Pattern Discovery in Protein Networks Reveals High-Confidence Predictions of Novel Interactions

Hazem Radwan Ahmed and Janice I. Glasgow

School of Computing, Queen's University, 25 Union Street, Kingston, Ontario K7L 3N6, Canada
{hazem, janice}@cs.queensu.ca

Abstract

Pattern discovery in protein interaction networks can reveal crucial biological knowledge on the inner workings of cellular machinery. Although far from complete, extracting meaningful patterns from proteomic networks is a non-trivial task due to their size-complexity. This paper proposes a computational framework to efficiently discover topologically-similar patterns from large proteomic networks using Particle Swarm Optimization (PSO). PSO is a robust and low-cost optimization technique that demonstrated to work effectively on the complex, mostly sparse proteomic networks. The resulting topologically-similar patterns of close proximity are utilized to systematically predict new high-confidence protein-protein interactions (PPIs). The proposed PSO-based PPI prediction method (3PI) managed to predict high-confidence PPIs, validated by more than one computational/experimental source, through a proposed PPI knowledge transfer process between topologically-similar interaction patterns of close proximity. In three case studies, over 50% of the predicted interactions for EFGR, ERBB2, ERBB3, GRB2 and UBC are overlapped with publically available interaction databases, ~80% of the predictions are found among the Top 1% results of another PPI prediction method and their genes are significantly co-expressed across different tissues. Moreover, the only single prediction example that did not overlap with any of our validation sources was recently experimentally supported by two PubMed publications.

Introduction

With a growing number of characterized protein sequences and a widening gap between known sequences and their structures and functions, computational prediction techniques for protein structure, function and interactions have become increasingly valuable in the past decades. Many proteins perform their functions only when they interact with a number of other partner proteins. Protein-Protein Interactions (PPIs) are, therefore, important in understanding almost all biological processes taking place in the cell. The study of PPIs can not only help predict the function of unknown proteins, but it can also help characterize essential pathways and cellular processes.

Unfortunately we do not have a complete and accurate picture of all PPIs within cells. It is estimated that our current knowledge of human PPIs could have as high as 64% false positives (noisy PPIs) and between 43% to 71% false negatives (missing PPIs) (Cannistraci, Alanis-Lobato, and Ravasi 2013). Our current knowledge of PPIs is mainly derived from experimental determination techniques or computational prediction methods. Examples of common experimental techniques for determining PPIs include Yeast two-hybrid (Y2H) (Suter, Kittanakom, and Stagljar 2008), Mass Spectrometry (MS) with Tandem Affinity Purification (TAP) (Aebersold and Mann 2003) and Protein Microarrays (MacBeath 2002). While these techniques offer good insights about large numbers of PPIs, they are expensive, lab-intensive and often include high false-positive and false-negative rates (Jurisica and Wigle 2010). Thus, machine learning-based prediction methods have been widely tried as less-expensive alternatives to expand, validate or denoise the current knowledge of PPIs (Browne et al. 2010).

Pattern discovery in PPI networks has been successfully applied to several useful applications, including projecting functional annotation (Sharan, Ulitsky, and Shamir 2007), identifying protein complexes (Li et al. 2012) and conserved functional modules across species (Dolinski and Botstein 2007). The discovered topologically-similar patterns from protein interaction networks have already shown to overlap in biologically-relevant functions/processes and have been, for example, used to effectively construct phylogeny (Kuchaiev et al. 2010). In this paper, we suggest using such patterns in a novel way to predict high-confidence PPIs. In particular, we propose a swarm intelligence-based computational framework for pattern discovery in proteomic networks, and use the resulting topologically-similar patterns of close proximity to systematically predict new high-confidence PPIs (validated by more than one computational/ experimental source). While a number of studies (Kotlyar and Jurisica 2006; Scott and Barton 2007) employed network topological features for PPI prediction (such as the number of shared neighbors and node degree), or some topologically-defined classes for the interaction behaviour of the immediate neighbour (Saito, Suzuki, and Hayashizaki 2003), no study in the literature, to the best of our knowledge, used topologically-similar interaction patterns and protein domains to predict novel PPIs. Protein domains are evolutionarily-conserved units within protein sequences that can fold and function independently. Protein sets in topologically-similar patterns of close proximity are repeatedly found to be statistically over-represented in a number of overlapping domain/motif interactions. Our method performs a systematic knowledge transfer of PPIs between pairs of protein sets that are significantly enriched (P-value < 0.001) in overlapping domains. These lists are also repeatedly found to be statistically over-represented in a number of overlapping molecular functions, pathways, biological processes and/or tissue expression (Milenkoviæ and Pržulj 2008).

The state-of-the-art computational methods for PPI prediction are mainly based on Artificial Neural Networks (Fariselli et al. 2002), Random Forests (Chen and Liu 2005), Association Mining (Kotlyar and Jurisica 2006), Bayesian Classifiers (Scott and Barton 2007), Support Vector Machines (Shen et al. 2007) or Genetic Algorithms (GAs) (Dimitrakopoulos et al. 2012). To the best of our knowledge, this is the first study to apply Particle Swarm Optimization (PSO) to the PPI prediction problem. PSO is known to be computationally less expensive and more robust than GAs (Hassan et al. 2005). Moreover, unlike neural networks-based approaches, the proposed PSO-based approach does not require a comprehensive training set of positive and negative interaction examples to learn patterns, nor does it require the heavy-handed feature construction that data mining-based approaches need.

Background

Particle Swarm Optimization (PSO)

The emergence of flocking and schooling in groups of interacting agents (such as birds, fish, penguins, etc.) have long intrigued a wide range of scientists from diverse disciplines including animal behavior, physics, social psychology, social science, and computer science for many decades. Bird flocking can be defined as the collective motion behavior of a large number of interacting birds with a common group objective. The local interactions between birds (particles) usually emerge the shared motion direction of the swarm. Such interactions are based on the "nearest neighbor principle," where birds follow certain flocking rules to adjust their motion (i.e., position and velocity) based only on their nearest neighbors, without any central coordination or one dedicated leader.

The standard PSO uses a number of flying particles, represented as a set of points in an n-dimensional solution space with dynamically-changing velocities according to their historically best performance and the swarm-wide best performance. Each particle stores its own experience in a private local memory, whereas the social swarm-wide experience is stored in a global public memory accessible to all particles. The experience-sharing behavior is what gradually guides the swarm motion towards the most promising areas detected so far in the search space. Therefore, particles iteratively evaluate their fitness based on their current position and compare it to their historically best fitness and the global best fitness in the swarm. Then, each particle updates its experience (if the current position is better than its historically best one), and adjusts its velocity to imitate the swarm's global best particle by moving closer towards it. Before the end of each iteration in PSO, the swarm's social experience and particles' private experiences are updated if the most recent change of particle positions happened to be better than the current position stored in the global and local memories. In this study, as discussed later in the methodology section, we used a multi-start variant of PSO to overcome the well-known premature convergence issue of the standard PSO, in which particles may get trapped in sub-optimal solutions in the early search stage. Interested readers may refer to this review paper (del Valle et al. 2008) for a detailed discussion of PSO concepts, variants and equations.

Protein Interaction Prediction Methods

Gene Co-Expression

It has been shown that the corresponding genes of many interacting protein pairs are co-expressed (Von Mering et al. 2002). Thus, gene co-expression has long been used to predict or verify PPIs (Kemmeren et al. 2002). The main premise of predicting or verifying PPIs from gene-co-expression is that significantly co-expressed genes (or, gene pairs with correlated expression profiles across different conditions/samples) are more likely to encode interacting proteins (Ge et al. 2001).

However, gene co-expression alone is not the best predictor of protein interactions, since it may lead to some false positives or false negatives. For instance, the corresponding gene pairs of transient PPIs are often not highly co-expressed. Thus, the use of gene expression data would potentially lead to false negatives for transient PPIs. Because of such prediction limitations of gene co-expression, it has often been coupled with other types of interaction evidence such as protein interaction domains, network topology data, etc.

PPI Network Topology

Network topology generally refers to the relative connectivity of its nodes. The topological structures of biological networks have been widely studied (Qi 2008), because major cellular functions and processes could be understood by analyzing the complex interaction patterns in PPI networks, as well as the relative positions of proteins within the PPI networks may indicate their functional importance (Qi 2008).

Learning from Information Integration

In practice, computational prediction methods of PPIs usually integrate more than one interaction evidences for making PPI predictions (Rhodes et al. 2005; Scott and Barton 2007), based on machine learning and statistical approaches. So the PPI prediction is typically inferred through supervised machine learning classifiers that learn from information integration of multiple interaction evidences, such as the *FpClass* algorithm.

The *FpClass* algorithm is a data mining-based approach that estimates interaction probabilities for all human protein pairs using several predictive features commonly used for PPIs prediction (Kotlyar and Jurisica 2006; Kotlyar 2011), such as protein sequence and structure, orthology, network topology, Gene Ontology, and gene co-expression. In particular, *FpClass* makes use of classification methods that are trained on such a diverse set of interaction evidences to recognize positive examples of truly interacting protein pairs from the negative examples of random, non-interacting pairs (Kotlyar and Jurisica 2006), where each protein pair is encoded as a feature vector of all used interaction evidences.

FpClass has shown to provide increased coverage of the interactome at 50% False Discovery Rate (FDR) (Kotlyar 2011), compared with probabilistic Bayesian-based models for PPI prediction with > 68% FDR (Rhodes et al. 2005; Scott and Barton 2007). Among such a comprehensive set of predictive features used to predict PPIs in the *FpClass* algorithm, it has been shown that network topology and protein domains are the two most powerful features for interaction prediction (Kotlyar 2011). Moreover, (Scott and Barton 2007) predicted human PPIs using diverse evidence types and also found that network topology was the most effective.

Methodology

We propose a novel particle swarm-based method for pattern discovery in PPI networks that combines both network topology and protein domains with network proximity information to make new, high-confidence PPI predictions. Over 80% of the predictions of the proposed method overlapped with the Top 1% predictions of the *FpClass* method (Kotlyar 2011). The proposed method also managed to predict a number of PPIs overlapped with publically-available interaction databases that were not discovered by the *FpClass* method, despite the comprehensive set of predictive features that *FpClass* uses.

Why immediate neighbors of target protein pairs?

Our proposed 3PI method attempts to discover topologically-similar patterns in the interactions among all immediate neighbors of a pair of interfacing proteins, because their interacting partners are likely to be biologically-relevant due to their close proximity in the interaction networks. Not only will discovering such topologically-similar patterns capture a strong similarity signal from protein interaction networks, but also the close proximity of proteins in these patterns (with at most 3 edges apart) will make them more likely to be functionally related. This is because proteins in these patterns are the immediate neighbors of a pair of physically interfacing proteins, which guarantees their Shortest Paths (SP) to be at most 3. SP = 3 if they are not common interaction neighbors of the pair of interfacing proteins (Fig. 1(a)), whereas SP = 2 if at least one of them is a common neighbor (Fig. 1(b)), or SP = 1 if they also happened to be directly interacting with one another (Fig. 1(c)).

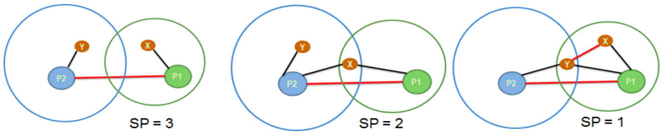


Figure 1: The minimum number of edges (SP) between X and Y ranges from 1 to 3 (max) in all possible cases.

PPI Predictions from Similar Patterns

For each pair of topologically-similar patterns in each case study, we generated a corresponding pair of non-overlapping protein lists. We ignored overlapping proteins at this stage in order not to bias our enrichment analysis results of the next stage. We used the DAVID annotation tool (Jiao et al. 2012) to perform InterPro domain enrichment analysis for each protein list. Enrichment results identified a number of overlapping domains that are statistically over-represented in both protein lists (P-value < 0.001). We selected the overlapping domain in both lists with the lowest P-value, and constructed a smaller pair of 'protein sets' of only those proteins enriched in the most-significant overlapping domain.

The corresponding members of pairs of enriched protein sets, A and B, are not only significantly over-represented in shared domains, but are already known to be interacting neighbors of the original pair of interfacing proteins, P$_1$

and P_2, respectively. Based on a proposed PPI knowledge transfer process between A and B, we hypothesize that P_2 will likely interact with all members in A, and P_1 will also likely interact with all members in B. Over 50% of our systematic predictions are overlapped with publically available interaction databases, and about 80% are not only significantly co-expressed in different gene expression datasets (as top one percentile in the co-expression matrices), but are also overlapped with the Top 1% results of the *FpClass* algorithm (Kotlyar and Jurisica 2006; Kotlyar 2011).

PSO-based Similar Local Pattern Discovery

Although far from complete, analyzing protein interaction networks and extracting meaningful patterns from the hundreds of thousands of currently known interactions among thousands of human proteins (which may only represent 10% of the entire interactome (Stumpf et al. 2008)) is undoubtedly a challenging task. PSO-based pattern discovery is, therefore, an important step in the proposed workflow of our methodology. The complete workflow of the 3PI method is shown in Fig. 2. We first extracted all interactions among the partners of each pair of interfacing proteins (e.g., UBC and GRB2 partners, as shown in Fig. 3), and represented them as two adjacency matrices (Fig. 4). This created a very sparse and challenging search space (e.g., about $2.9\text{x}10^{12}$ possible patterns of size *100x100* could be generated only from these two adjacency matrices). It is, therefore, impractical to apply exhaustive sliding window-based search to exactly discover the best similar patterns. To efficiently explore such a large search space, a PSO-based heuristic search is adopted, in which each particle has a four-dimensional position vector: X_1, Y_1, X_2, Y_2, which represents a pair of 2D points in the adjacency matrix pair. In particular, X_1 and Y_1 represent the upper-left corner of a flying *u x v* window on the first adjacency matrix, and X_2 and Y_2 represent the upper-left corner of a flying window (obviously of the same, user-defined size, e.g., *100x100*) on the other adjacency matrix.

However, a known issue in PSO-based search is swarm stagnation, which occurs when the rate of position changes (or velocities) that attract particles to the global best position is prematurely approaching zero. This situation, if left unhandled, may lead the swarm to being trapped in a local optimum, from which it cannot escape nor can it generate new better solutions. A common strategy to help recover the swarm from a stagnation situation is to restart the particles before premature convergence. The proposed methodology restarts particle positions and resets their velocity and memory using the swarm-wide performance on the objective function with relative to the global best performance, as shown in equation (1).

$$\frac{Average\ (f(P_i))}{f(P_g)} \geq \psi,\ \ f(P_g) > 0,\ \ f = |A \cap B| \qquad (1)$$

Where A & B are the flying window pair on both adjacency matrices, f is the objective (maximization) function, described as the size of the intersection between the two binary flying windows (or the size of the shared white positions representing common interaction patterns, or "common contacts"), $f(P_i)$ is the corresponding fitness vector of particles' historically best positions, $f(P_g)$ is the fitness value of the global best particle position, and ψ is the PSI threshold, which is a positive percentage value to be specified by the user. PSI value clearly affects the algorithm sensitivity to triggering the restart mechanism. The smaller the value of ψ, the more frequent the PSI condition will be satisfied during the algorithm run. Thus, the PSI threshold needs to be large enough to avoid unnecessarily restarting the swarm when the majority of particles are still exploring the search space and have not even become closer to the global best particle. In a previous study (Ahmed and Glasgow 2014), it was empirically found that a PSI threshold of 80% works better in the four-dimensional pair-wise pattern discovery problem.

Fig. 5 shows the search behavior of the proposed method over 1000 iterations, which within the first few iterations managed to discover patterns with over 80 common contacts. Not only that, but eventually, after a few restarts, the method discovered even better patterns with over 100 common contacts. In particular, the method retrieved several decent patterns with 103 - 121 common contacts – well before completing the specified maximum number of 1000 iterations. Sample of these interaction patterns are shown in Fig. 7. In order to get an idea if these patterns could just be generated by chance, we employed a random search method million times and checked how many common contacts were able to be identified just by random chance. As shown in Fig. 6, the *maximum* number of common contacts identified by random search was 75 in only 12 times, and the random search was able to identify patterns with only less than 10 common contacts most of the times, due to the complex and sparse nature of the problem search space. As shown in Fig. 8, we transformed all interactions in a similar pattern pair from the adjacency matrix representation to actual network topology, in order to visualize how the captured similarity in adjacency matrix has a corresponding similarity in PPI network topology. One way to measure the similarity between two sub-networks is to perform network clique analysis. Fig. 9 shows different groups of cliques (or sets of highly-interconnected nodes) in the two discovered sub-networks. As shown in Fig. 9, the highlighted "Cluste001" in both tables visualizes the local topological similarity of exactly the same number of 8 nodes, as well as almost the same number of edges (25 *vs.* 27) and density (0.96 *vs.* 0.89).

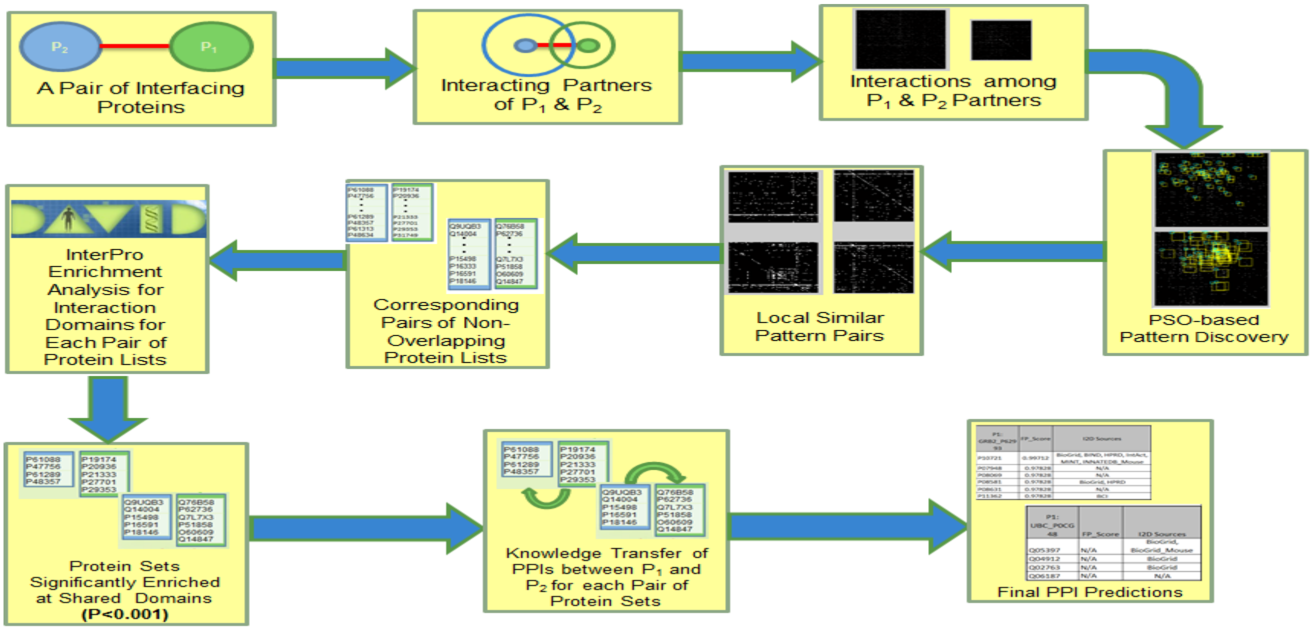

P₂
P₁
A Pair of Interfacing Proteins
Interacting Partners of P₁ & P₂
Interactions among P₁ & P₂ Partners
PSO-based Pattern Discovery
Local Similar Pattern Pairs
Corresponding Pairs of Non-Overlapping Protein Lists
InterPro Enrichment Analysis for Interaction Domains for Each Pair of Protein Lists
Protein Sets Significantly Enriched at Shared Domains (P<0.001)
Knowledge Transfer of PPIs between P₁ and P₂ for each Pair of Protein Sets
Final PPI Predictions

Figure 2: The complete workflow of the proposed method summarizing all 3PI steps. The three key steps in 3PI are: 1) PSO-based Pattern Discovery, 2) InterPro Enrichment Analysis for Interaction Domains, and 3) PPI Knowledge Transfer Process.

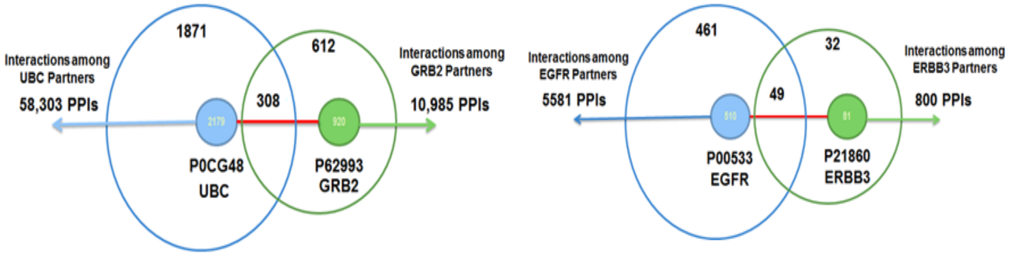
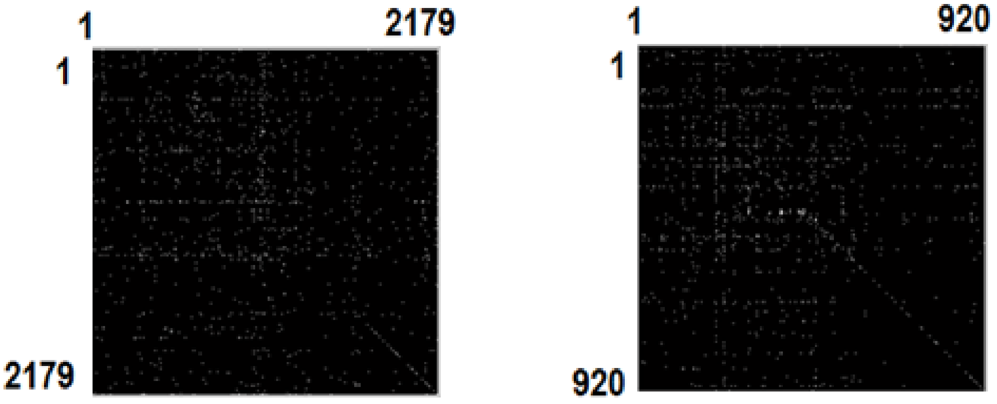

1871
612
308
Interactions among UBC Partners
58,303 PPIs
Interactions among GRB2 Partners
10,985 PPIs
2179
920
P0CG48
UBC
P62993
GRB2
461
32
49
Interactions among EGFR Partners
5581 PPIs
Interactions among ERBB3 Partners
800 PPIs
510
81
P00533
EGFR
P21860
ERBB3

Figure 3: Numbers of exclusive (and shared) immediate interaction partners for UBC & GRB2 (Left), and EGFR & ERBB3 (Right).



Interactions among UBC Immediate Partners (2179 Neighbour Proteins, 58,303 PPIs)

Interactions among GRB2 Immediate Partners (920 Neighbour Proteins, 10,985 PPIs)

Figure 4: A pair of adjacency matrices representing interactions among UBC and GBR2 partners.

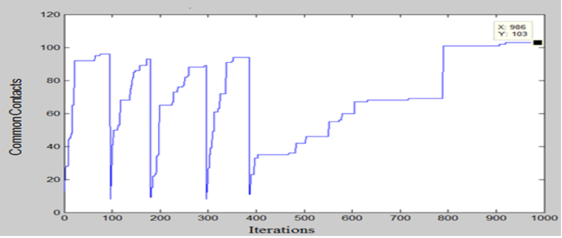


Figure 5: The proposed 3PI method discovered patterns with over 100 common contacts in less than 1000 iterations.

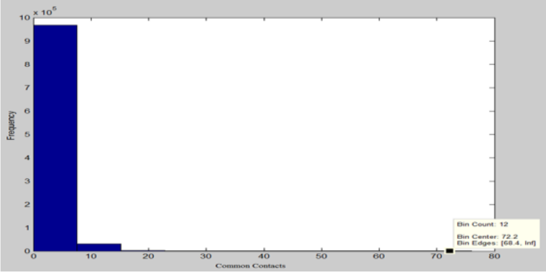


Figure 6: Random search failed to discover patterns with more than 75 common contacts over million iterations.

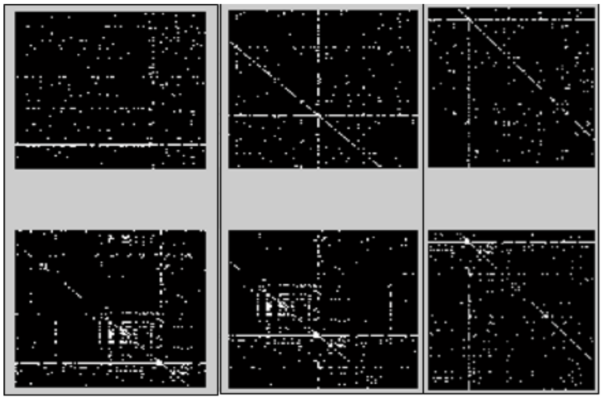


Figure 7: Samples of discovered local similar pattern pairs, each of size 100x100.

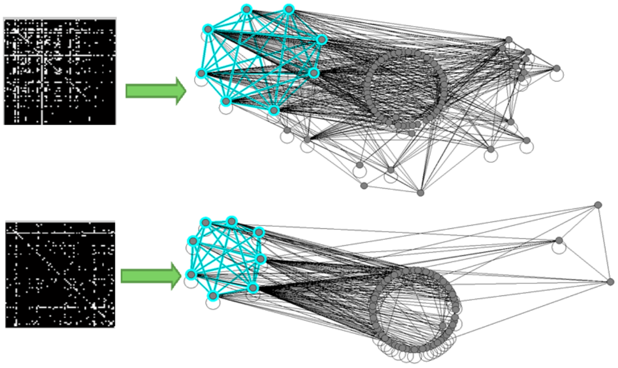
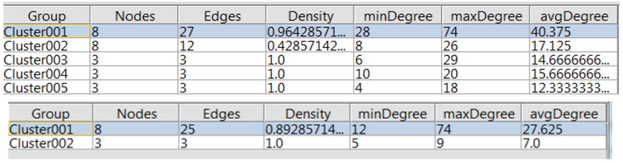


Figure 8: Corresponding local similarities in adjacency matrices and PPI network topology.



Group	Nodes	Edges	Density	minDegree	maxDegree	avgDegree
Cluster001	8	27	0.96428571...	28	74	40.375
Cluster002	8	12	0.42857142...	8	26	17.125
Cluster003	3	3	1.0	6	29	14.6666666...
Cluster004	3	3	1.0	10	20	15.6666666...
Cluster005	3	3	1.0	4	18	12.3333333...

Group	Nodes	Edges	Density	minDegree	maxDegree	avgDegree
Cluster001	8	25	0.89285714...	12	74	27.625
Cluster002	3	3	1.0	5	9	7.0

Figure 9: Network clique analysis showing groups of cliques (clusters) in the two sub-networks presented in the previous figure.

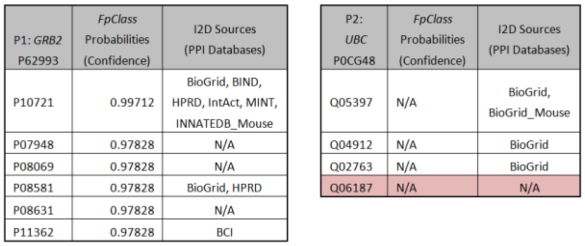
Results and Discussion

We first present the results of our proposed 3PI method on two case studies. The first case study analyzes the interactions among the immediate neighbors of a pair of interfacing 'hub' proteins. In this case study, we used the two most highly connected human proteins in the Interologous Interaction Database (I2D) (Brown and Jurisica 2007), namely UBC_P0CG48 and GRB2_P62993. I2D is one of the most comprehensive online sources of known and predicted PPIs for 6 different organisms; in this paper, we used the latest version of Human I2D 2.0, released in 2012. Next, we analyzed the interactions among the neighbors of a pair of interfacing less-connected 'bottleneck' proteins, which are often found to be biologically important (Pržulj, Wigle, and Jurisica 2004; Yu et al. 2007). The two interfacing bottleneck proteins used in the second case study are: EFGR_P00533 and ERBB3_P21860.

Based on such local topologically-similar patterns, several PPI predictions for UBC, GRB2, EFGR and ERBB3 were systematically generated for both case studies. A sample of these predictions are presented in Table 1 and Table 2, and summarized in Fig. 10. As shown in Table 1 and Table 2, 72% and 80% of the predictions overlapped with *FpClass* predictions with high interaction probabilities (> 0.97 and 0.90, respectively). This means 4 in 5 predictions are among the Top 1% of the *FpClass* predictions. Moreover, 52% of the predictions have been validated with at least one publically available interaction databases, e.g., BioGrid, BIND, IntAct, HPRD and MINT.

Although about 1 in 6 predictions received low interaction probabilities (< 0.74), which are reported under the '*FpClass'* column as "N/A" in Table 1(b), 75% of these predictions are overlapped with the BioGrid database. This suggests that the proposed 3PI method that only uses interaction domains and topology information can complement *FpClass* predictions, despite the comprehensive set of predictive features that *FpClass* uses to estimate interaction probabilities.

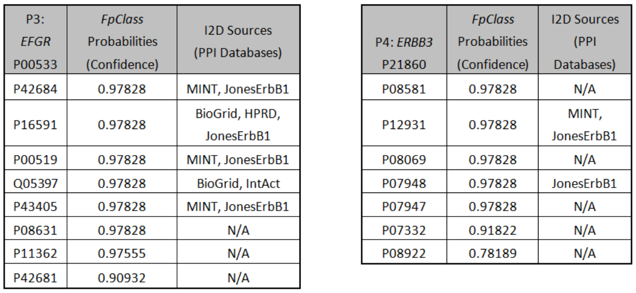
Third, the highlighted interaction in Table 1(b) between P0CG48 (*UBC*) and Q06187 (*BTK*) is the only PPI prediction that neither overlapped with *FpClass* high-confidence predictions, nor with any manually curated databases in Human I2D 2.0, released in 2012. However, a thorough PubMed search has revealed two PubMed publications (Kim et al. 2011; Wagner et al. 2011) that experimentally supported this particular interaction between *UBC* and *BTK*. Moreover, this interaction has been recently imported to the BioGrid database, with a comment saying that "this interaction was experimentally detected by Affinity Capture-MS assay and manually curated" based on these two PubMed publications. This literature validation gives yet another example on the ability of the proposed *3PI* method to discover high-confidence PPIs, and supplement the prediction results of *FpClass* method, based primarily on pattern discovery in protein networks and shared domain interactions among immediate neighbors of pairs of interfacing proteins. Furthermore, Fig. 10 shows that 80% the proposed PPI predictions overlap with those protein pairs whose gene co-expression values appear among the Top 1% in different gene expression datasets across normal and tumor tissues.



P1: *GRB2* P62993	*FpClass* Probabilities (Confidence)	I2D Sources (PPI Databases)
P10721	0.99712	BioGrid, BIND, HPRD, IntAct, MINT, INNATEDB_Mouse
P07948	0.97828	N/A
P08069	0.97828	N/A
P08581	0.97828	BioGrid, HPRD
P08631	0.97828	N/A
P11362	0.97828	BCI

P2: *UBC* P0CG48	*FpClass* Probabilities (Confidence)	I2D Sources (PPI Databases)
Q05397	N/A	BioGrid, BioGrid_Mouse
Q04912	N/A	BioGrid
Q02763	N/A	BioGrid
Q06187	N/A	N/A

Table 1: PPI predictions for (a) GBR2 & (b) UBC, along with their corresponding FpClass confidences.



P3: *EFGR* P00533	*FpClass* Probabilities (Confidence)	I2D Sources (PPI Databases)
P42684	0.97828	MINT, JonesErbB1
P16591	0.97828	BioGrid, HPRD, JonesErbB1
P00519	0.97828	MINT, JonesErbB1
Q05397	0.97828	BioGrid, IntAct
P43405	0.97828	MINT, JonesErbB1
P08631	0.97828	N/A
P11362	0.97555	N/A
P42681	0.90932	N/A

P4: *ERBB3* P21860	*FpClass* Probabilities (Confidence)	I2D Sources (PPI Databases)
P08581	0.97828	N/A
P12931	0.97828	MINT, JonesErbB1
P08069	0.97828	N/A
P07948	0.97828	JonesErbB1
P07947	0.97828	N/A
P07332	0.91822	N/A
P08922	0.78189	N/A

Table 2: PPI predictions for (a) EFGR & (b) ERBB3, along with their corresponding FpClass confidences.

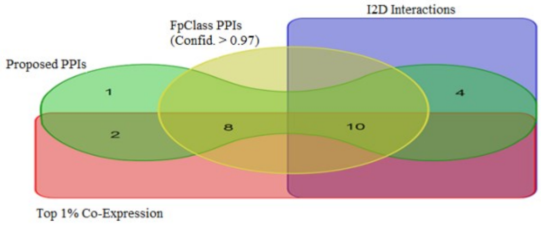


Figure 10: Overlapping proposed PPIs with I2D, the Top 1% Co-Expression and FpClass Results (with high-confidence > 0.97).

Further Analysis for EGFR and ERBB2

We further tested the proposed method on another pair of bottleneck, cancer-related proteins (EFGR_P00533 and ERBB2_ P04626), as shown in Fig. 11(a). Similar observations could be drawn from the prediction results of this third case study, as shown in Fig. 11(b). First, about 5 in 6 predictions overlapped with *FpClass* predictions with high interaction probabilities (> 0.97). This means over 80% of the prediction are again among the Top 1% of the *FpClass* predictions with interaction probability greater than 0.97. Moreover, the remaining ~16% of the predictions have fairly high interaction probabilities in *FpClass* (ranging between 0.90 and 0.93), and one of them is already overlapped with the MINT interaction database. Last, similar to previous case studies, about 50% of the predictions in the third case study are overlapped with at least one publically available interaction database in I2D.

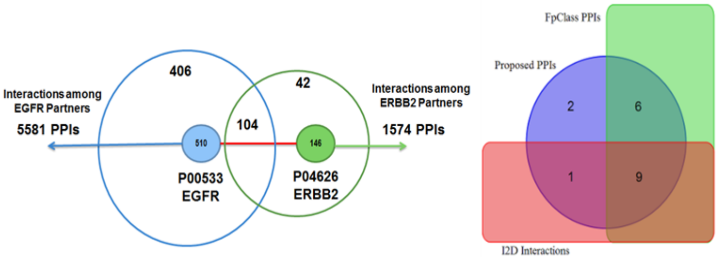


Figure 11: (a) Interaction partners of EGFR and ERBB2 (Case Study #3). (b) Overlapping proposed PPIs with I2D and the Top 1% FpClass PPIs (with high-confidence > 0.97)

Conclusions and Future Work

This paper, to the best of our knowledge, presents the first study in the literature that applies Particle Swarm Optimization to the pattern discovery problem in protein interaction networks, as well as the first to make use of the resulting topologically-similar patterns of close proximity and protein domains in predicting high-confidence PPIs. Discovering such similar interaction patterns is not the sole reason for capturing a strong similarity signal from PPI networks, with high biological relevance. Another reason for such a strong similarity signal is due to the close proximity of proteins in these patterns, which are at most 3 edges apart. While the study is still in its early stages, the encouraging high-confidence results, validated by more than one computational/experimental source, suggest a promising novel class of PPI prediction techniques using pattern discovery in PPI networks.

Future work includes conducting more in-depth analysis and fine-tuning for each step in the methodology workflow, as well as providing the community with an online tool to suggest novel potential interaction partners for interfacing protein pairs of interest to their studies. We also plan to apply the proposed method on a large proteome scale towards a more complete and accurate human interactome, which is currently far from complete (with up to 70% missing interactions estimated) and very noisy (with as high as 60% false interactions estimated). Our ultimate goal is expanding the knowledge of high-confidence PPIs starting from the known 10%, as well as the known interaction domains and proximity information, using an efficient, low-cost and robust machine learning technique for topology-based pattern discovery in PPI networks, which does not necessarily require protein structure or even sequence information.

Acknowledgments

In no particular order, we thank Sara Rahmati for useful discussions about gene co-expression values and their correlation with PPIs, Max Kotlyar for critical discussions about PPIs and *FpClass* method, Lili Wang for interesting chats about DAVID enrichment tool and network proximity, Renatas Minkstimas for the idea of network clique analysis and Chiara Pastrello for helping with the PubMed literature validation.

References

Aebersold, R., and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature* 422(6928): 198-207.

Ahmed, H. R., and Glasgow, J. I. 2014. An Improved Multi-Start Particle Swarm-based Algorithm for Protein Structure Comparison. In *2014 Genetic and Evolutionary Computation Conference (GECCO 2014)*, Vancouver, Canada

Brown, K. R., and Jurisica, I. 2007. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome biology* 8(5): R95.

Browne, F.; Zheng, H.; Wang, H.; and Azuaje, F. 2010. From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. *Advances in Artificial Intelligence* 2010: 15.

Cannistraci, C. V.; Alanis-Lobato, G.; and Ravasi, T. 2013. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* 29(13): i199-i209.

Chen, X.-W., and Liu, M. 2005. Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 21(24): 4394-4400.

del Valle, Y. et al. 2008. Particle swarm optimization: basic concepts, variants and applications in power systems. *Evolutionary Computation, IEEE Trans on* 12(2): 171-195.

Dimitrakopoulos, C. M. et al. 2012. Efficient Computational Construction of Weighted PPI Networks Using Adaptive Filtering Techniques Combined with Natural Selection-Based Heuristic Algorithms. *J of Sys Bio and Biomed Tech* 1(2): 20-34.

Dolinski, K. D., and Botstein, D. 2007. Orthology and functional conservation in eukaryotes. *Annu Review Genetics* 41: 465-507.

Fariselli, P.; Pazos, F.; Valencia, A.; and Casadio, R. 2002. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *European J of Biochem* 269(5): 1356-1361.

Ge, H.; Liu, Z.; Church, G. M.; and Vidal, M. 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature genetics* 29(4): 482-486.

Hassan, R.; Cohanim, B.; De Weck, O.; and Venter, G. 2005. A comparison of particle swarm optimization and the genetic algorithm. In *Proceedings of the 46TH AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Austin, Taxes.

Jiao, X. et al. 2012. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28(13): 1805-1806.

Jurisica, I., and Wigle, D. 2010. *Knowledge Discovery in Proteomics*. CRC press.

Kemmeren, P. et al. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell* 9(5): 1133-1143.

Kim, W. et al. 2011. Systematic and quantitative assessment of the ubiquitin-modified proteome. *Molecular cell* 44(2): 325-340.

Kotlyar, M. 2011. Prediction of PPIs and Essential Genes Through Data Integration *PhD Thesis, University of Toronto*.

Kotlyar, M., and Jurisica, I. 2006. Predicting PPIs by association mining. *Information systems frontiers* 8(1): 37-47.

Kuchaiev, O. et al. 2010. Topological network alignment uncovers biological function and phylogeny. *J R Soc Interface* 7(50): 1341-1354.

Li, M.; Wu, X.; Wang, J.; and Pan, Y. 2012. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. *BMC bioinformatics* 13: 109.

MacBeath, G. 2002. Protein microarrays and proteomics. *nature genetics* 32: 526-532.

Milenkoviæ, T., and Pržulj, N. 2008. Uncovering biological network function via graphlet degree signatures. *Cancer informatics* 6: 257.

Pržulj, N.; Wigle, D.; and Jurisica, I. 2004. Functional topology in a network of protein interactions. *Bioinformatics* 20(3): 340-348.

Qi, Y. 2008. Learning of protein interaction networks. PhD Thesis, Universitat Pompeu Fabra, Spain.

Rhodes, D. R. et al. 2005. Probabilistic model of the human protein-protein interaction network. *Nat Biotech* 23(8): 951-959.

Saito, R.; Suzuki, H.; and Hayashizaki, Y. 2003. Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics* 19(6): 756-763.

Scott, M. S., and Barton, G. J. 2007. Probabilistic prediction and ranking of human PPIs. *BMC bioinformatics* 8: 239.

Sharan, R.; Ulitsky, I.; and Shamir, R. 2007. Network-based prediction of protein function. *Molecular sys biology* 3(1): 88.

Shen, J. et al. 2007. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences* 104(11): 4337-4341.

Stumpf, M. P. et al. 2008. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America* 105(19): 6959-6964.

Suter, B.; Kittanakom, S.; and Stagljar, I. 2008. Two-hybrid technologies in proteomics research. *Current opinion in biotechnology* 19(4): 316-323.

Von Mering, C. et al. 2002. Comparative assessment of large-scale data sets of PPIs. *Nature* 417(6887): 399-403.

Wagner, S. A. et al. 2011. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Molecular & Cellular Proteomics* 10(10).

Yu, H. et al. 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology* 3(4): e59.