

Inference of Protein-Protein Interactions by Unlikely Profile Pair

Byung-Hoon Park, George Ostrouchov, Gong-Xin Yu,
Al Geist, Andrey Gorin, and Nagiza F. Samatova

*Computational Biology Group, Computer Science and Mathematics Division,
Oak Ridge National Laboratory
{parkbh, samatovan}@ornl.gov*

Abstract

We note that a set of statistically “unusual” protein-profile pairs in experimentally determined database of protein-protein interactions can typify protein-protein interactions, and propose a novel method called PICUPP that sifts such protein-profile pairs using a statistical simulation. It is demonstrated that unusual Pfam and InterPro profile pairs can be extracted from the DIP database using a bootstrapping approach. We particularly illustrate that such protein-profile pairs can be used for predicting putative pairs of interacting proteins. Their prediction accuracies are around 86% and 90% when InterPro and Pfam profiles are used, respectively at 75% confidence level.

1. 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. From a computational standpoint, the problem is how we can predict that two proteins interact from their structures or sequence information. Various computational methods using genomic context alone have recently been designed to address this problem (see reviews in [1, 2]). They are based on gene fusion events [3, 4], conservation of gene-order or co-occurrence of genes in potential operons [5, 6], and presence/absence of genes in different species [7]. 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 [8], which is more inherent to experimental approaches.

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 [9, 10]. Sprinzak and Margalit [10] tried to learn what typifies interacting protein pairs by analyzing over-representation of

sequence-signature pairs derived from available experimental data of interacting proteins. Bock and Gough [9] attempted to learn correlations between biochemical patterns of sequence pairs derived from experimentally verified positive set and artificially manufactured “negative” set (a set of putative “non-interacting” protein pairs).

Protein Interaction Classification by Unlikely Profile Pair (PICUPP) - our proposed method - extracts protein interaction indicators from positive protein-protein interaction data (no negative instances). It particularly 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. Here “correlated” indicates that co-occurrence of the two given profiles accounts for an interaction. We particularly choose to use *unusual* protein profile pairs, which are derived from statistical simulation using interacting protein pairs, as such correlated pairs. 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 domains [11], InterPro signatures [12] and Blocks motifs [13] using the Database of Interacting Protein (DIP) [14] of June 16, 2002.

2. Methods

A protein sequence is typically associated with multiple profiles. Thus each interacting protein pair may contribute to a number of profile pairs. A set of protein pairs creates a table, where each cell denoted by a row and column contains the frequency count of a profile pair. Note that it is different from an ordinary contingency table, because each protein pair would not contribute to exactly one cell of the table. Therefore significance analysis would not follow standard contingency table theory with a closed form solution [15]. Because of the non-standard setting of multiple cell contributions we evaluate significance by simulation.

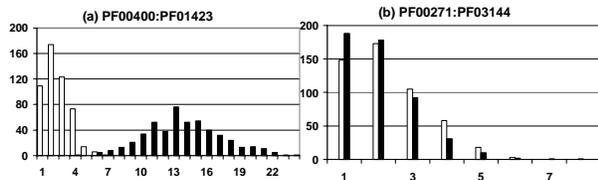


Figure 1: Example count distributions of unusual (a) and usual (b) Pfam profile pairs. In each case, white bars and black bars represent d_r and d_p distributions, respectively. X and Y axes represent count and its frequency.

The frequency counts in a profile pair table are random variables that are generated by the simulation. For each protein pair (pr_k, pr_l) in D , a set of profile pairs is identified by all pair-wise combinations of profiles between pr_k and pr_l . For example, let proteins pr_k and pr_l be associated with profiles (s_1, s_2) and (s_3, s_4) , respectively. Then, 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 a profile pair. 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 tables: one from bootstrapped samples of the same protein interaction data and the other from bootstrapped samples of the random protein pairings. This provides two count samples for each cell of the table: an interacting pairs sample, c_p , and a random pairs sample, c_r . We denote the corresponding empirical count distribution as d_p and d_r , respectively. The amount of overlap between d_p and d_r determines the usualness of the pair compared to what is expected at random. That is, PICUPP finds a profile pair (a cell in the table) to be unusual (or, correlated) if its frequency distribution constructed from interacting protein pair samples is significantly different from what

is constructed from random protein pairings. Computationally, PICUPP only keeps track of the d_p and d_r distributions for non-zero counts. Technically, the confidence score, or the degree of correlation of a profile pair (s_i, s_j) is computed from its two count distributions d_p and d_r , as:

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

where X_r and X_p are random variables of d_p and d_r , respectively. Given a set of identified unusual protein-profile pairs, PICUPP determines a possible interaction between a pair of proteins (pr_k, pr_l) , if any of its profile pair is unusual. An example of unusual and usual profile pair distribution is shown in Figure 1.

3. Results

PICUPP is modeled by comparing statistics of the given protein interaction pairs with those expected by random pairing of proteins. Therefore, it is expected to be sensitive to the positive protein interactions while being insensitive to those random protein couples. In this section, we report the sensitivity of PICUPP when InterPro and Pfam profiles are used over protein interactions in DIP database. We also discuss how different sizes of bootstrap samples and different confidence levels affect the overall performance.

DIP database contains a rich set of protein interactions from several species. From approximately 17,000 total protein interactions, we selected proteins that have Swiss-Prot annotations. As a result, the 7,655 and 6,652 protein interactions are left available for the experiment for cases with InterPro and Pfam, respectively. Figure 2 illustrates sensitivity to positive and randomly coupled interactions at the confidence levels of 0.7 and 0.8 with the different bootstrapped sample sizes. 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 protein-protein interactions.

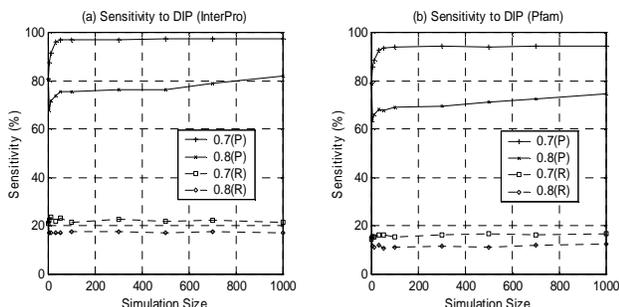


Figure 2: The sensitivity to protein interactions in DIP when PICUPP is trained with InterPro and Pfam respectively. In each figure, the sensitivities at confidence levels of 0.7 and 0.8 are shown to both positive (P) and randomly coupled (R) interactions.

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. Likewise, the sensitivity to protein pairs in E-coli was

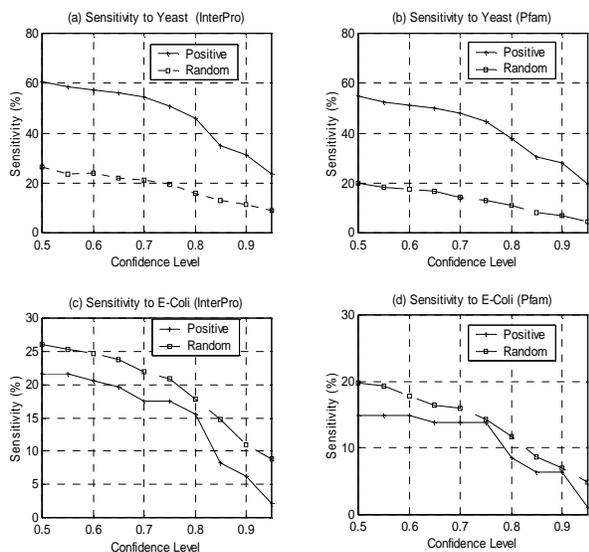


Figure 3: Cross coverage of PICUPP. Sub-figures (a),(b) and (c),(d) illustrate the sensitivities to yeast subset and *E-coli* when InterPro and Pfam are used. In each case, the solid line illustrates sensitivity to positive interactions at different threshold level, whereas dashed line illustrates sensitivity to randomly coupled interactions.

measured. Figure 3-(a) and (b) shows the sensitivity (coverage) of PICUPP to Yeast protein pairs in case of InterPro and Pfam, respectively. Similarly Figures 3-(c) and (d) show the coverage of E-coli protein pairs. As clearly illustrated in the Figures, PICUPP was able to differentiate interacting protein pairs from randomly coupled pairs in Yeast around 57% at confidence level of 70%. However, it is around 20% for interacting pairs in E-coli at the same confidence level, which indicate the unusual profile pairs that typify interaction in E-coli were not found by the process and may be different from what are 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 [9].

4. Discussion

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-interaction modules. It plays a vital role in a wide variety of biological processes. It increases the local concentration or altering the subcellular

localization of components of signaling pathways, and mediates the assembly of large multiprotein complexes [16]. The SH3 is found to be correlated with Actin in our analysis. In fact it has been verified that this domain is often closely associated with Actin (in cytoskeletal proteins, such as fodrin and yeast actincytoskeletal proteins, such as fodrin and yeast actin binding protein ABP-1 [17]).

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 [18]. 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 found in our prediction include actin and Cofilin/tropomyosin-type actin-binding protein [19], Protein kinase domain and Fibroblast growth factor [20], and EF hand and Myosin head (motor domain) [21].

5. Conclusion

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.

PICUPP needs to be further refined in several ways. First, its performance highly depends on the quality of the experimental data. Unfortunately genome-scale experimental methods, such as protein arrays and two-hybrid system, have many limitations intrinsic to the experimental design. Another limitation, and even more restrictive, is the binary nature of some of those experimental approaches, which potentially excludes many of the cellular machines that 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. Second, an accurate understanding of interactions between protein profiles and how these interactions affect interactions between proteins is very limited. A possible solution 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.

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.

6. References

- [1] Y. a. X. D. Chen, "Computational Analyses of High-Throughput Protein-Protein Interaction Data," *Current Protein and Peptide Science*, vol. 4, 2003.
- [2] A. Valencia and F. Pazos, "Computational methods for the prediction of protein interactions," *Current Opinion in Structural Biology*, vol. 12, pp. 368-373, 2002.
- [3] A. J. Enright, et al., "Protein interaction maps for complete genomes based on gene fusion events," *Nature*, vol. 402, pp. 86-90, 1999.
- [4] E. M. Marcotte, et al., "Detecting protein function and protein-protein interactions from genome sequences," *Science*, vol. 285, pp. 751-3., 1999.
- [5] R. Overbeek, Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., "Use of contiguity on the chromosome to predict functional coupling," *In Silico Biol.*, vol. 1, pp. 93-108, 1999.
- [6] T. Dandekar, et al., "Conservation of gene order: a fingerprint of proteins that physically interact," *Trends in Biochemical Sciences*, vol. 23, pp. 324-328, 1998
- [7] M. Pellegrini, et al., "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proc Natl Acad Sci U S A*, vol. 96, pp. 4285-8., 1999.
- [8] M. Huynen, et al., "Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences," *Genome Research*, vol. 10, pp. 1204-1210, 2000.
- [9] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, pp. 455-60., 2001.
- [10] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction," *J Mol Biol*, vol. 311, pp. 681-92., 2001.
- [11] A. Bateman, et al., "The Pfam protein families database," *Nucleic Acids Res*, vol. 30, pp. 276-80., 2002.
- [12] N. J. Mulder, et al., "The InterPro Database, 2003 brings increased coverage and new features," *Nucleic Acids Research*, vol. 31, pp. 315-318, 2003.
- [13] J. G. Henikoff, et al., "Blocks-based methods for detecting protein homology," *Electrophoresis*, vol. 21, pp. 1700-6. [pii], 2000.
- [14] I. Xenarios, et al., "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, pp. 303-305, 2002.
- [15] S. E. Fienberg, "Analysis of Cross-Classified Categorical Data," *Notices of the American Mathematical Society*, vol. 23, pp. A619-A619, 1976.
- [16] T. Pawson and J. Schlessinger, "Sh2 and Sh3 Domains," *Current Biology*, vol. 3, pp. 434-442, 1993.
- [17] D. G. Drubin, et al., "Homology of a Yeast Actin-Binding Protein to Signal Transduction Proteins and Myosin-I," *Nature*, vol. 343, pp. 288-290, 1990.
- [18] T. Achsel, et al., "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*, vol. 18, pp. 6756-6766, 1998.
- [19] E. Nishida, et al., "Cofilin, a Protein in Porcine Brain That Binds to Actin- Filaments and Inhibits Their Interactions with Myosin and Tropomyosin," *Biochemistry*, vol. 23, pp. 5307-5313, 1984.
- [20] F. Taniguchi, et al., "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*, vol. 88, pp. 773-780, 2003.
- [21] N. Messer and J. Kendrickjones, "Chimeric Myosin Regulatory Light-Chains - Subdomain Switching Experiments to Analyze the Function of the N-Terminal Ef Hand," *Journal of Molecular Biology*, vol. 218, pp. 825-835, 1991.

