Braberg H, Mehta M, Chechik G, Cagney G, Mukherjee P, Silva AC, Shales M, Collins SR, van Wageningen S *et al.* **Functional Organization of the *S. cerevisiae* Phosphorylation Network.** *Cell* 2009.
72. Jorgensen P, Rupes I, Sharom JR, Schneper L, Broach JR, Tyers M: **A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size.** *Genes Dev* 2004, **18**(20):2491-2505.
73. Urban J, Soulard A, Huber A, Lippman S, Mukhopadhyay D, Deloche O, Wanke V, Anrather D, Ammerer G, Riezman H *et al.* **Sch9 is a major target of TORC1 in *Saccharomyces cerevisiae*.** *Mol Cell* 2007, **26**(5):663-674.
74. Garcia-Barrio M, Dong J, Ufano S, Hinnebusch AG: **Association of GCN1-GCN20 regulatory complex with the N-terminus of eIF2alpha kinase GCN2 is required for GCN2 activation.** *Embo J* 2000, **19**(8):1887-1899.
75. Powell DW, Weaver CM, Jennings JL, McAfee KJ, He Y, Weil PA, Link AJ: **Cluster analysis of mass spectrometry data reveals a novel component of SAGA.** *Mol Cell*

Biol 2004, **24**(16):7249-7259.

76. Ingvarsdottir K, Krogan NJ, Emre NC, Wyce A, Thompson NJ, Emili A, Hughes TR, Greenblatt JF, Berger SL: **H2B ubiquitin protease Ubp8 and Sgf11 constitute a discrete functional module within the *Saccharomyces cerevisiae* SAGA complex.**

Mol Cell Biol 2005, **25**(3):1162-1172.

77. Lee KK, Florens L, Swanson SK, Washburn MP, Workman JL: **The deubiquitylation activity of Ubp8 is dependent upon Sgf11 and its association with the SAGA complex.** *Mol Cell Biol* 2005, **25**(3):1173-1182.

78. Caruthers JM, McKay DB: **Helicase structure and mechanism.** *Curr Opin Struct Biol* 2002, **12**(1):123-133.

79. Fazi B, Cope M, Douangamath A, Ferracuti S, Schirwitz K, Zucconi A, Drubin D, Wilmanns M, Cesareni G, Castagnoli L: **Unusual binding properties of the SH3 domain of the yeast actin-binding protein Abp1: structural and functional analysis.** *J Biol Chem* 2002, **277**(7):5290-5298.

80. Collins KA, Furuyama S, Biggins S: **Proteolysis contributes to the exclusive centromere localization of the yeast Cse4/CENP-A histone H3 variant.** *Curr Biol* 2004, **14**(21):1968-1972.

81. Bialkowska A, Kurlandzka A: **Proteins interacting with Lin 1p, a putative link between chromosome segregation, mRNA splicing and DNA replication in**

- Saccharomyces cerevisiae.** *Yeast* 2002, **19**(15):1323-1333.
82. Shi X, Kachirskaia I, Walter KL, Kuo JH, Lake A, Davrazou F, Chan SM, Martin DG, Fingerman IM, Briggs SD *et al.* **Proteome-wide analysis in Saccharomyces cerevisiae identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36.** *J Biol Chem* 2007, **282**(4):2450-2455.
83. Lottersberger F, Panza A, Lucchini G, Longhese MP: **Functional and physical interactions between yeast 14-3-3 proteins, acetyltransferases, and deacetylases in response to DNA replication perturbations.** *Mol Cell Biol* 2007, **27**(9):3266-3281.
84. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP *et al.* **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430**(6995):88-93.
85. He X, Zhang J: **Why do hubs tend to be essential in protein networks?** *PLoS Genet* 2006, **2**(6):e88.
86. Zotenko E, Mestre J, O'Leary DP, Przytycka TM: **Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality.** *PLoS Comput Biol* 2008, **4**(8):e1000140.
87. Brohee S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks.** *BMC Bioinformatics* 2006, **7**:488.

88. Palla G, Derenyi I, Farkas I, Vicsek T: **Uncovering the overlapping community structure of complex networks in nature and society.** *Nature* 2005, **435**(7043):814-818.
89. Yu H, Paccanaro A, Trifonov V, Gerstein M: **Predicting interactions in protein networks by completing defective cliques.** *Bioinformatics* 2006, **22**(7):823-829.
90. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M *et al*: **Large-scale mapping of human protein-protein interactions by mass spectrometry.** *Mol Syst Biol* 2007, **3**:89.
91. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S *et al*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122**(6):957-968.
92. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M *et al*: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13**(10):2363-2371.

Figure Legends

Figure 1: Accuracy in reconstructing reference complexes

Comparison of predicted complexes to other state-of-the-art methods in the ability to accurately reconstruct reference complexes.

(a) The number of reference complexes well-matched by our predictions (y-axis) for the different methods we compared (x-axis). The prediction quality is shown as bars: blue bars – perfect prediction; green bars – predictions that differ by a single protein (one extra or one fewer); red bars – predictions that differ by two proteins. Hart and Pu are state-of-the-art methods that outperform Gavin and Krogan. The method of Bader et al. has even lower accuracy (data not shown). Applying HAC to PE score (HAC PE) performed slightly better than Hart and Pu, which use MCL. Our model, which uses LogitBoost and clustering, is able to achieve significantly better results than any other method by integrating multiple sources of data. The results are better even when we use simple HAC (HAC all, 88 perfect matches) for the clustering, and improve further when we use HACO (HACO, 95 perfect matches). This improvement is consistent over all five folds in our cross-validation process: over the five folds, HAC PE recovers 15, 11, 16, 22, and 9 of the complexes; HAC all recovers 21, 13, 21, 23, 10; and HACO recovers 24, 13, 23, 23, 12. This consistency over folds demonstrates the robustness in the improvement we obtain using our method. In “Train all”, we trained on all data and tested on the same data; this method achieves only slightly higher accuracy, which indicates little overfitting to the training data, and supports evaluating biological coherence of our predictions on this set.

(b) The x-axis is the sensitivity of our predictions, which quantifies how likely a prediction is to match some reference complexes; the y-axis is the coverage of our predictions, which quantifies how many reference complexes are matched by our predictions (see Methods). Our approach has higher sensitivity and coverage than other methods. HACO has the highest product of sensitivity and coverage, except for “Train all”.

Figure 2: Contribution of each data source

To assess the contribution of each data source, we successively applied our pipeline with HAC to each data source alone and to all data sources except each data source; shown are the interesting cases (see also Figure S2b), using the same format as in Figure 1. The PE score by itself predicts most of the complexes, but we still get a significant improvement by integrating other data sources. Localization or expression are non-specific, and by themselves do not predict any complexes at all, but removing them decreases the accuracy, suggesting that they help clarify ambiguities in the TAP-MS data. Conversely, the yeast two-hybrid feature by itself predicts a reasonable number of complexes, but removing it does not decrease accuracy at all, suggesting that it is redundant with the TAP-MS data.

Figure 3: Coherence of our predicted complexes

We computed the functional coherence between proteins in the same complex against external data sources that are not used in training. More coherent complexes have a smaller difference in protein abundance, a smaller semantic distance in GO biological process, share more transcriptional regulators, and have a higher growth fitness correlation. The y-axis shows the values for these metrics of functional coherence; also shown is the performance of random pairs (green line). Our predicted set of complexes significantly outperforms other state-of-the-art methods. For GO biological processes, our complexes have a semantic distance 8% and 17% lower than the methods of Hart and Pu respectively. For protein abundance, the improvement over Hart and Pu is 5%, 10%, respectively; conversely, our complexes are 12% less coherent than the top affinity pairs, suggesting that proteins with lower affinity scores can be members of the complex, but also play other roles in the cell, reducing their correlation with other proteins in the same complex. For the correlation of growth phenotypes across different conditions, our predicted complexes are 19%, 31% more coherent, a very significant improvement. Finally, protein pairs within our complexes on average share 30%, 59% more transcription factors than those of Hart and Pu. The comparison to the reference complexes shows that our complexes are considerably more coherent on regulator overlap, and perform similarly on correlation of abundance and growth phenotype. Conversely, our complexes are 21% less coherent than the reference complexes on GO biological process annotations; this is not surprising, as the reference complexes and GO annotations are derived (at least partly) from similar data sources, such as literature and small scale experiments.

Figure 4: Validation of novel biological findings

(a) Pil1, Lsp1 and Mrp8 form a stable complex. TAP-tagged Pil1 was affinity purified from yeast. Highly enriched fractions were run on SDS PAGE and co-purifying proteins were indentified by LC-MS/MS as Lsp1 and Mrp8, indicated on the left; protein sizes are shown in kDa on the right. The result supports our prediction that Mrp8 is a component of the Eisosome complex.

(b-c) Support for newly-uncovered NSC complex, comprised of six genes (YCR095C, YHL029C, YNL032W, YNL056W, YNL099C/OCA1, YDR067C), four of which are phosphatases. Five of these components were predicted by HACO to be a stoichiometrically stable complex; based on other data (shown in this figure) we conjecture that the sixth (YDR067C) may also be a member of this complex .

(b) Support in chemical genomics data of Hillenmeyer et al. [45], which measured the fitness profiles of all non-essential homozygous yeast mutants under 418 conditions. (left) The fitness profiles of the six predicted NSC members cluster tightly together. (right) Shown are the conditions in which at least one of the six components had a fitness defect with $p < 1e-10$; the

consistently strong sensitivity to rapamycin, lysine dropout, and synthetic complete medium suggests the involvement of these proteins in response to nutrient starvation.

(c) Support in new EMAP data, which measured quantitative genetic interaction profiles with ~500 genes involved in signaling. (left) The genetic interaction profiles of the six components cluster tightly together. (right) The components have positive genetic interactions among them, and exhibit significant interactions with genes involved in nutrient starvation response, including Sch9 and Gcn2.

Figure 5: Verification of complex-based interaction network

(a) Verification of our complex-complex interaction predictions relative to our reference set. Complex pairs in the hidden set of a ten-fold cross validation are ranked based on their predicted interaction probabilities. Blue, green, and red curves are for the three models we tried. Light blue and purple curves are for the predictions using only PE score or InSite probabilities respectively. Each point on the curve corresponds to a different threshold, giving rise to a different number of predicted interactions. The value on the x-axis is the number of pairs not in the reference set, but predicted to interact. The value on the y-axis is the number of reference interactions that are predicted to interact. The bars on the right bottom corner are the areas under the ROC curves. Our Naïve Bayes model with EM achieves the highest accuracy. The prediction made by PE score alone is slightly worse than our integrated models.

(b) Functional coherence of interacting complexes, measured by joint membership in the same MIPS functional category, a feature not used in training. We only consider those interacting complexes if both of them are assigned to some MIPS category. We picked the top 500 predictions from our NB+EM model, and the top 500 obtained from the PE score alone. We compared them to complex pairs in our reference set and randomly selected pairs. The y-axis shows the proportion of interacting complexes that are assigned to the same MIPS category. As we can see, 59.2% of our predicted interacting complexes share the same MIPS category, while only 35.2% and 45.5% share the same category for random complex pairs and for those predicted by PE score alone, respectively. The reference complexes are the most coherent, a fact which is not surprising as the functional classification of reference complexes is sometimes derived from the same literature sources as the interactions between those complexes.

(c) Verification against a reference set of our unified predictions of protein-complex and complex-complex interactions set. Complex pairs in the hidden set of a ten-fold cross validation are ranked based on their predicted interaction probabilities. Blue curve is for our Naïve Bayes model with EM. Green curve is for the predictions using only PE score and red curve is for the prediction using only InSite probability. Each point on the curve corresponds to a different threshold, giving rise to a different number of predicted interactions. The value on the x-axis is the number of pairs not in the reference set, but predicted to interact. The value on the y-axis is the number of reference interactions that are predicted to interact. The areas under the curves are

shown by the bars on the bottom-right corner.

Figure 6: A complex-level interaction network

(a) A fragment of our ComplexNet, comprising a subset of the interactions between the largest complexes. Shown are the 500 highest-confidence predictions plus the reference interactions, restricted to interactions between complexes of size ≥ 3 . The color of each complex indicates the fraction of essential components, demonstrating the enrichment of essential proteins in larger complexes. The complexes are placed in regions based on their cellular localization, determined by majority vote based on the data of Huh et al. [37]. The inset shows all interactions that involve the kinetochore complex.

(b) A list of the complexes associated with the numbers in the figure. Complexes are associated with a name of a known complex when they overlap with that complex with F-score > 0.5 . Otherwise, the number associated with a complex is a unique identifier used in our supporting website [32].

Figure 7: Relationship between complex size and essentiality

(a) Fraction of complexes with different essentiality fractions. Each complex is represented by its size and the fraction of essential components. The different colors represent different ratio of essentiality in a complex, discretized into five bins. The x axis represents the complex size, and the y axis the fraction of complexes of that size that have this particular essentiality ratio. We can see that the large majority of complexes of size 2 have essentiality ratio in the range 0-0.2, whereas larger complexes tend to have a larger essentiality ratio. Also shown on the x axis, in parentheses, is the number of complexes in each category (e.g., there are 54 complexes of size 3).

(b) The relationship between complex size and the *proportion* of essential proteins in complexes of that size. The x-axis is the size bin of the complexes. The y-axis is the proportion of essential proteins in all complexes within the size bin. As we can see, larger complexes tend to have a higher proportion of essential proteins.

(c) Evaluation of different metrics as predictive of essentiality: size of the largest enclosing complex vs. degree in the protein-protein interaction network (hubness). For the red and light blue curves, we rank each protein based on the size of the largest complex to which it belongs; the red curve uses predicted complexes and the light blue curve uses the reference complexes. For the blue curve and green curve, we use the hubness – degree of protein in a protein-protein interaction network; the blue curve uses the yeast two-hybrid protein-protein interaction network, and the green curve uses a network where pairs are connected if they have scaled PE score > 0.5 . The x-axis is the number of essential proteins in the K top-ranked proteins (for different values of K) and the y-axis is the number of non-essential proteins. Complex size in our predicted complexes (red) is the best predictor for essentiality. The hubness based on PE score (green) performs better than the other metrics, presumably because it also correlates directly with co-membership in a

complex. The reference complexes (light blue) perform slightly worse, but considerably better than interactions in the Y2H data.

Supplementary Figures

Figure S1: Set of reference complexes and predicted complexes

(a) We compile a comprehensive set of reference complexes by combining literature-derived results from small-scale experiments in MIPS and SGD with a hand-curated list that we generated. The MIPS, SGD, and hand-curated set contain 225, 195, and 164 complexes respectively. We establish correspondence between the three lists and combining them into a high-confidence reference set of 340 complexes. The reference set is suitable for training our method and for evaluating the accuracy of its predictions. We predicted a set of 418 complexes on average in the cross validation runs and a set of 383 complexes in the run when we train on all data.

(b) Size distribution of complexes in reference set and predicted set.

Figure S2: Contribution of each data source

(a) Same as Figure 2, but showing results when we attempt to reconstruct only SGD complexes. The results are qualitatively the same as Figure 2, showing that, although MIPS complexes were used (in a minor way) to tune the parameters of the PE scores, there is no overfitting to those data.

(b) Same as Figure 2, but showing complete analysis. To assess the contribution of each data source, we successively applied our pipeline with HAC to the data source alone and to all data sources except that data source. The x-axis shows the different runs. The y-axis is the number of reference complexes that are well matched by our predictions. Blue bars are for reference complexes that are perfectly matched by our predictions. Green bars are for reference complexes that differ with some of our predicted complexes by one protein, either one extra or one fewer. Red bars are for reference complexes that differ with some of our predicted complexes by two proteins.

Figure S3: Illustration of the HACO intuition

Set A is merged to form set M in an earlier step of HACO. Instead of removing A from the candidate pool, as in standard HAC, we keep it and consider its proposed merger with another set C . If $N = A \cup C$ is almost as coherent as M , as shown on the left panel, we merge A and C to create N so we have overlapping sets of M and N . On the other hand, if N is much less coherent than M , as shown on the right panel, we do not merge A and C . Instead, we consider the potential merge between M and C .

Figure S4: Metrics for overlap between two complexes

R_j is a reference complex and C_i is our predicted complex. The overlap between the two complexes can be quantified by the Jaccard coefficient and the Hamming distance as illustrated in the figure. The better the overlap, the greater the Jaccard coefficient and the smaller the Hamming distance, with perfect match has a Jaccard coefficient 1 and Hamming distance 0.

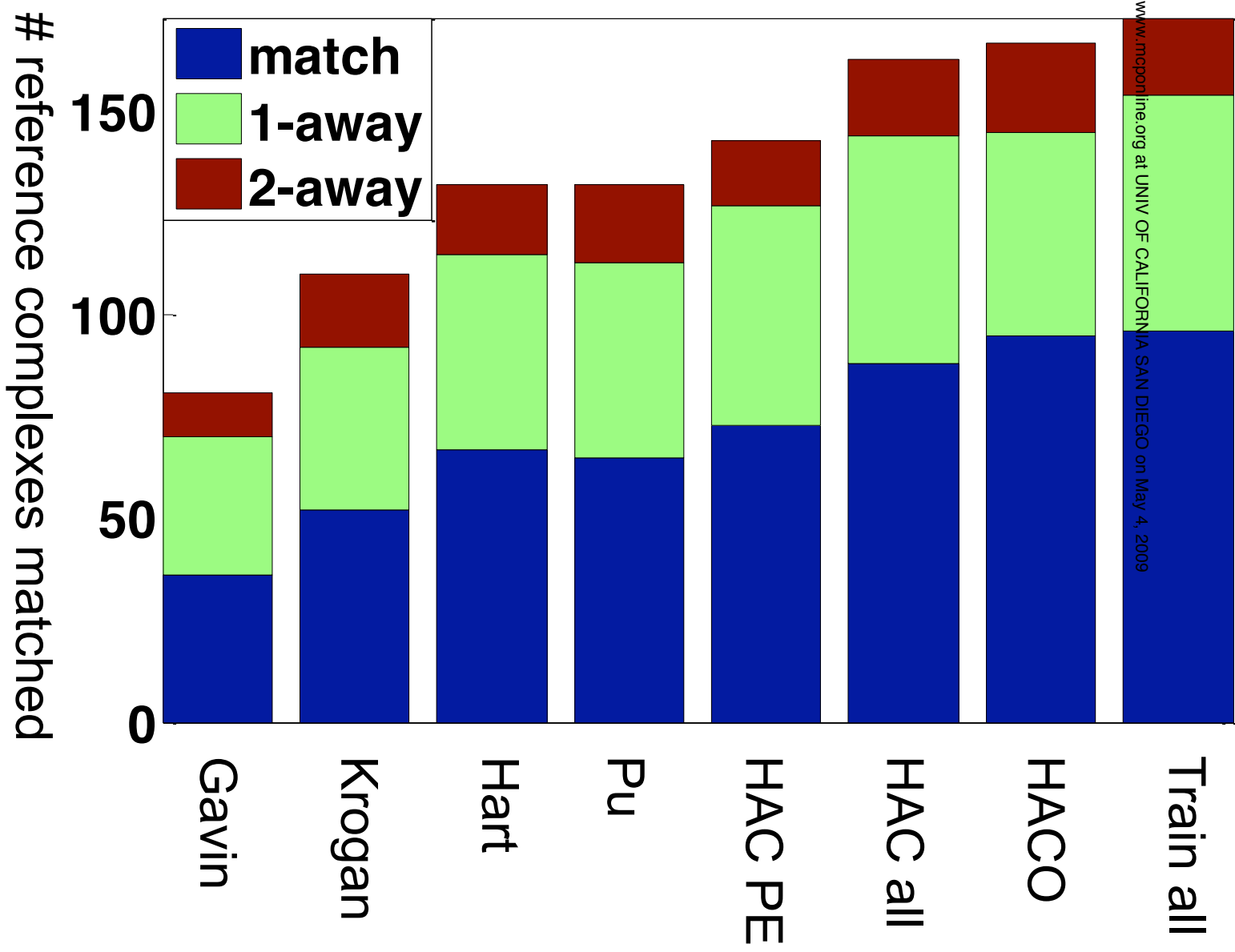
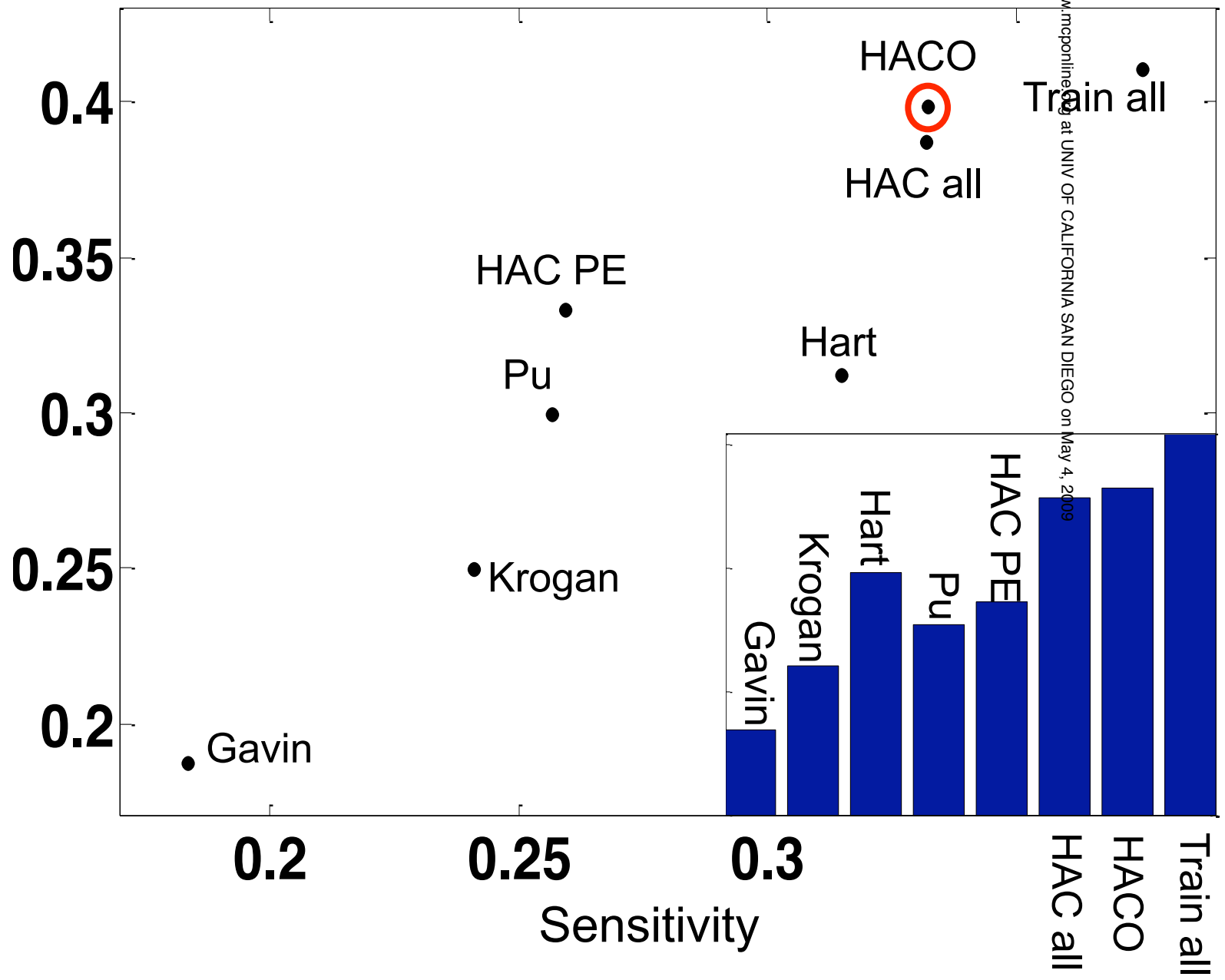


Figure 1a



Coverage



reference complexes matched

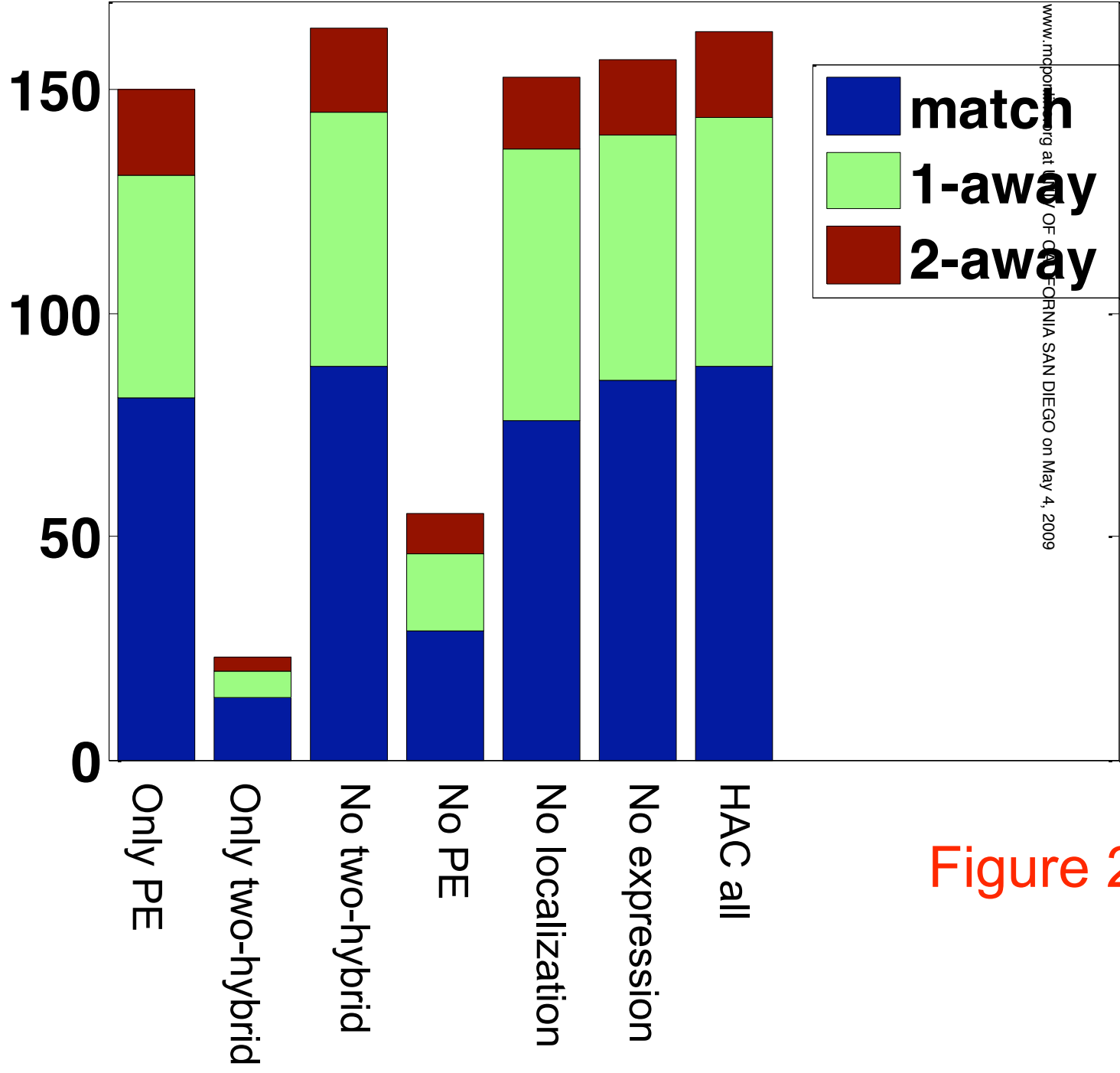
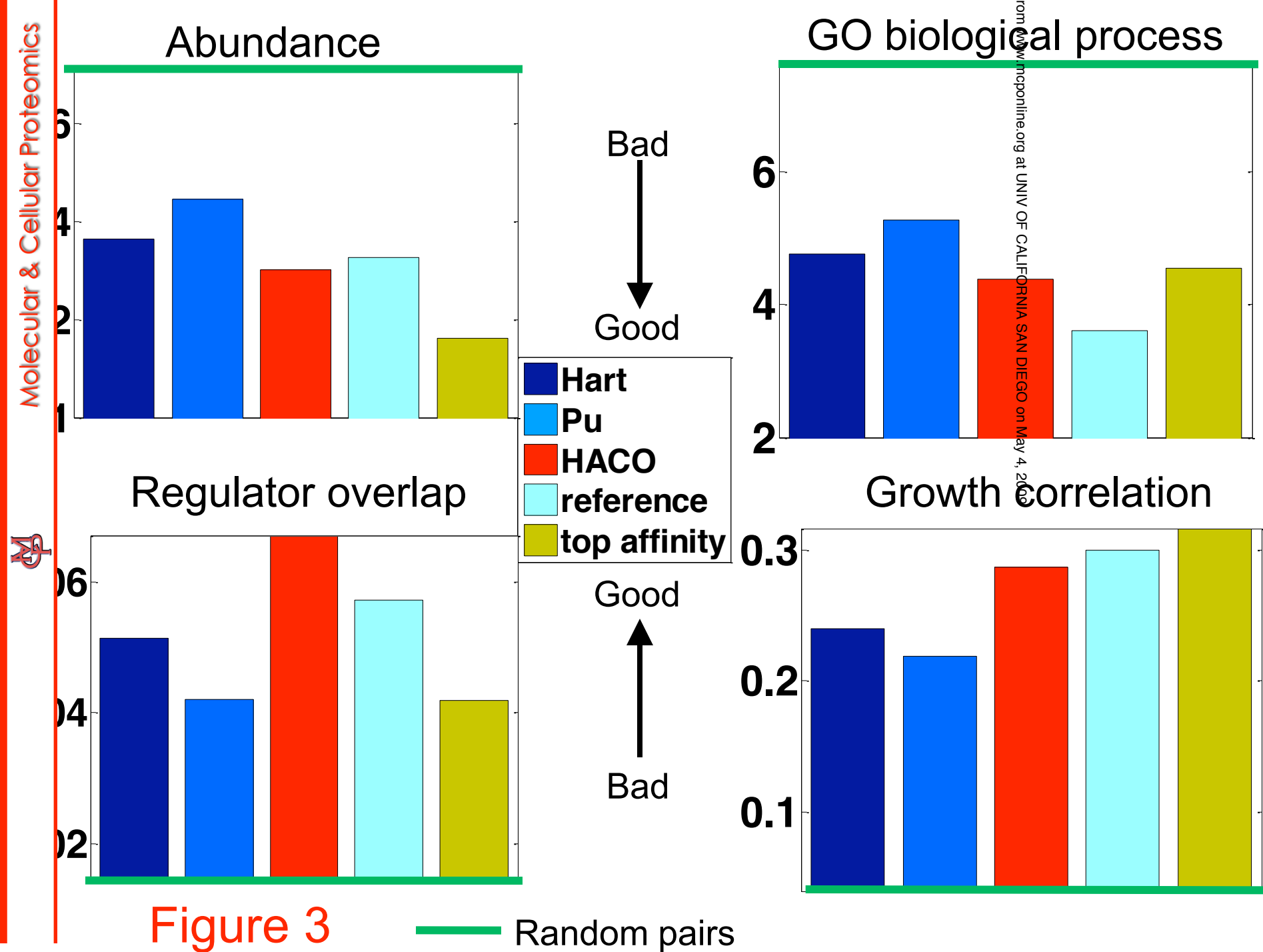


Figure 2



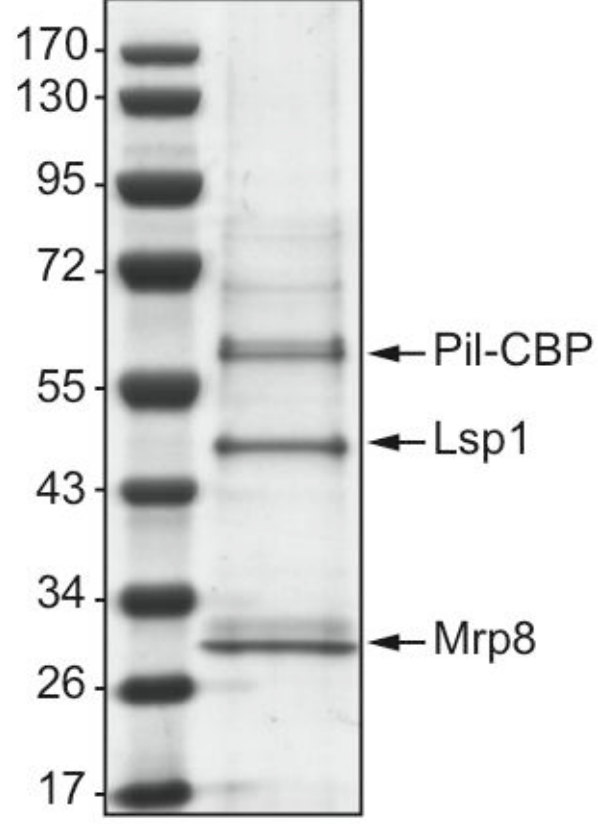
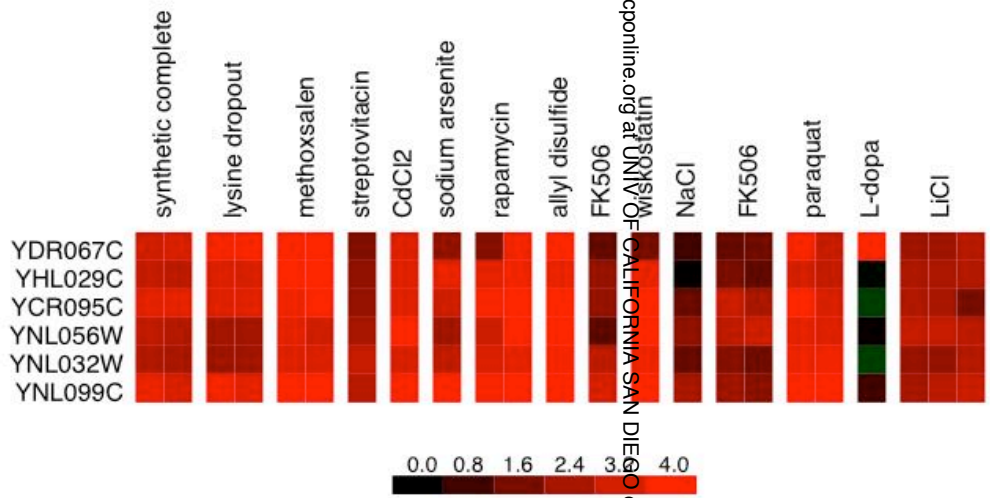
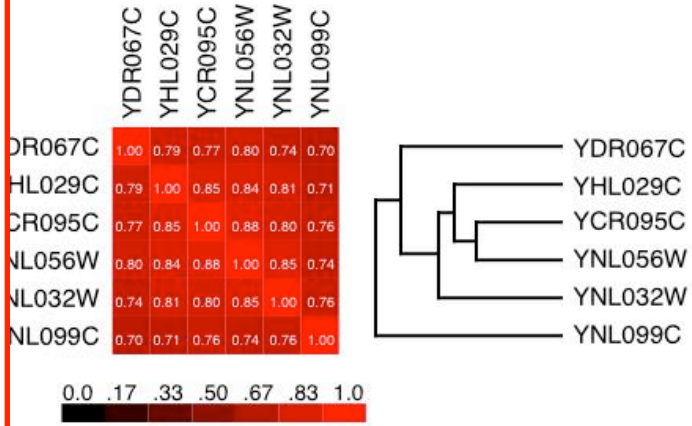
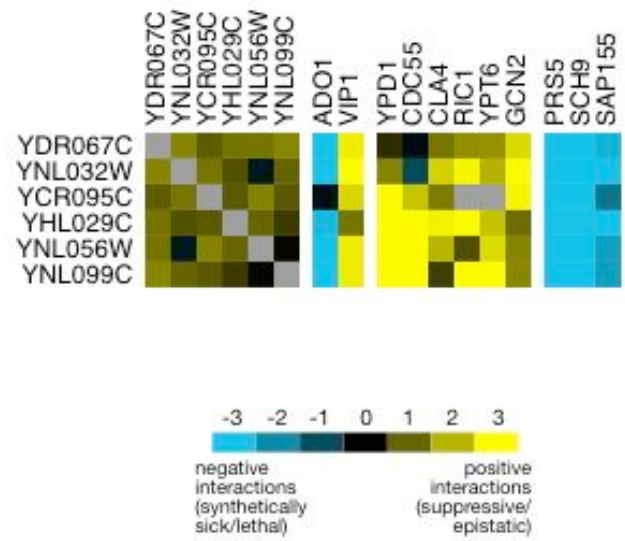
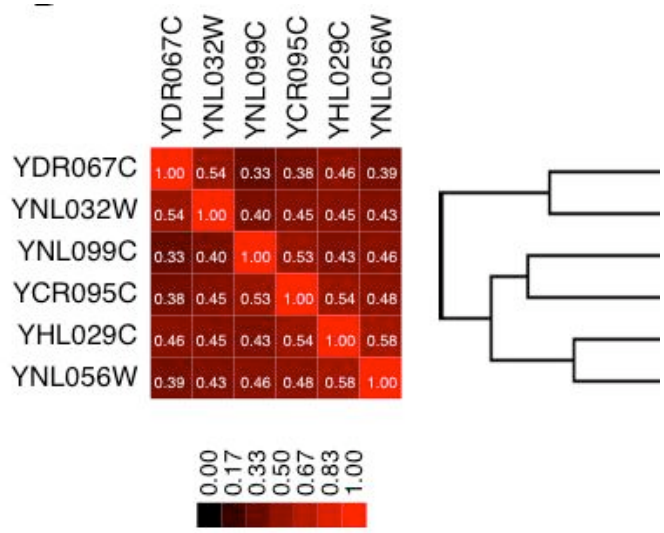


Figure 4a

b)



c)



www.molproteomics.org at UNIV OF CALIFORNIA SAN DIEGO on May 4, 2009

Figure 4b,c

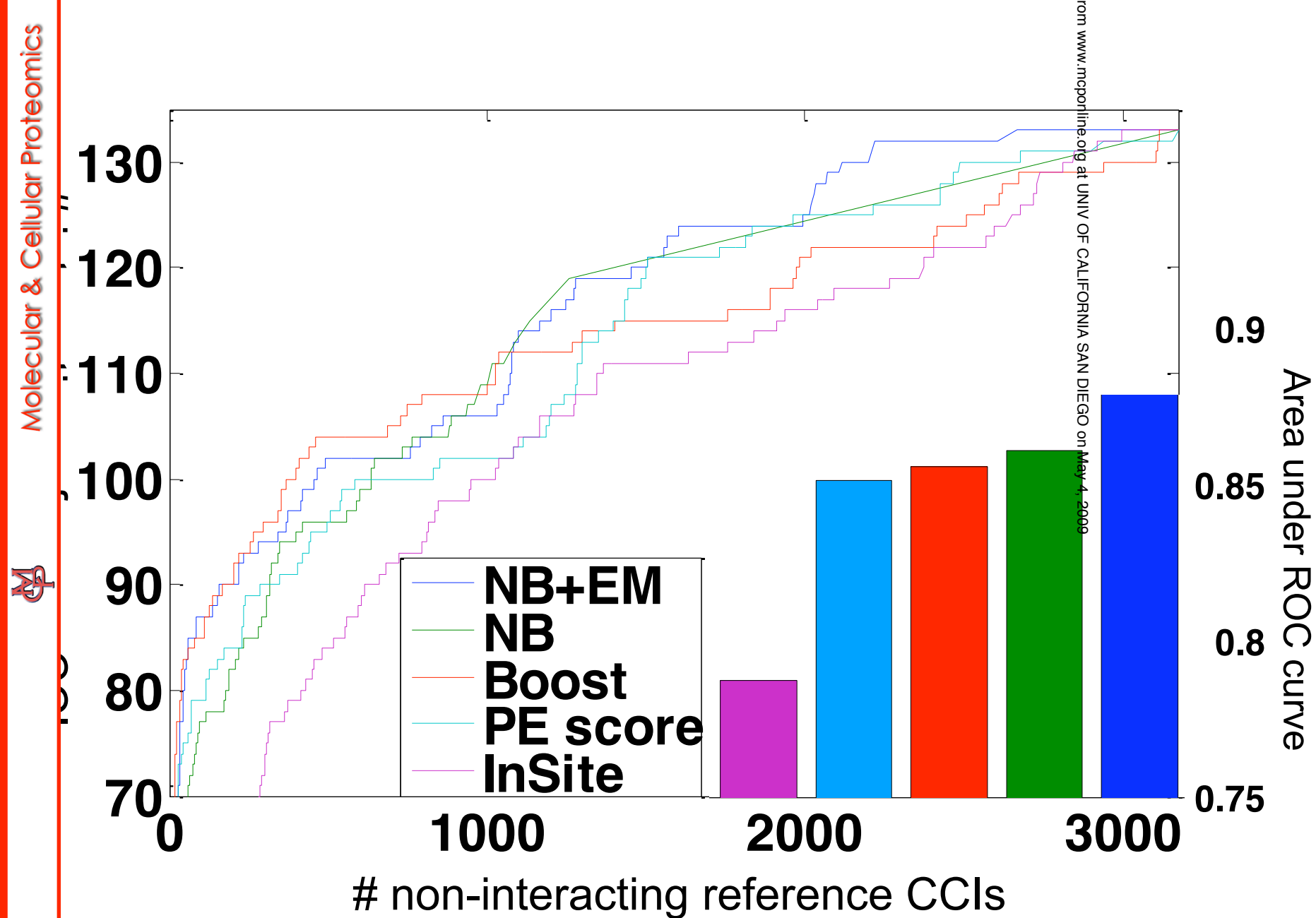


Figure 5a

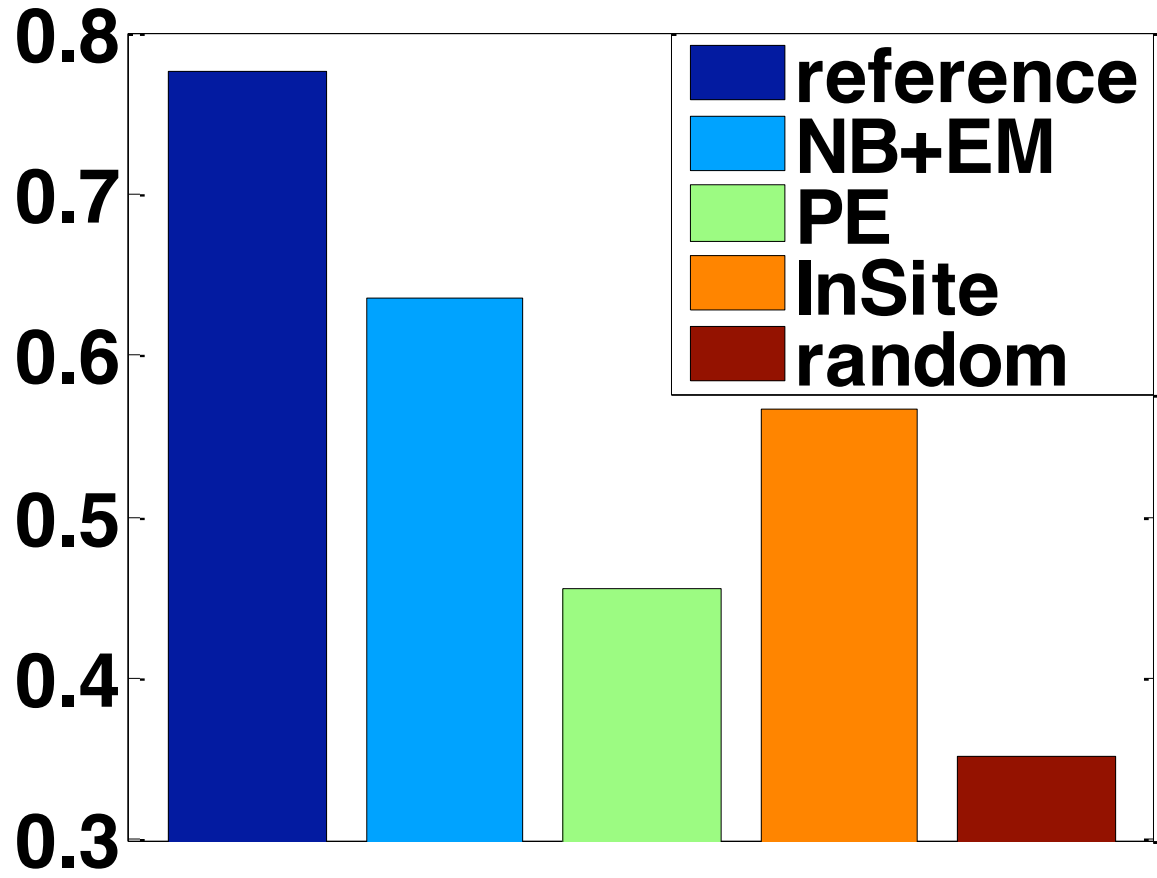
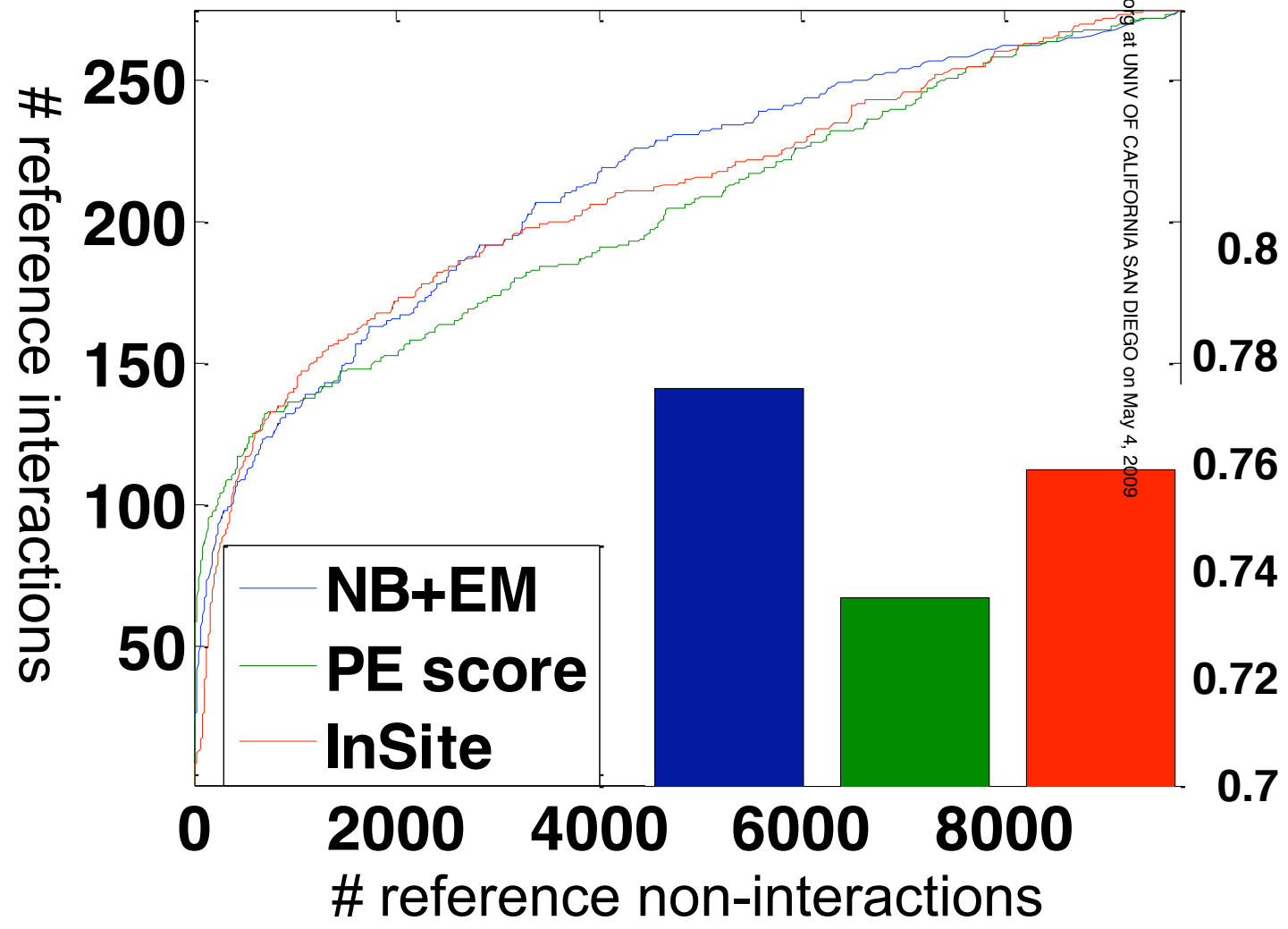
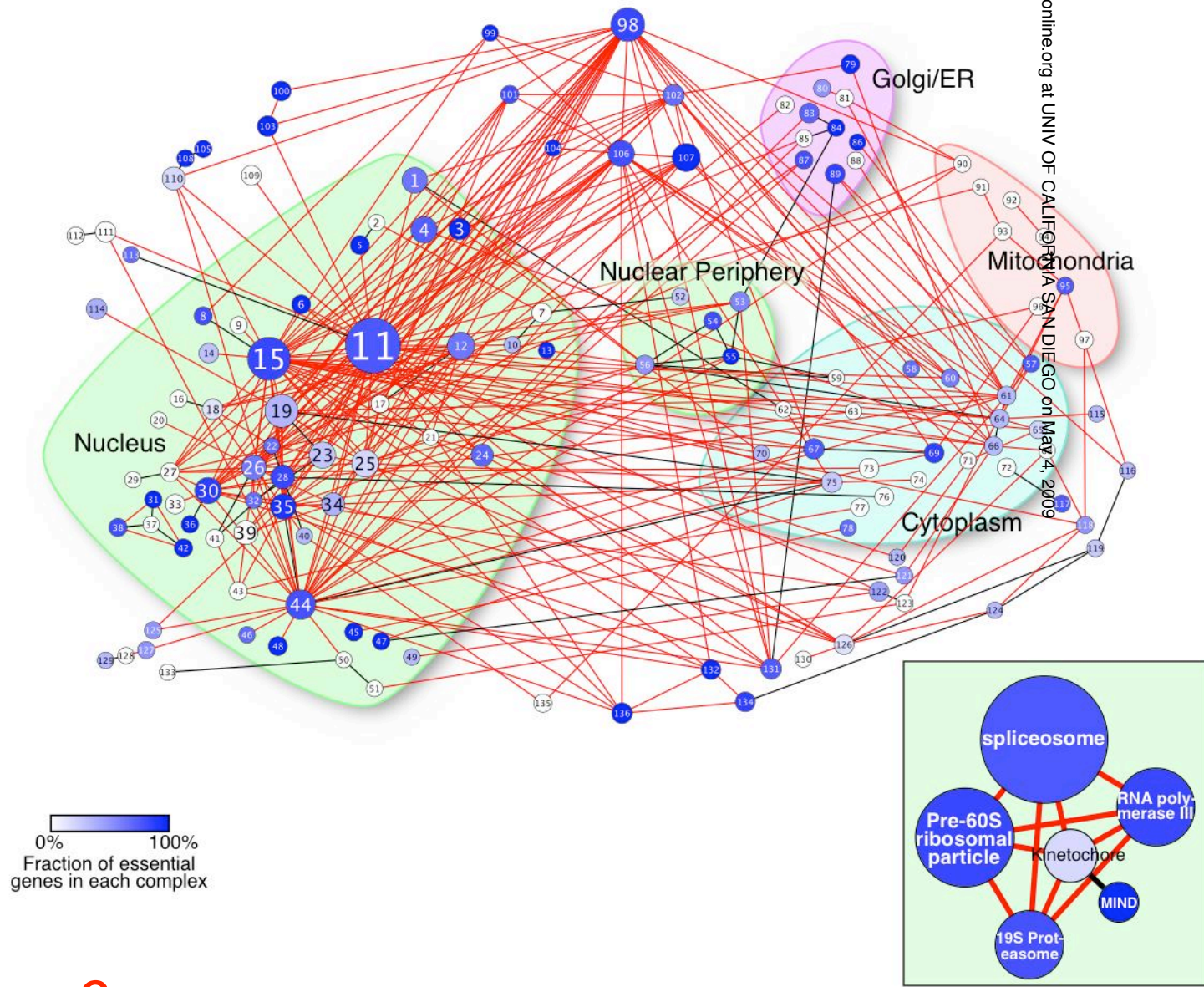


Figure 5b



from www.mcponline.org at UNIV OF CALIFORNIA SAN DIEGO on May 4, 2009

Figure 5c

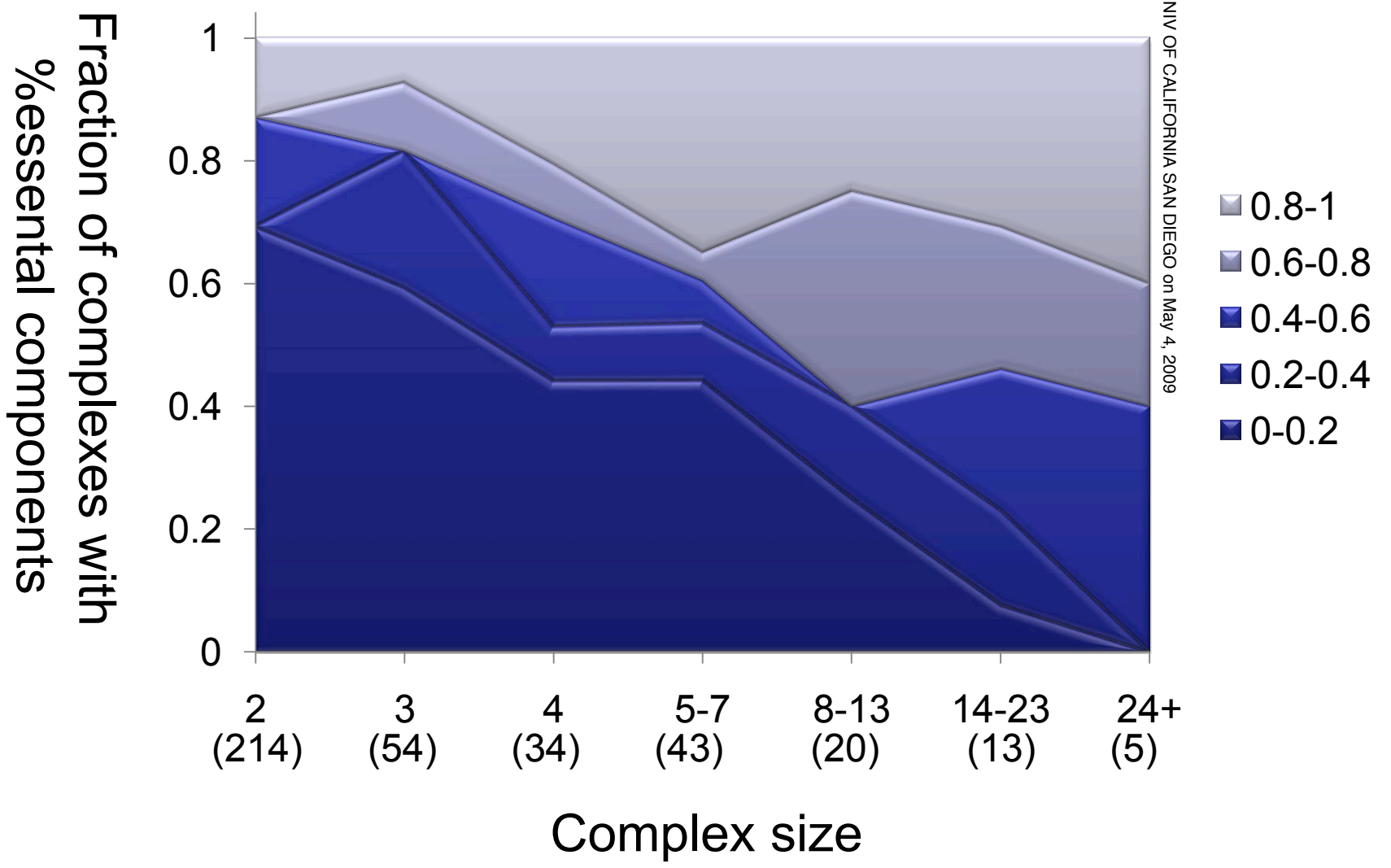


from www.mcponline.org at UNIV OF CALIFORNIA SAN DIEGO on May 4, 2009

Figure 6a

1	EXOSOME	67	EIF1
2	HAT1C	68	984
3	ribonuclease MRP complex	69	EIF2
4	1104	70	KG_58
5	ORC	71	1098
6	TFIIH	72	Gim complexes prefoldin
7	Nucleosomal protein complex	73	KG_118
8	1147	74	nascent polypeptide-associated complex
9	987	75	CCR4_NOT
10	1021	76	ELONGATOR
11	1126	77	signalosome complex
12	RSC	78	1095
13	RFA	79	Signal recognition particle (SRP)
14	TREX	80	Sec62/Sec63
15	1088	81	alpha-1,6-mannosyltransferase complex
16	HIRC	82	AP-1 adaptor
17	ISW1	83	TRAPP1
18	SWI/SNF	84	COPII_COAT
19	MEDIATOR	85	806
20	HTT-C	86	GPI-anchor transamidase complex
21	chromatin accessibility	87	oligosaccharyl transferase
22	TFIIF	88	AP-3
23	SAGA	89	925
24	APC	90	903
25	1125	91	oxoglutarate dehydrogenase
26	1004	92	respiratory chain complex IV
27	SN_SET3_COMPLEX	93	Pyruvate dehydrogenase
28	RNAPII	94	respiratory chain complex III
29	955	95	1130
30	mRNA cleavage and polyadenylation specificity factor	96	F0/F1 ATP synthase (complex V)
31	PROCOM_RAD24_RFC2_RFC3_RFC4_RFC5	97	1022
32	954	98	RNA polymerase III
33	COMPASS	99	Noc1p-Noc2p complex
34	NUA4	100	Exocyst
35	20S_PROTEASOME	101	1003
36	mRNA cleavage factor	102	997
37	PROCOM_CTF18_CTF8_DCC1	103	Exocyst
38	ALTERNATE_RFC	104	990
39	KG_15	105	SPINDLE_POLE
40	TORPEDO	106	1106
41	PAF1C	107	1129
42	RFC	108	MIND
43	CHD1_CKII	109	HOPS complex
44	19S_PROTEASOME	110	KINETOCHORE
45	nuclear condensin complex	111	KG_152
46	delta DNA polymerase	112	965
47	DNA polymerase alpha (I) - primase complex SN_PRIMASE	113	KG_147
48	TFIIIC	114	Golgi transport
49	TRAMP complex	115	1070
50	MSH2/MSH6 complex	116	963
51	KG_168	117	gamma-tubulin
52	KG_94	118	1042
53	NUP84 complex	119	989
54	924	120	1107
55	NSP1 complex	121	epsilon DNA polymerase
56	1101	122	Septin filaments
57	1148	123	1039
58	methionyl glutamyl tRNA synthetase	124	898
59	1015	125	1020
60	1025	126	1036
61	877	127	878
62	SKIC	128	heterotrimeric G-protein
63	KG_120	129	866
64	1066	130	879
65	eEF1	131	KG_70 COP1
66	1138	132	chaperonin-containing T-complex
		133	1127
		134	Arp2p/Arp3p
		135	retromer complex
		136	chaperonin-containing T-complex

Figure 6b



from www.mcponline.org at UNIV OF CALIFORNIA SAN DIEGO on May 4, 2009

Figure 7a

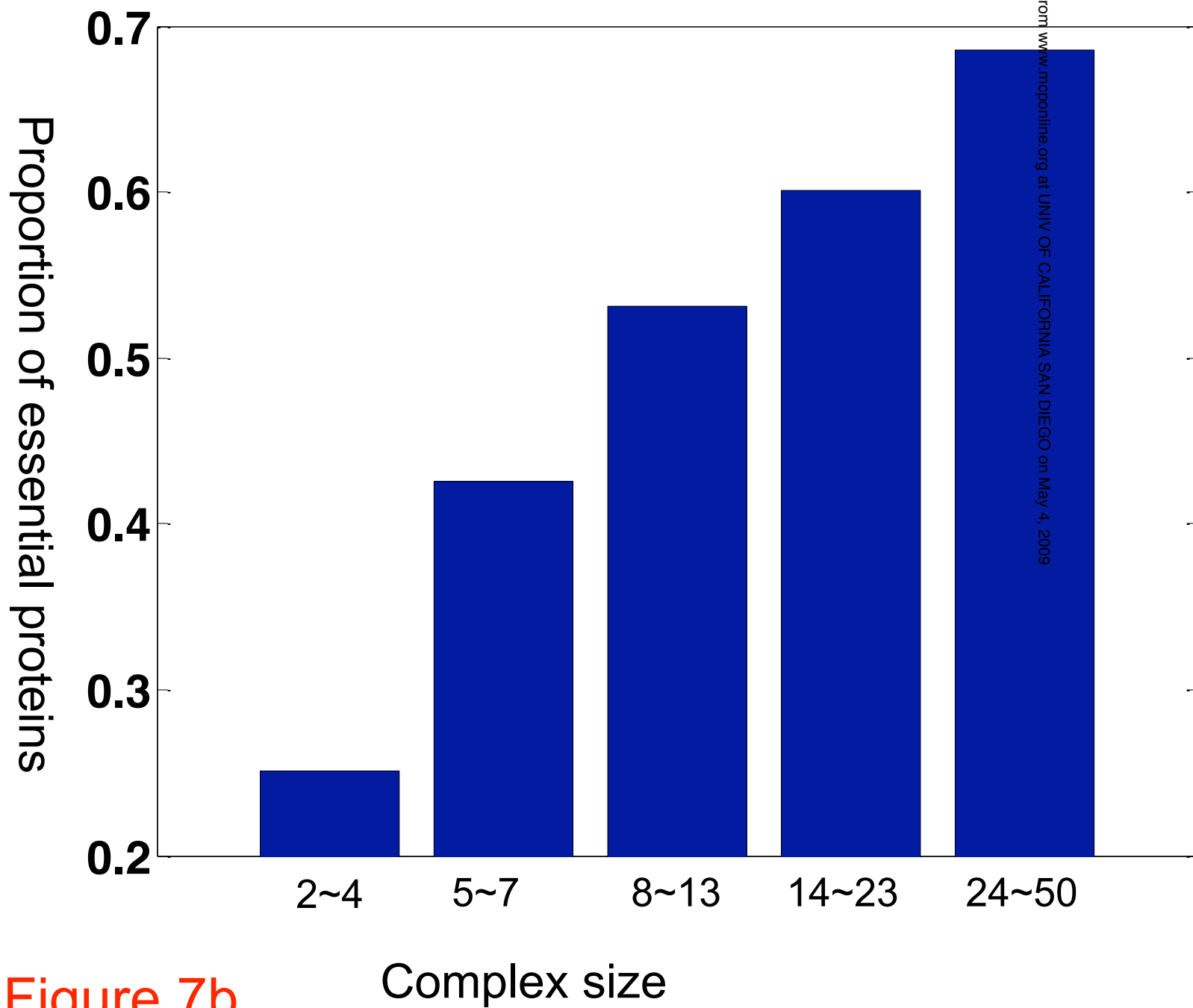
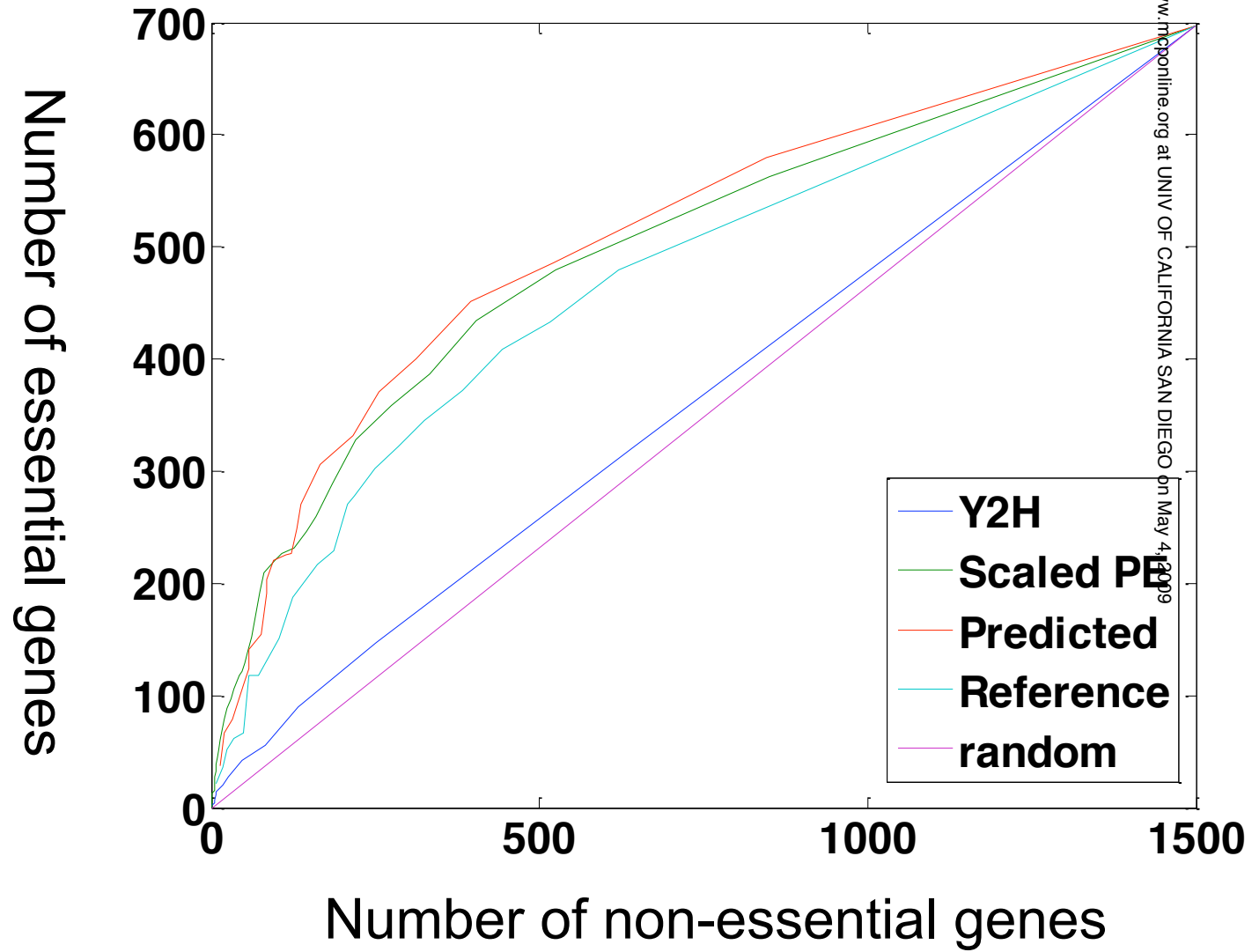


Figure 7b



from www.mcponline.org at UNIV OF CALIFORNIA SAN DIEGO on May 4, 2009

Figure 7c



Supplementary Figures

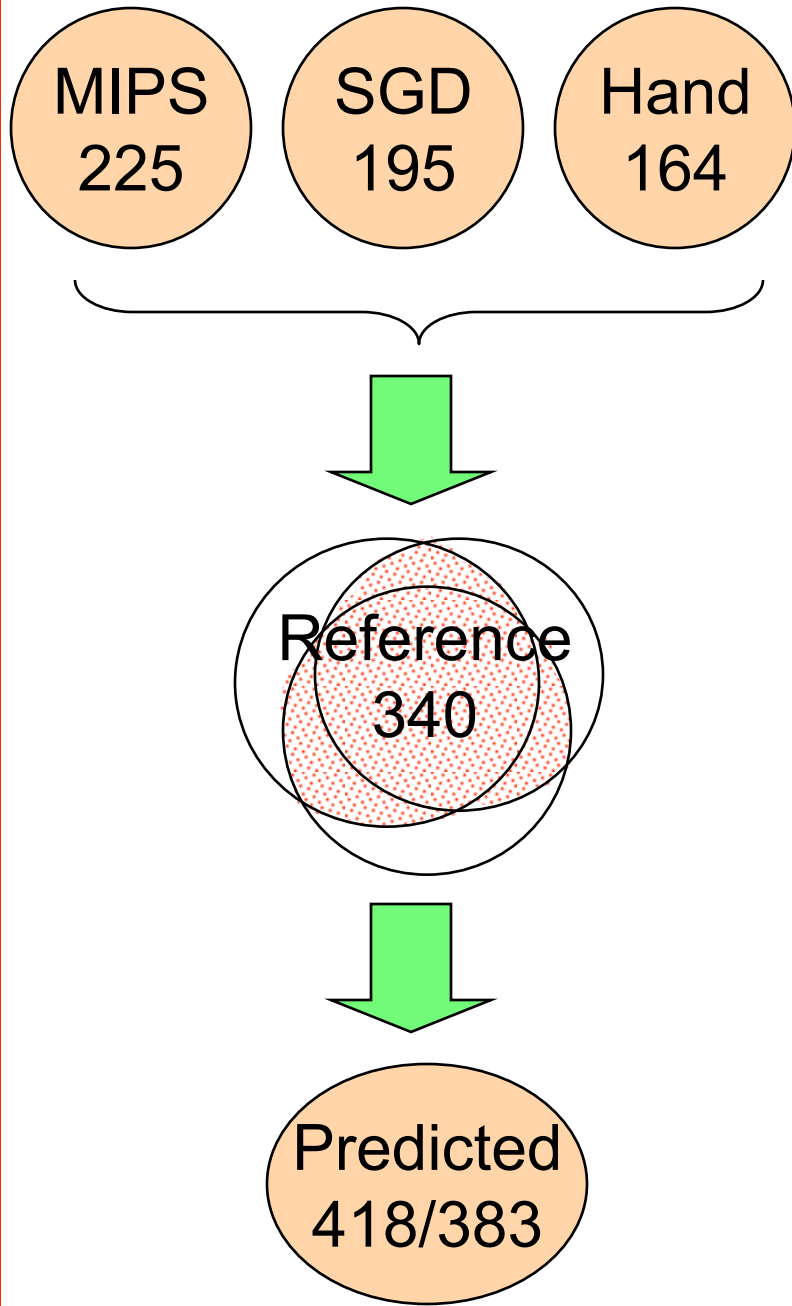
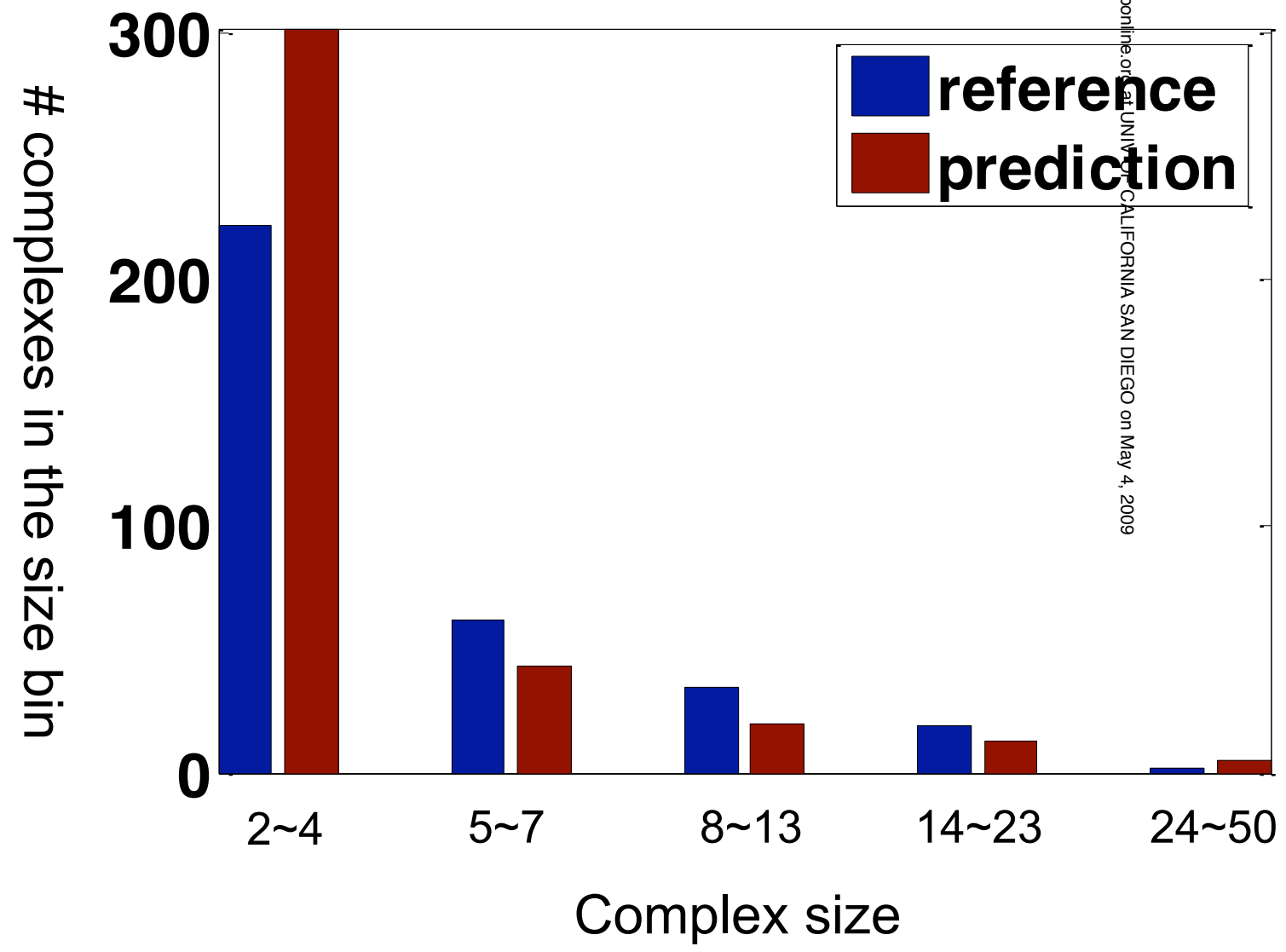


Figure S1(a)



from www.mcponline.org at UNIV OF CALIFORNIA SAN DIEGO on May 4, 2009

Figure S1(b)

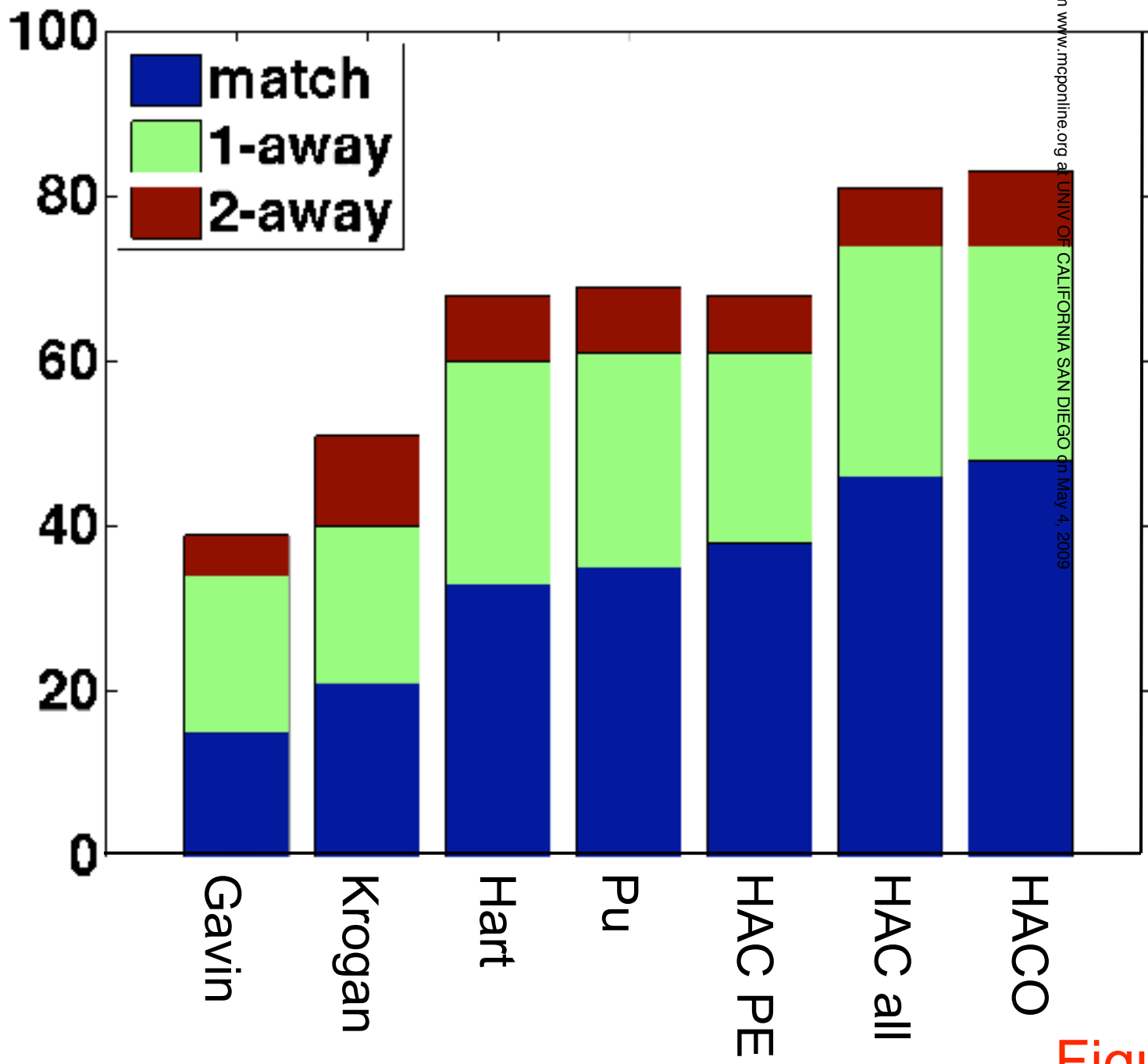
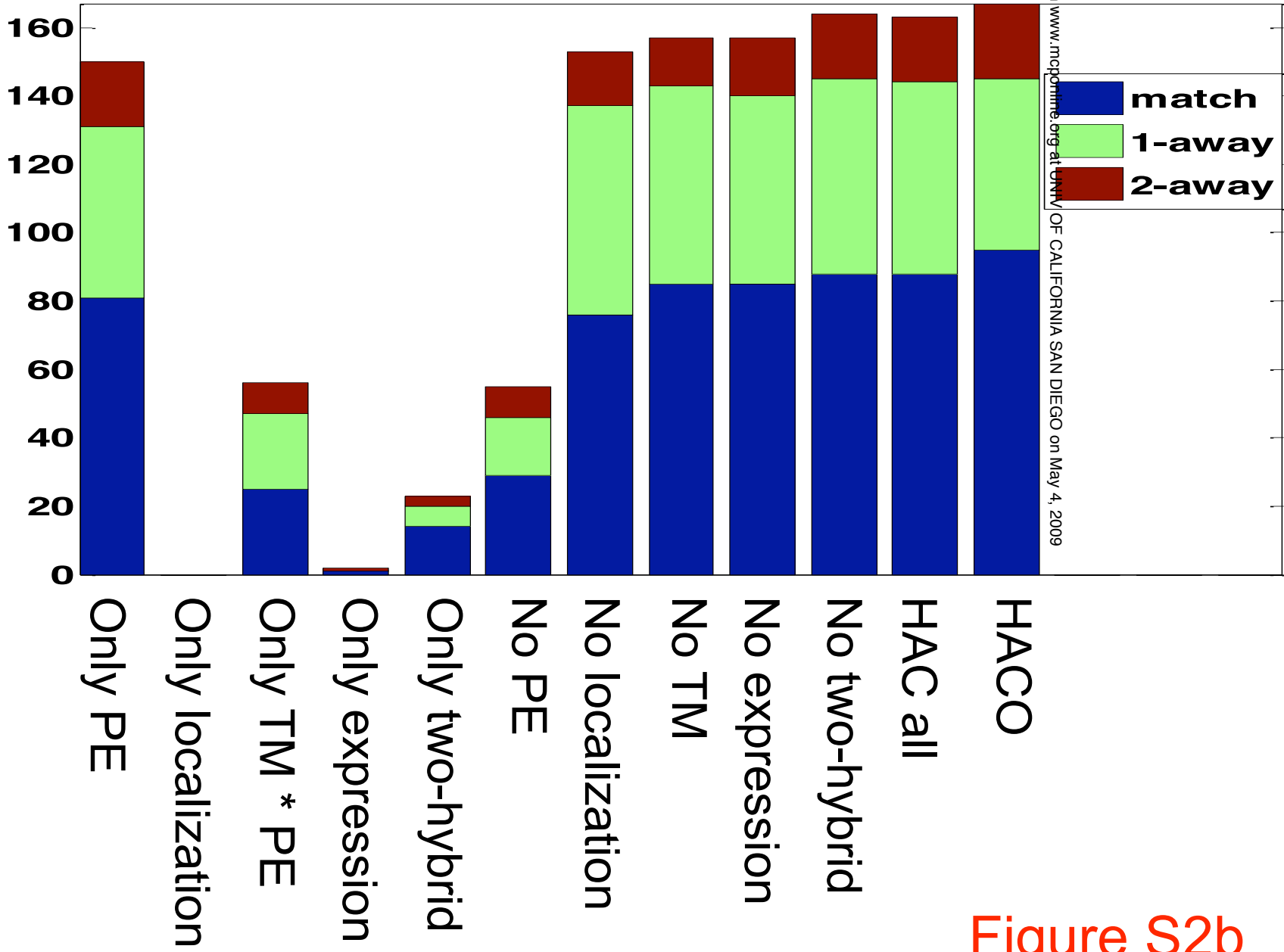


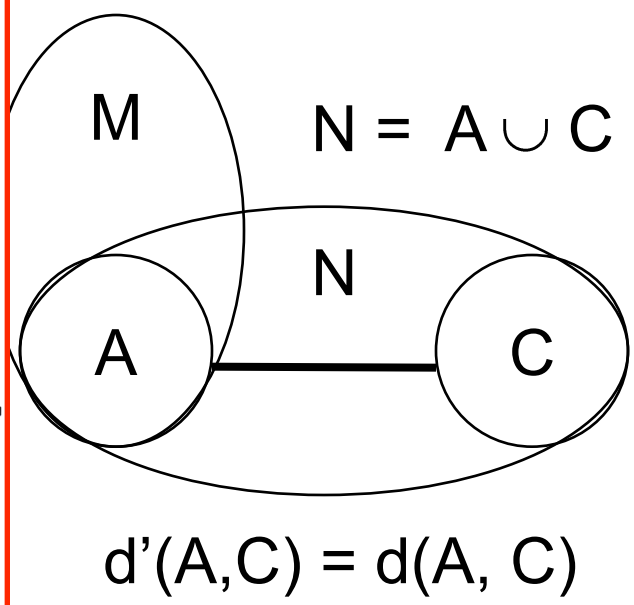
Figure S2a

reference complexes

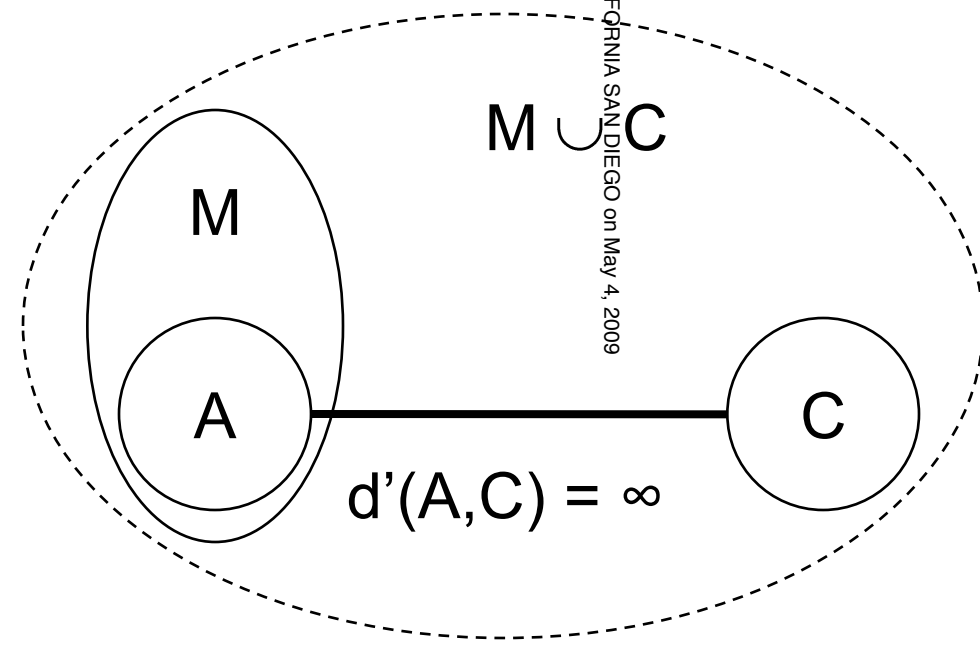


from www.mcp.jhu.edu on May 4, 2009

Figure S2b



is almost as coherent as M



N much less coherent than M
M \cup C preferred over N

Figure S3

Jaccard coefficient:

$$|C_i \cap R_j| \div |C_i \cup R_j| = 3/5$$

Hamming distance:

$$|C_i \cup R_j| - |C_i \cap R_j| = 5 - 3 = 2$$

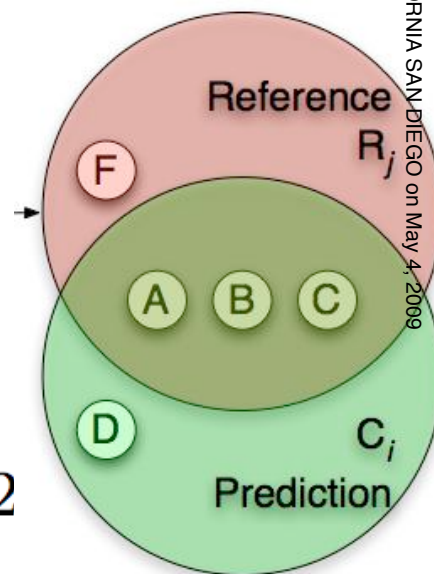


Figure S4