

PICUPP: Protein Interaction Classification by Unlikely Profile Pair

Byung-Hoon Park^a, George Ostrouchov^a, Gong-Xin Yu^a,
Al Geist^a, Andrey Gorin^a, and Nagiza F. Samatova^{a,b}

Abstract

A computational approach that infers protein-protein interactions from genome sequences is proposed in this paper. It is based on our recent observation that protein-protein interactions can be identified by a set of “unusual” protein-profile pairs in experimentally determined protein interactions. A pair of protein-profiles is considered to be unusual if its occurrence in the given data is statistically unusual from what is expected at random. The proposed method, called PICUPP, sifts out unusual protein-profile pairs by comparing frequency distributions of their occurrences in the given data to what may result from random appearances. It is demonstrated that such unusual protein-profile pairs with statistically assessed confidences can be learned efficiently from the DIP database using a bootstrapping approach. We particularly illustrate that unusual (or, significant) protein-profile pairs can be characterized as pair wise interactions between the Pfam domains, Blocks protein families, or InterPro signatures. Such statistically significant protein-profile pairs can be used for predicting putative pairs of interacting proteins. Their prediction accuracy is around 86% and 90% when InterPro and Pfam profiles are used, respectively at 75% confidence level.

Keywords: protein-protein interaction, protein complex, interacting domains, functional genomics, proteomics

Introduction

Protein-protein interactions are fundamental to cellular processes. They are responsible for phenomena like DNA replication/transcription, regulation of metabolic pathways, immunologic recognition, signal transduction, etc. The identification of interacting proteins is therefore an important prerequisite step in understanding its physiological function.

High-throughput experimental techniques to detect and analyze protein-protein interactions are already tackling this problem systematically. The leading experimental methods are yeast two-hybrid system (Fields et al. 1989), protein arrays (Finley et al. 1994; Zhu et al. 2001), phage display (Rodi et al. 1999), and mass spectrometry (Gavin et al. 2002; Ho et al. 2002). Using these techniques, the protein interaction networks for a few cellular systems, such as yeast (Ito et al. 2000; Uetz et al. 2000) and *Helicobacter pylori* (Rain et al. 2001), were obtained.

Although, all these experiments generate rich data resources of interacting proteins, they still have a limited range of applicability because they are extremely laborious and expensive. Moreover, they are inaccurate due to many limitations intrinsic to the experimental design (see Discussion). For example, only few of the existing experimental techniques would be able to detect transient protein complexes that comprise a significant fraction of all regulatory interactions in the cell [Chen, 2003 #100]. The accuracy and coverage of these techniques have been recently compared (Lakey et al. 1998; Legrain et al. 2001) and constitute a strong argument

^a Computational Biology Group, Computer Science and Mathematics Division, Oak Ridge National Laboratory^c, P.O. Box 2008, Oak Ridge, TN 37830

^b To whom correspondence should be addressed: samatovan@ornl.gov

^c Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under Contract No. DE-AC05-00OR22725.

for the development of computational approaches to prediction of protein-protein interactions in order to enhance and supplement experimental approaches.

From a computational standpoint, the problem is how we can predict that two proteins interact from their structure or sequence information. Various computational methods using genomic context alone have recently been designed to address this problem (see reviews in (Valencia et al. 2002; Chen 2003)). They are based on gene fusion events (Enright et al. 1999; Marcotte et al. 1999), conservation of gene-order or co-occurrence of genes in potential operons (Dandekar et al. 1998; Overbeek 1999), and presence/absence of genes in different species (Pellegrini et al. 1999). All these methods attempt to identify functionally associated genes (for example, involvement in the same biochemical pathway or similar gene regulation). However, they provide only a small coverage of direct physical interactions (Huynen et al. 2000), which is more inherent to experimental approaches described above.

In parallel to genomic context based developments, a number of computational methods that attempt to “learn” from experimental data of interacting proteins have been reported in the literature (Bock et al. 2001; Sprinzak et al. 2001). Sprinzak and Margalit (Sprinzak et al. 2001) tried to learn what typifies interacting protein pairs by analyzing over-representation of sequence-signature pairs derived from available experimental data of interacting proteins. For this, they used a log-odds value that captures the ratio between the expected at random and the observed frequency of a sequence-signature pair. The pair whose log-odds value is above a certain threshold is identified as over-represented. However, there are a number of limitations with this approach. First, statistical significance of the computed log-odds values is hard to be drawn, especially for sequence-signature pairs that have high log-odds values but very low actual frequencies. Note that in their study more than a half of the sequence-signature pairs in *Saccharomyces cerevisiae* had actual frequency values of 1. Second, the reported classification of protein-protein interactions is biased toward positive interactions. The sensitivity of 94% using “leave-one-out” cross-validation was reported on a positive set (i.e., the set of experimentally determined protein interacting pairs) for a certain threshold value. But at the same time our calculations showed that about 93% of protein pairs from a random set (the set of randomly generated protein pairs) of similar size as the positive set were predicted as interacting for the same threshold value. This is largely due to the way log-odds values are updated when classifying an undetermined protein pair by treating it as an interacting pair.

Bock and Gough (Bock et al. 2001) attempted to learn correlations between biochemical patterns of sequence pairs derived from both the positive set and “negative” set (a set of putative “non-interacting” protein pairs). Because an accurate understanding of those proteins that do not interact is very limited, the authors generate a “negative” set by randomizing amino acids sequences sampled from the positive set while preserving amino acid composition as well as di- and tri-peptide ‘k-let’ frequencies (Kandel et al. 1996; Coward 1999). Shuffling biological sequences was found useful for estimating statistical significance of sequence alignments when the obtained alignments are judged relative to a background level (provided by shuffled sequences) expected by chance alone (Kandel et al. 1996). In a sense, the randomization was done against what was evaluated. However, using a set of shuffled sequence pairs as a negative set should be done with caution. Specifically, three questions need to be addressed. First, whether sequences in a negative set are part of a real protein space or part of some artificially

created space of sequences of non-existing proteins. For example, we observed that shuffling protein sequences often results in sequences with no putative conserved domains detected by the NCBI Blast searches (i.e., the E-value was greater than 2.0 on average). Second, whether random data is constructed properly. Namely, whether randomization allows estimating statistical significance of observed correlations compared to the expected random correlations learned by a classifier. Finally, whether any biological evidence exists that indicates that sequence pairs under study correspond to proteins that do not interact. Otherwise, the performance of a classifier becomes quite questionable.

Instead of adopting a traditional classification approach by fabricating negative instances whose biological meanings are in question, we use a statistical approach that sifts out a set of correlated profile pairs that account for protein-protein interactions using positive data only. Here “correlated” indicates that co-occurrence of the two given profiles accounts for an interaction. We particularly choose to use *unusual* protein profile pairs as such correlated pairs. Put in more detail, given an interacting protein pair and the two sets of associated profiles, we assess the importance of each profile combination by performing a statistical simulation. A pair of profiles is then identified as a correlated pair, if its occurrence(s) in the data is statistically unusual relative to its occurrence generated by random (independent) protein pairings. Whenever the context is clear, we will use *correlated* and *unusual* interchangeably henceforth. The proposed approach is investigated with various protein profiles: Pfam (Bateman et al. 2002), InterPro (Mulder et al. 2003) and Blocks (Henikoff et al. 2000) using the Database of Interacting Protein (DIP) (Xenarios et al. 2000; Xenarios et al. 2002) of June 16, 2002.

Materials and Methods

The identification of reliable indicators that characterize protein-protein interactions largely depends on the quality of the experimental data under consideration. Given the absence of negative protein-protein interaction data (or, pairs of proteins that are known to be non-interacting), statistical approaches that compare observed pairings to independent pairings are natural candidates to extract such indicators from positive protein-protein interaction data. However, experimental data are incomplete. This means that certain indicators may be too sparsely represented in the data to get a statistically significant assignment, even though they may be biologically important. Experimental data can also contain incorrect protein-protein interactions (false positives) due to limitations of experimental technologies used to generate this data. In this section, we describe **Protein Interaction Classification by Unlikely Profile Pair (PICUPP)** method that extracts reliable protein interaction indicators from positive and possibly incomplete protein-protein interaction data.

PICUPP seeks to identify correlated protein-profile pairs as indicators of protein-protein interactions. A profile describes a protein domain or family that may perform biologically important functions. Profile examples include Pfam, InterPro, Blocks, etc. It is generally constructed from multiple sequence alignment and is maintained in a tabular form (or, matrix of position-specific values and gap penalties) or as a hidden Markov model. A protein is typically associated with a number of profiles depending on the value of a similarity threshold. Each protein pair contributes to a table with profiles as the row and column labels. Because each protein usually has more than one profile, a pair contributes to several cells of the table, each a combination of the pair’s profiles. A set of protein pairs creates a table of profile pair counts. If

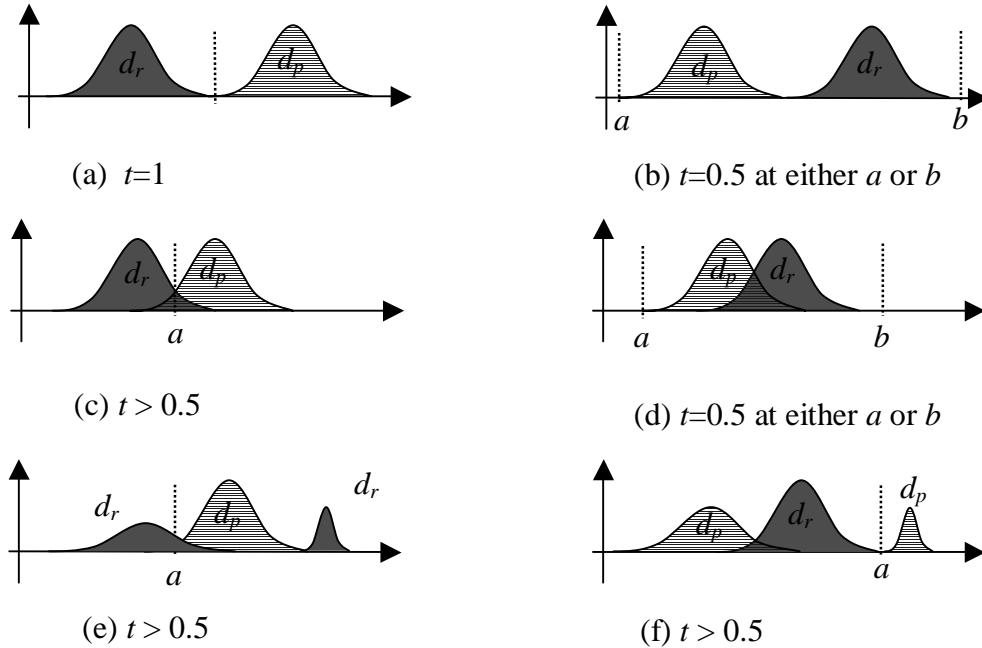


Figure 1: Cut-points that maximally separate d_p and d_r . Either a and b is a cut-point in (b) and (d).

this were true contingency table, each protein pair would contribute to exactly one cell of the table and significance analysis would follow standard contingency table theory with a closed form solution (Fienberg 1976). Because of the non-standard setting of multiple cell contributions we evaluate significance by simulation.

Sifting correlated profile pairs by comparing bootstrapped profile pair distributions

The frequency counts in a profile pair table are random variables that are generated by the available interacting protein pair data set D and by the complex process of assigning profiles. Since a meaningful evaluation of a random variable can only be made through its distribution, we bootstrap the table cell distributions by resampling the protein pair data D and computing many instances of the table. We compute two ensembles of tables: one ensemble is based on sampling the interacting protein pair data D and the other ensemble is based on sampling random pairings of the same set of unique proteins. This provides two count samples for each cell of the

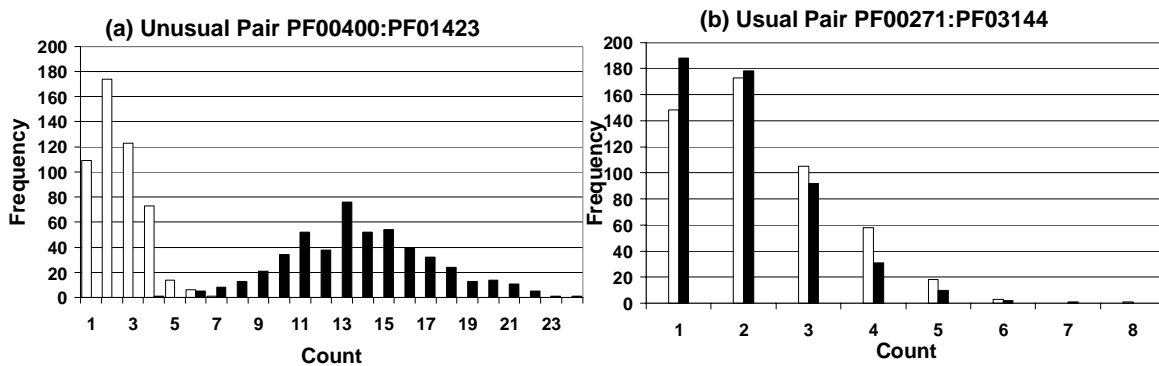


Figure 2: Example count distributions of unusual (a) and usual (b) Pfam profile pairs. In each case, white bars and black bars represent d_r and d_p distributions, respectively.

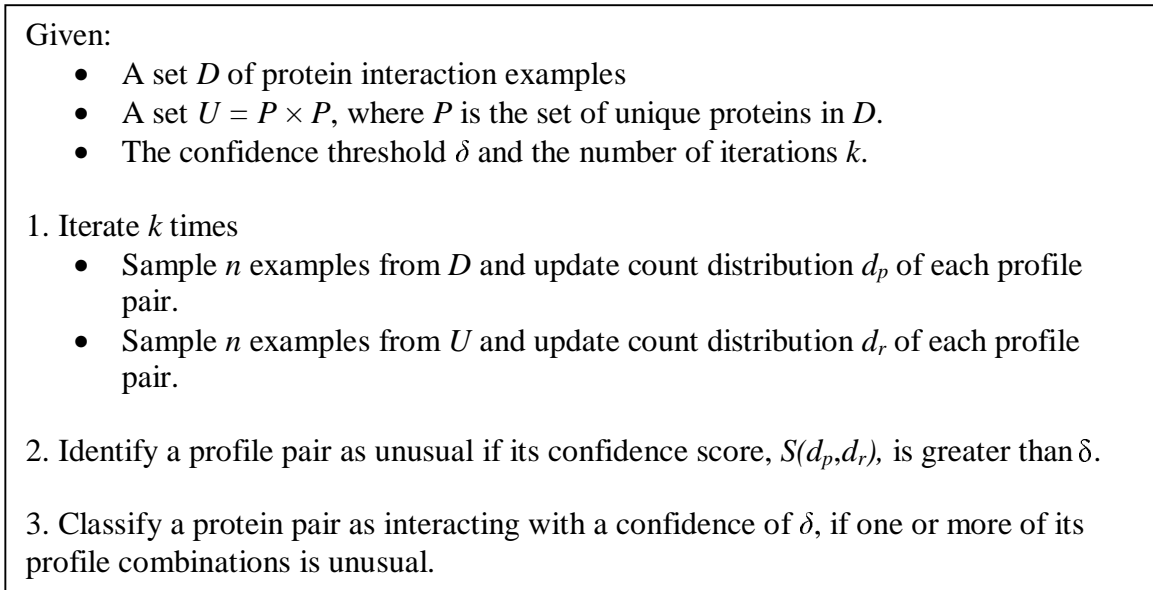


Figure 3: PICUPP: Bootstrapping approach to identify correlated profile pairs.

table: an interacting pairs sample, c_p , and a random pairs sample, c_r . We denote the corresponding empirical count distributions as d_p and d_r , respectively. The amount of overlap in d_p and d_r determines the significance of a profile pair. That is, PICUPP finds a profile pair (a cell in the table) to be correlated if its frequency among interacting protein pairs is significantly different from the frequency among random pairs. Computationally, PICUPP only keeps track of the d_p and d_r count distributions for non-zero counts.

The confidence score of a pair of profiles (s_i, s_j) is computed from its two count distributions d_p and d_r as:

$$S(d_p, d_r) = \max_a \left(\frac{1}{2} P(X_r \leq a) + \frac{1}{2} P(X_p \geq a) \right)$$

Note that $S()$ essentially computes the average proportions of d_p and d_r that lie on right and left of the given cut-point a : it assigns different values depending on how d_p and d_r are placed, even though amounts of separation are identical. $S()$ takes values between 0.5 and 1.0. For example, consider two cases when d_p and d_r are completely separated as shown in Figure 1(a) and 1(b). $S()$ assigns 1 when d_r lies in the left of d_p , whereas it assigns 0.5 for the other case. This is due to the observation that a pair that occurs more frequently in randomly coupled protein pairs cannot account for a protein-protein interaction. We consider a profile pair as correlated (unusual) if its confidence score $S()$ is greater than some threshold δ . Figure 2 shows two examples of d_r and d_p distributions for usual (Left) and unusual (Right) Pfam profile pairs. Figure 1 shows how the maximal cut-points (i.e., a) are found in several cases.

Given a set of identified unusual protein-profile pairs, PICUPP determines a possible interaction between a pair of proteins (pr_k, pr_l) , if all pair-wise combinations of their profiles include at least one such unusual pair. For example, let proteins pr_k and pr_l be associated with profiles (s_1, s_2) and (s_3, s_4) , respectively. Then, if any combination $(s_i, s_j) \in (s_1, s_2) \times (s_3, s_4)$, where \times denotes Cartesian product operator, is identified as correlated, the pair (pr_k, pr_l) is classified as interacting. The overall procedure of PICUPP is detailed in Figure 3.

Results

In this section, we discuss the performance of PICUPP in several aspects. First, its efficacy as a reliable classifier for protein-protein interactions is studied in terms of its sensitivity (i.e., the proportion of protein pairs that are identified as interacting) to both positively interacting and randomly coupled protein pairs. The aim is to estimate ratios of true-positive and false-positive cases that PICUPP produces. Since PICUPP is modeled by comparing statistics of the given protein interaction pairs with those expected at random, it is expected that PICUPP will be sensitive to the positive protein interaction pairs while being insensitive to randomly coupled protein pairs. Second, we monitor changes of sensitivity when PICUPP is trained with different numbers of bootstrapped samples. This is particularly helpful to learn how sensitivity of PICUPP at different confidence levels change with different simulation sizes, and to identify when the overall performance is stabilized. Third, the sensitivity of PICUPP to protein interactions in different genomes, which are not used during the training stage, is studied.

Comparative analysis of various profiles

We trained PICUPP with different profiles to study the suitability of several profiles as means to induce unusual profile pairs. For this experiment, Blocks, InterPro and Pfam protein profiles were individually tested using interacting protein pairs in the yeast *Saccharomyces cerevisiae* subset from DIP database. Experiments were repeated with different simulation sizes (the

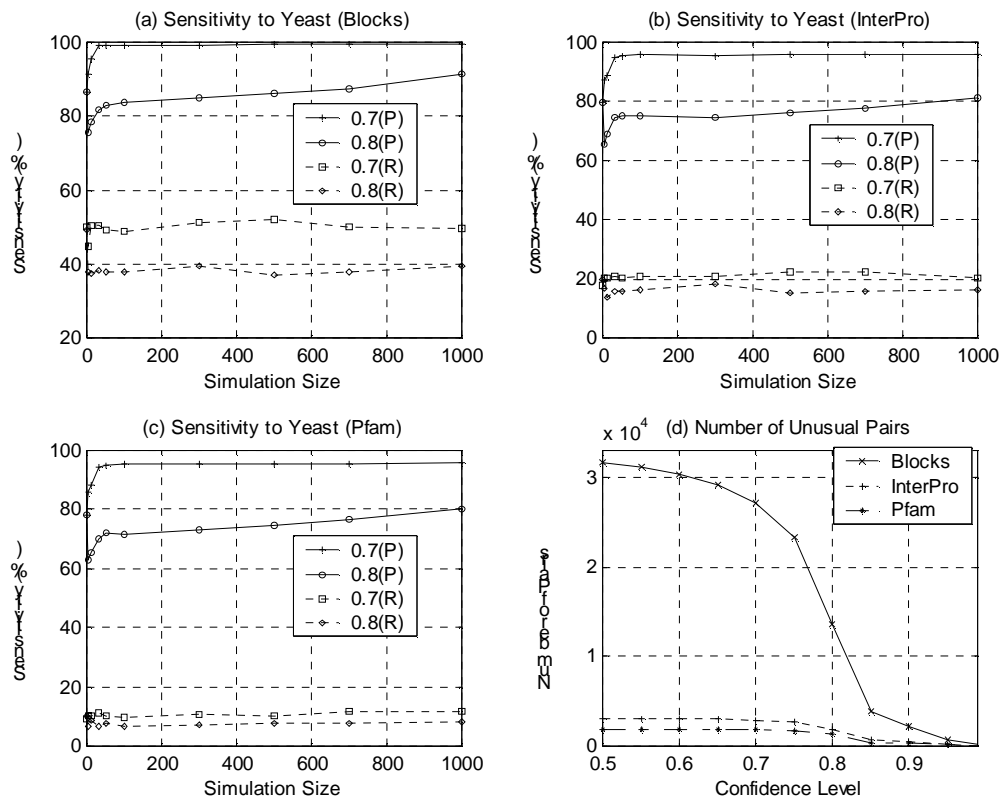


Figure 4: The sensitivity to Yeast subset (a)-(c) and model sizes (d) using various profiles. In each sub-figure (a)-(c), sensitivities at confidence levels of 0.7 and 0.8 are shown to both positive (P) and randomly coupled (R) interactions. In (d), the number of unusual profiles at different confidence levels is presented.

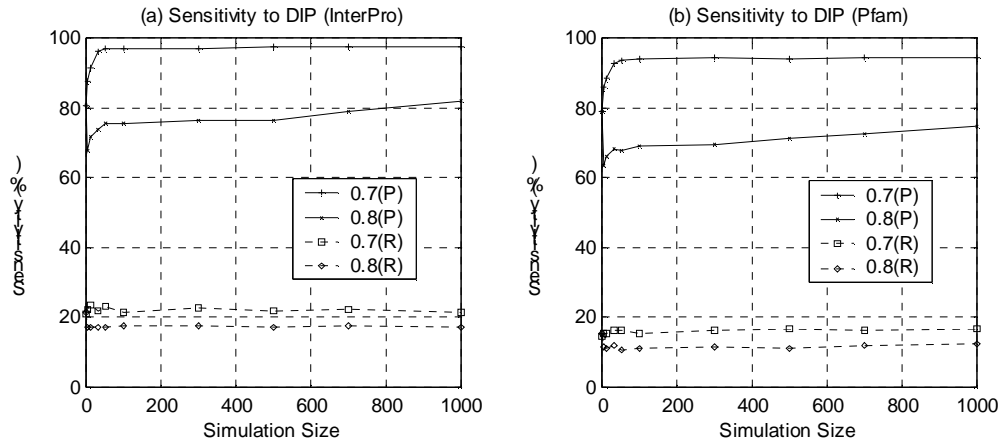


Figure 5: The sensitivity to DIP protein interactions when InterPro and Pfam profiles are used respectively for different simulation sizes. In each figure, the sensitivities at confidence levels of 0.7 and 0.8 are shown to both positive (P) and randomly coupled (R) interactions.

number of bootstrapped samples). Figure 4 shows the sensitivity to yeast subset for each profile used. In each case, sensitivity to both positive interactions and randomly coupled interactions are illustrated at the confidence level of 0.7 and 0.8. For both positive and randomly coupled interactions, the sensitivity with Blocks is higher. This may be due to the fact that Blocks profiles are selected with high E-value, which is in fact set to 1.0. Therefore it is very likely that many uncorrelated profiles are introduced, and consequently a large number of usual profile pairs are falsely identified as unusual. This is indirectly verified in Figures 4-(d) where the histograms of unusual protein pairs are plotted with respect to different confidence levels. The Figures show that the numbers of unusual pairs are roughly 10 times larger with Blocks than the others. Each histogram is a result from a simulation of size 500. In summary, all three profiles produced compatible performance.

Analysis on the entire DIP data

A protein domain often embraces multiple species. It is therefore important to investigate possible profile pairs that can be found across multiple species. DIP database includes protein interactions from 5 different organisms. To find out potentially correlated inter-species protein pairs as well as intra-species pairs, we performed a larger scale experiment using all interacting protein pairs in DIP database. From DIP database, we selected proteins that have Swiss-Prot annotations. As a result, the total of 17,000 interacting protein pairs in DIP has been reduced to 7,655 for cases with InterPro and 6,652 for cases with Pfam, respectively.

Again, experiments were repeated with different simulation sizes. Figures 5 illustrate sensitivity to positive and randomly coupled interactions at the same confidence levels as in Figure 4. In both cases with (a) InterPro and (b) Pfam profiles, PICUPP shows high sensitivity to positive interactions, whereas low to randomly coupled protein pairs. With confidence of 80%, the sensitivity of PICUPP to positive interactions is around 82% (InterPro) and 75% (Pfam) from the simulation of size 1,000. On the other hand, it is around 17% (InterPro) and 13% (Pfam) to randomly coupled protein pairs. The result illustrates that PICUPP effectively identifies inter-species protein-protein interactions that are reflected in DIP database.

Cross Coverage/Validation

We also measured the performance of PICUPP when it is applied to a list of interacting protein pairs that is left out during the training stage. For this, we excluded all interactions of Yeast from DIP database and trained PICUPP. Then the sensitivity (or, accuracy) of PICUPP to the interacting protein pairs in Yeast was measured. Then the same experiment was repeated with interacting protein pairs in E-coli being left out.

Figure 6-(a) and (b) shows the sensitivity (coverage) of PICUPP to Yeast protein pairs in cases of InterPro and Pfam, respectively. Similarly Figure 6-(c) and (d) show the coverage of E-coli protein pairs. As clearly illustrated in the Figures, PICUPP was able to identify a big portion of interacting protein pairs in Yeast. However, it failed to classify majority of interacting protein pairs in E-coli, which indicates that unusual profile pairs that typify interactions in E-coli were not found by the process and may be different from that observed in the remaining data. This can be understood by the fact that DIP database is largely biased toward Eukaryotes. Although the bacterium *Escherichia coli* (E-coli) is the most dominant non-eukaryote proteome in DIP database, it only accounts for 1.3% of the proteins therein (Bock et al. 2001).

High confident interaction-related Pfam pairs

Approximately 108 pairs of Pfam domains are identified unusual at a confidence level of 98%. A case-by-case investigation discovered many well-established Pfam-Pfam associations among these interactions. The SH3 domain is perhaps the best-characterized member of protein-

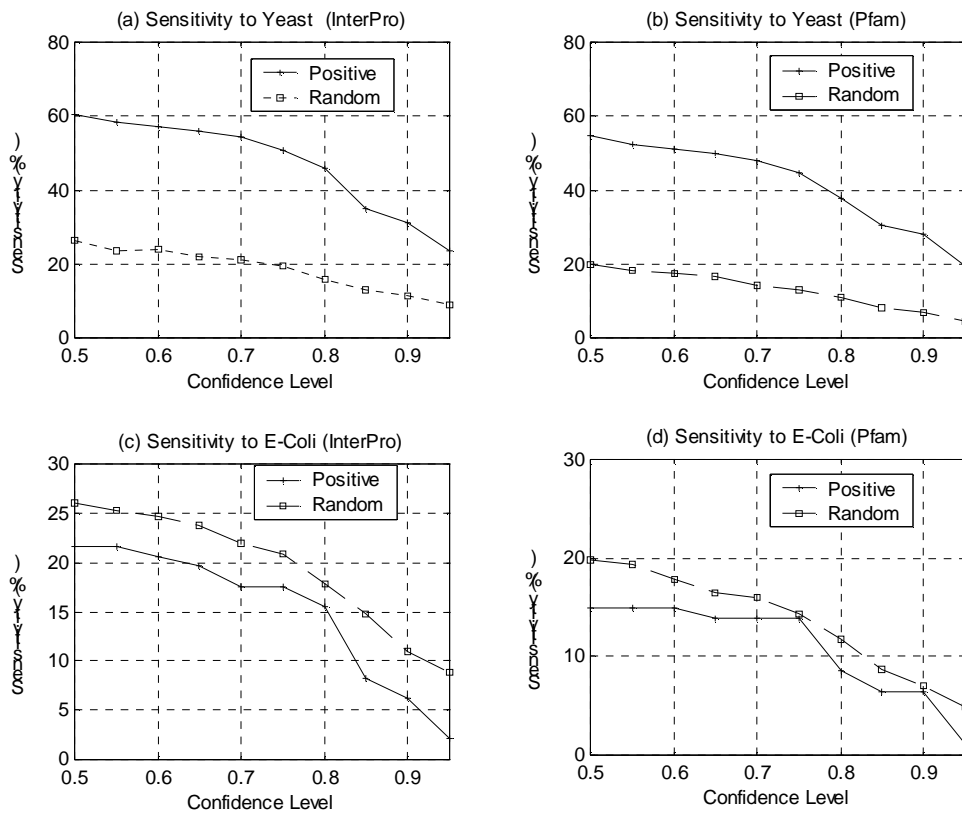


Figure 6: Cross coverage of PICUPP. Sub-figures (a),(b) and (c),(d) illustrate the sensitivities to yeast subset and *E-coli*, when InterPro and Pfam profiles are used, respectively. In each case, solid line stands for sensitivity to positive interactions with respect to different threshold level, whereas dashed line to randomly coupled interactions.

interaction modules. By binding with moderate affinity and selectivity to proline-rich ligands, the domain plays a vital role in a wide variety of biological processes ranging from regulation of enzymes, increasing the local concentration or altering the subcellular localization of components of signaling pathways, and mediating the assembly of large multiprotein complexes (Pawson et al. 1993). The SH3 is found to pair in unusually high frequency with Actin in our analysis. This domain is often closely associated with Actin (in cytoskeletal proteins, such as fodrin and yeast actin binding protein ABP-1 (Drubin et al. 1990)).

G-protein beta WD-40 repeat (G- β), another well-known interaction module, is one of the three subunits (α , β , and gamma) of the guanine nucleotide-binding proteins (G proteins) which act as intermediaries in the transduction of signals generated by transmembrane receptors. The α subunit binds to and hydrolyzes GTP; the β and gamma subunits seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition. We found that this domain is highly coupled with Small nuclear ribonucleoprotein (sm protein) in our analysis. This finding is consistent with previous research results (Achsel et al. 1998). Both the sm proteins and G- β possibly mediate regulated protein-protein interactions essential for the functions of small nuclear ribonucleoproteins (snRNPs). Additional well-known Pfam pairs occurred in our prediction include actin and Cofilin/tropomyosin-type actin-binding protein (Nishida et al. 1984), Protein kinase domain and Fibroblast growth factor (Taniguchi et al. 2003), and EF hand and Myosin head (motor domain) (Messer et al. 1991).

Discussion and Conclusion

Considering the enormous combinatorial space of protein-protein interactions, a reliable high-throughput method is desired not only for predicting a possible interaction, but also for identifying characteristic protein domains responsible for putative protein-protein interactions. A statistical approach to identify correlated protein-profile pairs that account for protein interactions is presented with experimental validation. We demonstrate that the proposed approach, PICUPP, effectively maximizes statistical confidences given to correlated protein-profile pairs by applying a bootstrapping approach to an incomplete data. We show that a set of unusual protein-profile pairs inferred from experimentally determined protein interactions can indeed epitomize putative protein-protein interactions. Such unusual protein-profile pairs reveal interacting domains and uncover relationships between highly correlated/uncorrelated domains for protein interactions.

Since we have not compared the coverage of PICUPP with that of other methods such as the domain fusion approach, it is still early to assert whether the coverage of PICUPP is complementary or subset of others. While the fusion method captures similar domains, our results illustrate many counter-examples. For this reason, we conjecture that our approach produces complementary coverage.

Despite the promising results obtained in the different tests carried out, a number of problems are still present in the current approach. The current performance of our approach depends on the quality of the experimental data of interacting proteins that are used to train the system. Unfortunately genome-scale experimental methods, such as protein arrays and two-hybrid system, have many limitations intrinsic to the experimental design. For instance, the protein should preserve a correct fold while attached to the chip surface (or linked to the hybrid domain).

Otherwise, non-native interactions could be captured by the method. Another limitation, and even more restrictive, is the binary nature of some of those experimental approaches. But many of the cellular machines are multi-protein complexes. Moreover, transient (short-living) protein complexes probably comprising a significant fraction of all regulatory interactions in the cell may need additional stabilization for detection by these experimental methods. Thus, expanding the training set with various types of annotated protein interactions will potentially address this problem.

The DIP-based training set of interacting proteins is heavily biased towards the Eukaryotes. Note that the yeast *Saccharomyces cerevisiae* accounts for 64% of the interactions. The bacterium *Escherichia coli* constitutes only 1.3% of the proteins found in the database. As a result, a factor of three differences in the predictions (55% vs. 17%) was observed for the yeast and the *E.coli* genomes (Figure 6). It is to be expected that the increase of data of interacting proteins for the other kingdoms of life would alleviate this problem, allowing us to increase the number of predictions.

Given the data of protein interactions, the performance of PICUPP may vary depending on the resolution and specificity of the underlying protein profiles that are generated with certain similarity threshold value (e.g., E-value). With lower threshold (less similar) values, proteins will be assigned more profiles, which possibly include non-related ones. This will result in increased false-positive rate. On the other hand, with higher threshold values, a more compact set of correlated pairs will be identified, which, in turn, will result in decreased sensitivity to positive protein interactions. Therefore, the similarity threshold value should be chosen with care.

The type of the underlying protein profiles (e.g., Pfam domains, Blocks families, InterPro signatures, ProSite regular expressions) is another factor whose influence on the quality of our approach still remains to be evaluated. For example, no statistically significant ProSite expressions in terms of log-odds values were observed. Unfortunately, an accurate understanding of interactions between protein profiles and how these interactions affect interactions between proteins is very limited. A systematic study that led to the classification of interacting proteins in terms of the repertoires of interactions of structural domains was determined for the PDB and the yeast genome (Park et al. 2001; Teichmann et al. 2001). A possible extension may involve moving away from somewhat “ad hoc” utilization of protein profiles to more systematic approaches leading to a comprehensive understanding of relationships between interacting protein profiles and interacting proteins.

It is very difficult to assess the accuracy and coverage of any of the protein interaction prediction techniques in the absence of large and representative enough collection of well-annotated protein interactions. Current efforts to develop databases of protein interactions (Xenarios et al. 2000; Bader et al. 2001; Zanzoni et al. 2002) and to establish standards for the exchange of information between these databases will, certainly, be critical in the evaluation of prediction methods.

Acknowledgements

This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org). The work of G.O. was sponsored by the

Laboratory Directed Research and Development Program of Oak Ridge National Laboratory. This research used resources of the Center for Computational Sciences at Oak Ridge National Laboratory.

References:

- Achsel, T., K. Ahrens, et al. (1998). "The human U5-220kD protein (hPrp8) forms a stable RNA-free complex with several US-specific proteins, including an RNA unwindase, a homologue of ribosomal elongation factor EF-2, and a novel WD-40 protein." Molecular and Cellular Biology **18**(11): 6756-6766.
- Bader, G. D., I. Donaldson, et al. (2001). "BIND - The Biomolecular Interaction Network Database." Nucleic Acids Research **29**(1): 242-245.
- Bateman, A., E. Birney, et al. (2002). "The Pfam protein families database." Nucleic Acids Res **30**(1): 276-80.
- Bock, J. R. and D. A. Gough (2001). "Predicting protein--protein interactions from primary structure." Bioinformatics **17**(5): 455-60.
- Chen, Y. a. X. D. (2003). "Computational Analyses of High-Throughput Protein-Protein Interaction Data." Current Protein and Peptide Science **4**.
- Coward, E. (1999). "Shufflers: shuffling sequences while conserving the k-let counts." Bioinformatics **15**(12): 1058-1059.
- Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends in Biochemical Sciences **23**(9): 324-328.
- Drubin, D. G., J. Mulholland, et al. (1990). "Homology of a Yeast Actin-Binding Protein to Signal Transduction Proteins and Myosin-I." Nature **343**(6255): 288-290.
- Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.
- Fields, S. and O. K. Song (1989). "A Novel Genetic System to Detect Protein Protein Interactions." Nature **340**(6230): 245-246.
- Fienberg, S. E. (1976). "Analysis of Cross-Classified Categorical Data." Notices of the American Mathematical Society **23**(6): A619-A619.
- Finley, R. L. and R. Brent (1994). "Interaction Mating Reveals Binary and Ternary Connections between Drosophila Cell-Cycle Regulators." Proceedings of the National Academy of Sciences of the United States of America **91**(26): 12980-12984.
- Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-147.
- Henikoff, J. G., S. Pietrokovski, et al. (2000). "Blocks-based methods for detecting protein homology." Electrophoresis **21**(9): 1700-6. [pii].
- Ho, Y., A. Gruhler, et al. (2002). "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry." Nature **415**(6868): 180-183.
- Huynen, M., B. Snel, et al. (2000). "Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences." Genome Research **10**(8): 1204-1210.
- Ito, T., K. Tashiro, et al. (2000). "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins." Proceedings of the National Academy of Sciences of the United States of America **97**(3): 1143-1147.
- Kandel, D., Y. Matias, et al. (1996). "Shuffling biological sequences." Discrete Applied Mathematics **71**(1-3): 171-185.
- Lakey, J. H. and E. M. Raggett (1998). "Measuring protein-protein interactions." Current Opinion in Structural Biology **8**(1): 119-123.

- Legrain, P., J. Wojcik, et al. (2001). "Protein-protein interaction maps: a lead towards cellular functions." Trends in Genetics **17**(6): 346-352.
- Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-3.
- Messer, N. and J. Kendrickjones (1991). "Chimeric Myosin Regulatory Light-Chains - Subdomain Switching Experiments to Analyze the Function of the N-Terminal Ef Hand." Journal of Molecular Biology **218**(4): 825-835.
- Mulder, N. J., R. Apweiler, et al. (2003). "The InterPro Database, 2003 brings increased coverage and new features." Nucleic Acids Research **31**(1): 315-318.
- Nishida, E., S. Maekawa, et al. (1984). "Cofilin, a Protein in Porcine Brain That Binds to Actin-Filaments and Inhibits Their Interactions with Myosin and Tropomyosin." Biochemistry **23**(22): 5307-5313.
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N. (1999). "Use of contiguity on the chromosome to predict functional coupling." In Silico Biol. **1**: 93-108.
- Park, J., M. Lappe, et al. (2001). "Mapping protein family interactions: Intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast." Journal of Molecular Biology **307**(3): 929-938.
- Pawson, T. and J. Schlessinger (1993). "Sh2 and Sh3 Domains." Current Biology **3**(7): 434-442.
- Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A **96**(8): 4285-8.
- Rain, J. C., L. Selig, et al. (2001). "The protein-protein interaction map of Helicobacter pylori." Nature **409**(6817): 211-215.
- Rodi, D. J. and L. Makowski (1999). "Phage-display technology - finding a needle in a vast molecular haystack." Current Opinion in Biotechnology **10**(1): 87-93.
- Sprinzak, E. and H. Margalit (2001). "Correlated sequence-signatures as markers of protein-protein interaction." J Mol Biol **311**(4): 681-92.
- Taniguchi, F., T. Harada, et al. (2003). "Activation of mitogen-activated protein kinase pathway by keratinocyte growth factor or fibroblast growth factor-10 promotes cell proliferation in human endometrial carcinoma cells." Journal of Clinical Endocrinology and Metabolism **88**(2): 773-780.
- Teichmann, S. A., A. G. Murzin, et al. (2001). "Determination of protein function, evolution and interactions by structural genomics." Current Opinion in Structural Biology **11**(3): 354-363.
- Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-627.
- Valencia, A. and F. Pazos (2002). "Computational methods for the prediction of protein interactions." Current Opinion in Structural Biology **12**(3): 368-373.
- Xenarios, I., D. W. Rice, et al. (2000). "DIP: the Database of Interacting Proteins." Nucleic Acids Research **28**(1): 289-291.
- Xenarios, I., L. Salwinski, et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucleic Acids Research **30**(1): 303-305.
- Zanzoni, A., L. Montecchi-Palazzi, et al. (2002). "MINT: a Molecular INTERaction database." Febs Letters **513**(1): 135-140.
- Zhu, H., M. Bilgin, et al. (2001). "Global analysis of protein activities using proteome chips." Science **293**(5537): 2101-2105.