Markov Random Field Models of Transient Interactions Between Protein Complexes in Yeast

Boyko Kakaradov
Department of Computer Science, Stanford University
June 10, 2008

Motivation:

Mapping all transient interactions between protein complexes in a cell is an open problem in Systems Biology which has recently been illuminated by the availability of noisy but comprehensive maps of the yeast proteome. Several works have analyzed and clustered these maps into functional complexes, but have not identified transient interactions between them. We combine insights from both of these approaches into a statistical framework that can infer both physical interactions and other functional relationships between protein complexes.

Contribution:

We present a framework for combining protein-level data into meta-features which capture the relationship between pairs of complexes based on the relationships between their member proteins. The biological relevance of these high-level features is determined by a Naive Bayes classifier which predicts complex-complex interactions (CCIs) and a Quadratic classifier which predicts complex-pathway participation with reasonable accuracy. We use these predictions as priors in a Markov Random Field (MRF) model which captures context-specific features of the data which escape the NB classifier. MRFs are well suited for this purpose because, unlike conventional classifiers, they can be used to perform inference over all data instances collectively, enabling our model to capture domain-specific structure such as transitivity in the CCI network and other types of relationships such as common regulation and function between complexes.

Conclusion:

Our MRF models outperform their respective baseline Naive Bayes and Quadratic classifiers on both hold-out labeled annotations and novel unlabeled complex-complex pairs. The latter are perfect candidates for laboratory validation due to the high confidence in positive predictions achieved by our transitive interaction model.

1. 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, faithful reconstructions of the entire set of complexes in the cell as well as the interactions between them are essential to identifying the function of individual proteins and complexes and their role as building blocks for understanding the higher level organization of the cell.

Complexes of proteins are the fundamental functional units of cellular pathways, and transient interactions between them are the vehicles which carry out their function to support virtually all cellular processes. Until recently, cellular pathways were mapped out exclusively via focused small-scale experiments at a great expense of time and labor. However, the growing field of systems biology aims to understand biological processes from an integrative rather than reductionist perspective with the help of high-throughput experiments and computational analysis. The field is being revolutionized by the increasing availability of enormous amounts of genome-wide measurements from high-throughput assays in conjunction with sophisticated statistical methods that deal with the noisy data and computational models that integrate multiple sources of evidence.

Recently, two high-quality data sets used tandem affinity purification followed by mass spectrometry (TAP_MS) to measure pairwise chemical interactions between many of the proteins in yeast [1,2]. Several works [3,4,5] cluster these maps (in conjunction with other high-throughput data) into disjoint or overlapping sets identified as complexes. While such approaches are able to identify the protein members in each complex with increasing accuracy, they give only a partial view of functional landscape in the yeast cell. This work focuses on determining the transient interactions between those complexes in order to complete the picture by specifying the way in which they carry out their function. The current knowledge of CCI networks is limited to well-studied pathways which are hand-curated from focused and labor intensive small-scale experiments [12]. Although it is biased towards a small set of well-known interactions, we observe this knowledge as ground truth and complement it with less conclusive evidence from multiple high-throughput assays in a principled semi-supervised approach to mapping out the entire CCI network in yeast.

A Markov Random Field is a probabilistic graphical model that efficiently represents the joint probability of a set of random variables by encoding dependencies between them. Such dependencies can be learned from data or derived from prior knowledge about the domain which is modeled. Unlike a standard classifier, an MRF enables collective inference over the entire set of known and unknown variables, which makes it particularly suitable to the task of modeling complex-complex interactions. A recent computational analysis of protein-protein interactions (PPIs) in yeast [6] proposed several MRF models that effectively captured prior knowledge about transitivity in the PPI network and mitigated

noise in the high-throughput data by including hidden variables.

Inspired by the statistical power of those MRF models for PPIs, my earlier work examined the benefit of adding more data and domain structure to capture the intuition that proteins that are regulated by a common transcription factor (TF) should be more likely to interact. The extended model, informed by Chromatin Immuno-Precipitation (ChIP) assays [19], achieved only slightly better performance most likely because transcription regulation works on a functional level and thus co-regulation would be a property more relevant to CCI models. This consideration sheds light on the broader relationship between physical interactions and common function between complexes. While not as specific as ChIP-chip data, expression microarray measurements have been widely used to infer functional relationships [15]. Using accurate but incomplete annotations of genes in protein complexes [23] and cellular pathways [12] we define a co-activation score between them based on the correlation of the expression profiles of their constituent genes (see Section 3.3 for more details). Elevating the biological intuition that co-expressed genes have a functional relationship [18], we create another MRF model which captures to what extent complexes which share a common function also exhibit physical interactions.

The rest of this report is organized around presenting an abstract treatment of the baseline and Markov Random Field models in Section 2, followed by technical details of the data used for those models in Section 3. Section 4 contains a brief summary of our current results, which are analyzed in Section 5.

2. Methods

2.1 Construction of complex-complex features from protein-protein data:

Complexes are not necessarily disjoint sets of proteins. We use several high-throughput assays (detailed in Section 4: Experiment Setup) which define pairwise scores over proteins: a real number for each pair of distinct proteins. Given the protein contents of two complexes C_i and C_j, we generate the set S_{ij} of pairwise scores between all pairs (p_1,p_2) in $\{C_1 \times C_2 \mid p_1 \neq p_2\}$. Finally, we calculate the set of features for each pair of complexes C_i and C_j as:

$$F_{ij} = \{\text{sum}(S_{ij}), \text{avg}(S_{ij}), \text{min}(S_{ij}), \text{max}(S_{ij}), \text{sum}(\text{top}(k,S_{ij})), \text{avg}(\text{top}(k,S_{ij}))\}$$

where sum,ave,min,max are self-explanatory, and top(k,S) returns the k largest scores in S. We used the last operator with a parameterization of k=3 in order to capture the notion of core proteins in a complex as the average size for a complex is 4.9

2.2 Naive Bayes classification of complex-complex interactions:

Concatenating the feature sets F_{ij} for each assay defines a feature vector for each pair of complexes (C_i, C_j) which can be used in a standard classification framework such as logistic regression [7] or Support Vector Machines (SVMs) [8]. In order to use the features in a Naive Bayes (NB) classifier which assumes conditional independence between features given the instance label, only the most predictive of the features in F_{ij} is selected for each assay, as determined by the weights learned by logistic

regression or an SVM. The NB classifier calculates the probability that two complexes interact as the posterior of the binary interaction variable given the observations at the continuous feature variables modeled by independent Gaussian distributions. Given this marginal probability, we can classify the complex pair as interacting or not and but lose information on how confident we are in our prediction. Instead, we use the marginals from the NB model as priors in a joint MRF model.

2.3 Quadratic Discriminant Analysis for classification of complex-pathway participation:
Quadratic Discriminant Analysis (QDA) is a generalization of Linear Discriminant Analysis (LDA) which can be used for binary classification when the data in each class is normally distributed with a full-rank covariance matrix [24]. In contrast, a Naive Bayes classifier assumes that the covariance matrix is diagonal. Using gene expression measurements processed according to Section 3.3, for each complex-pathway pair, we construct a 19-dimensional feature vector of co-activation scores (corresponding to 19 experimental categories). Using standard 4-fold cross-validation, we obtained a classifier whose separating hyperplane is a quadratic surface in the feature space. This Quadratic classifier exhibited superior performance over other common frameworks such as Logistic Regression and SVM. Therefore, we chose it as the baseline model for functional annotation of protein complexes and used its outputs as priors in the joint MRF model described in Section 2.4

2.4 Joint MRF models:
We construct an MRF model of complex-complex interactions called Interaction MRF, which has a single type of nodes and three types of potentials:

1) Nodes I_{ij} corresponds to the binary random variable encoding an interaction between complex i and complex j with value 1 and a non-interaction with value 0
2) A fixed interaction node potential for each pair (i,j) of complexes: $f_{ij}(I_{ij}) = Pr(I_{ij})$ which is derived from the Naive Bayes model (see Section 2.2)
3) A learned interaction node potential shared by all nodes I_{ij} to offset bias in labeled data
4) A learned transitive interaction triplet potential $g(I_{ij}, I_{jk}, I_{ik})$ shared for all values of i,j,k. It represents the intuition that if complexes i and j interact with a common complex k, then they should be more likely to interact with each other.

We construct a second model called Participation MRF, which in addition to the CCI nodes (1) and their singleton interaction potentials (2) and (3) has another type of node and three types of potentials:

5) P_{it} corresponds to the binary random variable encoding participation of complex i in pathway t with value 1 and lack thereof with 0.
6) A fixed participation node potential for each pair (i,t) of complex i and pathway t $f_{ij}(P_{it}) = Pr(Pit)$ which is is derived from the QDA model (see Section 2.3)
7) A learned participation node potential shared by all nodes Pit to offset bias in labeled data
8) A learned co-participation triplet potential $g(I_{ij}, P_{it}, P_{jt})$ shared for all values of i,j,t. It represents the intuition that if complexes i and j physically interact, then they should be more likely to participate in the same pathway.

During training, all nodes are observed (except those corresponding to unlabeled CCI variables), all fixed potentials remain unchanged, and the shared potentials are estimated by the a MAP-based learning algorithm described in Section 2.5

During testing, we hide the labels of the query variables in the hold-out set. The query variables for the the Interaction MRF are I_{ij} and those for the Participation MRF are P_{it}. We then perform inference over the entire model to predict MAP assignments to the query variables, and any hidden CCI variables corresponding to unlabeled complex pairs. All potentials remain fixed, as there is no learning.

2.5 Inference and Learning in MRFs:

Prediction of novel (or hidden) complex-complex interactions is made via inference in our MRF model. There are two types of inference queries: *conditional probability* and *most likely assignment*, that can be posed for the unobserved variables in a graphical model given the model's parameters and any observed data. The MRF models proposed by Jaimovich *et. al.* [6] use the first type of query to jointly calculate marginal probabilities for each complex-complex interaction variable. They use Loopy Belief Propagation (LBP), a standard inference algorithm which is approximate due to the triplet structure of the model. While achieving good results, LBP with its comparatively slow running time puts a limitation on the scope of their models. Because we desire to model complex-complex interactions on a larger proteome-wide scale, we use a faster inference algorithm based on the second type of query which is also known as *maximum a posteriori* (MAP) inference. We use a MAP inference algorithm that is based on quadratic pseudo boolean optimization (QPBO) [9]. It is guaranteed to find the exact MAP assignment when the model parameters follow specific submodularity constraints. Relaxation of those constraints results in conflicting assignments for some variables in the MRF. We fix this situation with a greedy algorithm that assigns any unresolved variables according to the maximum marginal for each value.

While we can perform inference in a graphical model given arbitrary initial parameters, our predictions are necessarily more accurate when the the model parameters are learned from training data. We use the Perceptron Learning algorithm [10] to estimate the parameters of our model from fully-observed data. As virtually all learning algorithms, Perceptron uses inference as a subroutine to guide its search through the space of parameter values. Depending on their initial values, parameters take several iterations to converge, which necessitates the use of a fast inference algorithm. When our model trains on unobserved data (unlabeled instances of CCIs), we initialize the free parameters (fixed parameters do not change) and use the hard-assignment Expectation Maximization [11] algorithm to alternate the inference (calculating the MAP given the current values of the parameters) and learning steps (estimating the parameters as to maximize the current observation).

3. Experimental Setup

3.1 Sources of protein-protein scores:

As described in the Methods section, our model uses high-level features between pairs of complexes derived from protein-protein scores. Those scores are derived from seven different data sources: the purification enrichment (PE) score from the consolidated network of Collins et al. [3], a cellular component from a truncated version of the Gene Ontology (GO) [13], trans-membrane proteins [14], co-expression [15], and yeast two-hybrid (Y2H) interactions [16,17], genetic interaction [18], and transcription regulation [19].

Our highest-coverage source regarding direct physical interaction between proteins comes from high-throughput TAP-MS data of the Gavin [2] and Krogan [1] data sets. The recent work of Collins et al. [3] 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 : 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.

The Gene Ontology (GO) cellular component hierarchy [13] 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. 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 [22], 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' (Y') to be the category one level down the path from Z to X (Y). We conclude that A and B belong to two different categories at X' and Y'. Thus, the smallest possible common category of A and B is X' \cup Y' assuming X' and Y' can form a coherent category by merging themselves. Therefore, our second feature is log(|X'\cup Y'|) .

A list of membrane proteins are obtained by parsing the trans-membrane annotations in SGD [14]. 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.

Yeast two-hybrid protein-protein interactions are obtained from the assays of Ito et al. [16] and Uetz et al. [17]. 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.

Microarray data were downloaded from Stanford Microarray Database (SMD) [15] on Dec. 5, 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 separately for each category and then pick the maximum coefficient. This processing is necessary because unlike stable co-complex interactions, transient interactions between proteins are highly dependent on the experimental condition.

Genetic interaction data in the form of Epistatic MiniArray Profile (E-MAP) scores from double mutant experiments were downloaded from the Interactome Database on April 11, 2007 [18]. These data consist of quantitative pairwise measurements of the genetic interactions between 743 yeast genes involved in various aspects of chromosome biology. We construct three protein-protein features for each pair: the E-MAP score of the pair, its absolute value, and the cosine distance between the genetic interaction profiles of the two proteins, in a similar fashion to the PE distance metric.

Transcription regulation data in the form of measurements from Chromatin ImmunoPrecipitation microarray (ChIP-chip) assays were downloaded on January 28, 2007 [19]. The data consists of p-values for binding affinities between transcription factors and binding sites associated with genes. Processed at a particular confidence threshold, it yields a binary vector signifying presence or lack of regulation by all TFs for each gene. For the Naive Bayes model, we derive two protein-protein features from this data: number of overlapping TFs between the two proteins, and the mutual information between the TF profiles of the two proteins.

3.2 Reference set of complex-complex interactions:

First, we compiled a list of 420 protein complexes, 81 of which are well known and the rest are high quality computational predictions [23] validated against reference sets hand-curated by experts, MIPS and SGD. From this list, we form all ${}_{420}C_2 = 87{,}990$ unordered pairs of distinct complexes. Of them, we have 133 positive labels which are trusted complex-complex interactions identified by combining a hand-curated set of 59 interactions with a set of 82 pairs with significant enrichment for small-scale interactions between their member proteins (see next paragraph). We have 3173 negative labels which are trusted non-interactions derived from all pairs between the 81 well-known complexes which were not identified as positive labels. The remaining 85104 pairs of complexes remain unlabeled, as we cannot conclusively say whether they form transient interactions or not. We used the Naive Bayes classifier to make marginal predictions for all pairs of complexes.

In order to expand our set of positive labels, we calculated the enrichment for small-scale physical interactions between all pairs of the 420 proteins. A list of small-scale interactions was downloaded from MIPS [20] and DIP [21] on March 21, 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.

3.3 Functional Annotation of complexes: participation labels and co-activation score

Functional annotation of complexes was established based on their participation in known pathways. Referencing pathway annotations from KEGG [12], we define a complex as participating in a pathway if at least half of its member proteins are annotated in that pathway. This definition yields 284 positive complex-pathway pairs containing 66 unique pathways and 120 unique complexes. We then sample 284 negative labels from the remaining 66*120 - 376 = 7,544 complex-pathway pairs.

Co-activation is a predictor of functional annotation derived from expression data. In order to construct the 19-dimensional feature vectors for the quadratic classifier in Section 2.3, we pre-process the expression data from SMD differently than when deriving the complex-complex features (Section 3.2). First we define a gene to be active in a particular experimental category if its activity is at least two standard deviations above the mean activity for all genes in that category. Then we combine the activity profiles of all genes belonging to some set. More specifically, we define a set of genes such as a complex or a pathway to be active in a particular experimental category if at least half of its member genes are active in that category. Finally, we calculate the co-activation score for each pair of complex and pathway in a particular experimental conditions as the product of their activations in that experimental condition. This processing wields a 19-dimensional feature vector for each complex-pathway pair, enabling us to use a wide variety of binary classifiers to predict which complexes participate in which cellular pathways.

3.4 Cross-validation of model predictions:

4-fold cross-validation was performed on the reference set of complex-complex interactions for the Interaction MRF and that of complex-pathway participations for the Participation MRF. For each fold, ¼ of the labeled set was held out. Our models were trained on the remaining ¾ of the labeled set, in addition to 1000 unlabeled CCIs for one instance of the Interaction MRF model. All models were tested on the held-out set in each fold, thereby producing unbiased results on the entire labeled set, when the predictions from each fold were combined. The standard test evaluation metrics of Sensitivity = TP / P and Specificity = TN / N were calculated from the trusted versus the predicted labels (where TP is the number of true positive predictions, TN is that of true negatives, P the number of positive labels, and N that of negative labels). The Naive Bayes and QDA models' marginal predictions were validated at all classification cutoffs, yielding a continuous Receiver Operating Characteristic (ROC) curve, while the MRF models' MAP inference output could only be evaluated to a point in ROC space.

4. Results

Figures 1 and 2 compare the classification performance of our Interaction and Participation MRF models with that of their respective baseline models. In Figure 1, the baseline Naive Bayes model ,whose area under the curve is 0.881, is compared to the equivalent disconnected MRF model, and to the Interaction MRF model, which was trained both with fully- and partially- labeled data. In Figure 2, the QDA model, whose area under the curve is 0.736, is compared to the Participation MRF (trained on fully-observed data only).

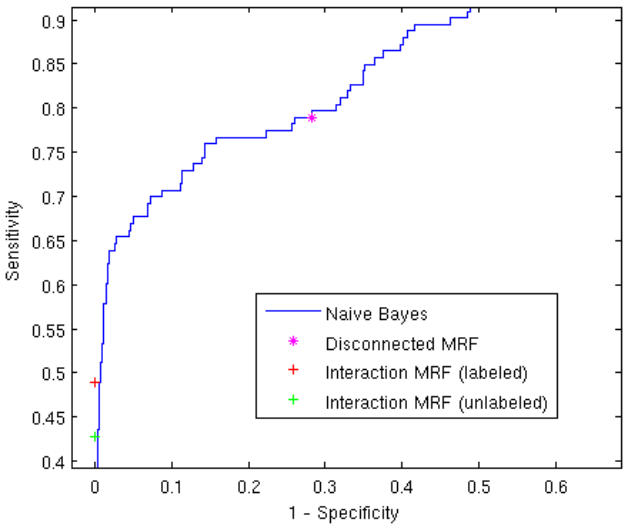

Figure 1 compares Naive Bayes model (blue curve) with Interaction MRF model (red and green cross) in ROC space. The MRF model was trained on fully labeled (red) and partially unlabeled (green) CCI instances.

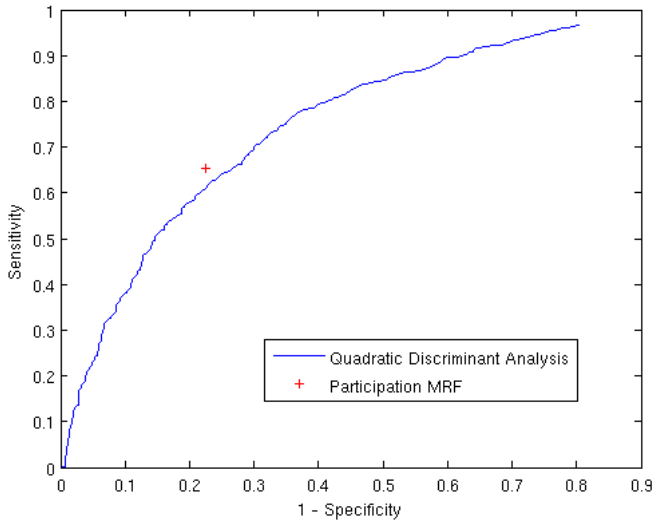

Figure 2 compares QDA model (blue curve) with Participation MRF model (red cross) in ROC space. The MRF model was trained on fully labeled data for both interaction and participation variables.

5. Discussion

5.1 MRF models benefit from domain-specific data

The above results show that our MRF models outperform their respective baseline models by encoding domain-specific dependencies in the data. As evidenced by Figure 1, the performance of our Interaction MRF model is slightly better than that of the baseline Naive Bayes classifier, in both the fully-supervised and partially unsupervised training regimes. Despite their smaller scope, the MRF models trained on fully labeled data perform slightly better on the hold out set than those trained on partially unlabeled data. This is due to the knowledge bias of the labeled interactions which was discussed in Section 1. More prominently, Figure 2 shows a relatively bigger gap in performance between the Participation MRF model and its baseline – QDA. This is due to the biological intuition that pathways are more or less a linear chain of interactions between pairs rather than triplets of complexes.

5.2 Specific predictions of Interaction MRF model

The limited gain for the Interaction MRF model is due to the absence of many transitive interactions in the labeled set. In fact, there are only 4 such sets of three pairs whose variables form a triplet clique. Despite the lack of enrichment for the Interaction triplet structure, the learned potentials for the interaction MRF model exhibit a strong preference for all-positive assignments (111) over two-positive plus a negative assignments (011). This enables the Interaction MRF to resolve an unlabeled variable with prior probability of being true equal to 0.3 from the NB model, because it is involved in 15 different 3-cliques whose other assignments are either labeled as positive or are very likely to be positive (with probability > 0.97 as estimated by the NB model). This variable corresponds to an alleged transient interaction between two complexes involved in the cell cycle pathway, and identifies a candidate for validation in the wet lab.

While the predictions of the Interaction MRF model fall above the baseline curve of the NB model, they are both strongly biased towards predicting negative labels due to the unbalanced reference set. This results in zero false positive predictions by our Interaction MRF model at the cost of many false negatives. The perfect Specificity of the MRF models leads us to conclude that they are very conservative in predicting transient interactions between complexes. This is a desirable property, as any positive predictions on unlabeled pairs of complexes correspond to very high-confidence candidates for wet lab validation. In fact, the MRF model makes just 5 positive predictions from the set of 1000 unlabeled complex-complex interactions. Of those, two pairs of complexes are assigned a low probability of interacting by the NB model. Both pairs contain a novel complex which is implicated in catalyzing the de-amination of adenosine monophosphate (AMP), an important molecule in energy transfer [14]. Its proposed interaction partners in the two pairs are the single protein URH1 (which cleaves bonds in nucleosides) and the well-studied 20S core particle of the PROTEASOME complex (which degrades unneeded or damaged proteins), both of which require energy to execute their function [14].

5.3 Specific predictions of Participation MRF model

We trained and tested the Participation MRF model only on labeled complex-pathway annotations because we were unable to distinguish potential negative labels from that of unlabeled instances. This is a common problem in systems biology which is due to the inherent lack of a high-confidence negative reference set. Therefore, unlike Section 5.2 we make no predictions of novel complex-pathway annotations in this analysis but focus on our model's ability to predict held-out positive labels. There are 46 out of all 284 gold positive complex-pathway pairs which are predicted incorrectly as false negatives by the QDA model, but are resolved as true positives by the Participation MRF model. 23 of them are given a prior of less than 0.1 by QDA, which suggests the utility of the Participation triplet structure. Of the 21 pathways involved in those 23 pairs, two contain more than one difficult false negative for QDA which is corrected by our MRF. The two pathways have very distinct functions: “Regulation of autophagy” and “Valine, leucine and isoleucine biosynthesis”, but share the property that at least two of the complexes participating in them interact with each other.

6. Conclusion

Identifying transient interactions between protein complexes on a genome-wide scale and cataloging them into cellular pathways are two primary challenges in Systems Biology. In this report, we presented two principled Markov Random Field models that address those challenges by encoding domain-specific knowledge of the complex-complex interaction network in yeast. While both MRFs outperform their respective baseline models, the Interaction MRF makes several confident predictions on potentially novel complex-complex interactions.

7. Acknowledgments

This work was supervised by Professor Daphne Koller and Ph.D. student, Haidong Wang, of the Stanford Artificial Intelligence Laboratory.

8. References

[1] Krogan, N. J. et al. Nature, 2006
Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*.

[2] Gavin, A. C. et al. Nature, 2006
Proteome survey reveals modularity of the yeast cell machinery.

[3] Collins, S. et al. Molecular Cell Proteomics, 2007
Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces cerevisiae

[4] Pu, S. et al. Proteomics, 2007
Identifying functional modules in the physical interactome of Saccharomyces cerevisiae

[5] Hart, G.T. et al. Bioinformatics, 2007
A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality BMC

[6] Ariel Jaimovich, Hanah Margalit and Nir Friedman. Journal of Computational Biology, 2006
Towards an Integrated Protein-Protein Interaction Network: A Relational Markov Network Approach

[7] Logistic Regression, Wikipedia article: http://en.wikipedia.org/wiki/Logistic_regression

[8] Support Vector Machine, Wikipedia article: http://en.wikipedia.org/wiki/Support_vector_machine

[9] Vladimir Kolmogorov, Carsten Rother. IEEE Transactions on Pattern Analysis and Machine Intelligence. July, 2007
Minimizing Nonsubmodular Functions with Graph Cuts-A Review

[10] Perceptron, Wikipedia article: http://en.wikipedia.org/wiki/Perceptron

[11] Dempster, A.P., Laird, N.M., and Rubin, D.B.
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. pp. 1-38.1977.

[12] Kyoto Encyclopedia of Genes and Genomes, Pathway Database: http://www.genome.ad.jp/kegg/pathway.html

[13] The Gene Ontology, Cellular Component Hierarchy for *Saccharomyces cerevisiae*:
http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/gene_association.sgd.gz?rev=HEAD

[14] SGD project. Saccharomyces Genome Database. ftp://ftp.yeastgenome.org/yeast/

[15] Demeter J, et. al. Nucleic Acids Research 2007
The Stanford Microarray Database: implementation of new analysis tools and open source release of software.

[16] Ito, T. et. al. Proceedings of the National Academy of Science of the USA, 2001.
A comprehensive two-hybrid analysis to explore the yeast protein interactome.

[17] Uetz, P. et. al. Nature, 2000.
A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae.

[18] Collins SR et. al. Nature, 2007
Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map.

[19] MacIsaac KD et. al. BMC Bioinformatics, 2006
An improved map of conserved regulatory sites for Saccharomyces cerevisiae.

[20] Mewes HW, et. al. Nucleic Acids Research, 2002
MIPS: a database for genomes and protein sequences.

[21] Ioannis Xenarios, et. al. Nucleic Acids Research, 2000
DIP: the Database of Interacting Proteins

[22] Cooper, Martin C. et. al. Semantic Distance Measures. Computational Intelligence, 2000

[23] Boyko Kakaradov et. al. Identifying Protein Complexes in *Saccharomyces cerevisiae*. Biocomputation at Stanford (BCATS) 2007

[24] Quadratic Classifier, Wikipedia article: http://en.wikipedia.org/wiki/Quadratic_classifier