Comparison of Sequence- and Structure-Based Protein-Protein Interaction Sites

Kevin Dick
Systems and Computer Engineering
Carleton University
Ottawa, Ontario
Email: kevin.dick@carleton.ca

Dr. James Green
Systems and Computer Engineering
Carleton University
Ottawa, Ontario
Email: jrgreen@sce.carleton.ca

***Abstract*—Computational protein-protein interaction (PPI) prediction is a diverse field with multiple paradigms generating insightful interaction interface information. The shortcomings of one approach are often the strength of another and establishing the agreement between methodologies is valuable for the development of novel PPI prediction techniques. This study represents the first large-scale comparison of PPI sites determined through a sequence-based method (PIPE-Sites) and a structure-based method (PiSITEs). A set of interactions (n = 3,109) amenable to analysis by both methods was examined. Interestingly, the distributions of the sizes of the predicted interaction sites have similar means and identical median values. Using the Sørensen-Dice similarity coefficient and independent randomization testing, we determined the degree of agreement of the predicted sites of interaction for both methods to be statistically significant ($p < 0.001$). Finally, applying the hypergeometric test and Q-Analysis, we identified 491 interactions with significantly heightened agreement ($p < 0.002$). These interactions represent a broad range of biological function including transcriptional regulation, cell proliferation, cytoskeletal dynamics, and apoptosis. These findings corroborate the joint application of these paradigms for future PPI prediction studies.**

I. Introduction

Protein-protein Interactions (PPIs) are fundamental cellular dynamics enabling life, occurring in biomolecular processes such as protein transport, signal transduction, cellular metabolism and cellular division. Experimental validation techniques of PPIs are resource-expensive and often result in high error rates [1]. Provided the recent rise in available protein sequences and validated PPIs for a number of species, the computational prediction of PPIs aims to compliment wet-lab validation techniques by assigning an objective measure to the confidence of a putative interaction. A variety of approaches exist given the diversity of available data [2], however this work focuses on sequence-based and structure-based approaches. Specifically, this paper examines two methods which not only predict the presence of a PPI, but also the protein subsequence responsible for the PPI.

Sequence-based methods are broadly applicable, generally requiring primary sequence information only and are thus suited to proteome-wide prediction. PIPE is one such method, which is highly tuned to achieve both high sensitivity and specificity, and is also highly optimized such that a protein pair can be evaluated in a fraction of a second [3]. Structure-based methods base their prediction on the three-dimensional coordinates of all atoms within a protein and its known interactions [4]. When such information is available, these methods exhibit high rates of accuracy. However, determining the 3D structure of a protein is complex and costly, and this information is typically only available for a small fraction of proteins in an organism. Critically evaluating the agreement between sequence- and structure-based methods of PPI site determination would provide indication of complementarity and highlight the importance of the regions of predicted interaction. Determining sequence-based performance on the limited set of well characterized PPIs available to a structure-based method has important implication for the validity of sequence-based prediction on uncharacterized PPIs.

A. PIPE-Sites: A Sequence-Based Method for PPIs

The Protein-protein Interaction Prediction Engine (PIPE) is a sequence-based PPI predictor which examines a sliding window (w = 20) along the amino acid (AA) sequence of two proteins, predicting the likelihood of interactions by comparing these windows to a database of protein pairs known to interact [5]. PIPE-Sites is an extension the PIPE algorithm which not only predicts PPIs but also suggesting which regions, if any, mediate this interaction. By examining the weight of evidence for the predicted PPI corresponding to each window within a protein, a scoring method identifies up to three sequence regions (PIPE "sites") believed to support the interaction [6]. This approach leverages sequence conservation information and can therefore define regions that not only make up the defined interface of interaction, but also identify those surrounding regions required to structurally maintain an interaction. As a sequence-based method, this method is applicable to any two proteins, independent of availability of structural information.

B. PiSITE: A Structure-Based Method for PPIs

Sequence regions determined by PIPE-Sites correspond to those regions that must necessarily be conserved to allow two proteins to interact. By altering our definition to mean only those residues that form the actual binding interface requires consideration of three-dimensional structures, available from protein database (PDB) files. PiSITEs comprises of 110,325

978-1-5090-0935-0/16/$31.00 ©2016 IEEE

proteins from 51,482 PDB files [7]. For a given protein, BLASTn sequence analysis is used to identify multiple PDB structures capturing interactions with that protein. The interaction interfaces from all relevant structures are considered and their physical and chemical properties then used to infer PPIs based on interface similarity. An interface residue was identified when the minimum distance between atoms for a given residue pair was <4.0 Å. This dataset therefore identifies those AAs likely to physically interact in a PPI.

II. Methods

The PIPE-Site *Homo sapiens* data was acquired from previous work by Pitre *et al.* [6], [8]. The 110,325 available PiSITEs PDB chain files were acquired from the online database [7] (accessed: February 29th, 2016). Interactions specific to *Homo sapiens* were extracted and each PDB ID was converted to its UniProtKB Accession number [9]. In cases where multiple PDB ID's mapped to the same Uniprot Accession number, pairwise sequence alignment was applied between the PDB file sequence and the canonical sequence, acquired from Uniprot. The BioPython package [10] *pairwise2* module, applying the Needleman-Wunsch algorithm [11] with the PAM30 substitution matrix, was used, selecting for the highest scoring sequence as the representative. The intersection of proteins present in both datasets was acquired (n = 3,080) as well as the intersection of predicted PPIs (n = 3,109), hereafter referred to as the 'IntersectionPPI' dataset. The sizes of predicted interaction regions/interfaces were compared for the entire dataset of both methods (Fig. 1) while the analysis of agreement between PIPE-Sites and PiSITEs interaction regions within the IntersectionPPI dataset was performed according to Algorithm 1. PIPE-Sites generates a continuous region of interaction whereas PiSITEs reports discrete AA positions (*i.e.* a non-contiguous region) known to interact between two proteins. Of the three predicted PIPE-Sites, the "BestSite" for a given PPI was determined as the site with maximum degree of agreement (DoA), or minimum p-value, depending on the context of study. The DoA was computed using Eq. 1, assuming the PIPE-Site continuous data to be a discretized sequence of AAs.

$$DoA = \frac{2k}{PS + Pi} \tag{1}$$

The metric is adapted from the Sørensen-Dice coefficient [12] which has been widely applied in studies of genetic relationships [13], image segmentation [14], and ecology [15]. For a given interaction between two proteins, the value k represents the number of AAs in agreement with both methods: the number of PiSITEs AAs known to physically participate in the interaction that fall within the predicted PIPE-Site region. The PS value is the length (in AA) of the predicted site of interaction between the two proteins from PIPE-Sites while Pi represents the total number of AAs that physically participate in the PPI. N is the total length (in AA) of the protein sequence participating in the PPI.

Randomization testing (RT) with the DoA was used to determine whether or not the PIPE-Sites significantly overlapped with PiSITEs defined residues. The analysis in Algorithm 1 was repeated on a pseudo PiSITEs dataset where the AAs participating in a given PPI were randomly assigned to new positions. The average DoA value over all RT PPIs was computed for 1,000 independent tests and the distribution of these means was compared against the original (Fig. 2).

Algorithm 1 Site Agreement Analysis: Computes the Degree of Agreement and p-value between all predicted sites of interaction common to both the PIPE-Sites and PiSITEs, the IntersectionPPI dataset.

for each PPI ∈ IntersectionPPI **do**
  Sites = Extract start,end position of three PIPE-Sites
  for each Site ∈ Sites **do**
    x = getTotalPiSITEAminoAcids()
    PS = Site.end - Site.start
    Pi = size(x)
    k = numSuccessfulInSite(x, Site.start, Site.end)
    $DoA = \frac{2*k}{PS+Pi}$
    p = hypergeo(x, PS, Pi, k)
  end for
  $BestSite$ = max(Sites.DoA)
  $BestDoA$ = $BestSite$.DoA
  $BestPval$ = min(Sites.p)
end for

Establishing that the two methods significantly overlap (p < 0.001) from RT, the hypergeometric test (Eq. 2) was then applied to each interaction in the IntersectionPPI dataset to identify PPIs with statistically significant overlap for further inquiry. Considering that multiple tests were conducted, Q-Analysis [16] was used to limit the number of false discoveries among all p-value. A q-value ≤ 0.01 was applied to all p-values to limit the false discovery rate (FDR) to less than 5%, implying $\alpha = 0.002$ should be used to define statistical significance (Fig. 3).

Finally, inquiry into the set of statistically significant PPIs (n = 491, p < 0.002) was undertaken to identify the nature of the PPIs in high agreement.

$$P(k \mid N, Pi, PS) = \frac{\binom{Pi}{k}\binom{N-Pi}{PS-k}}{\binom{N}{PS}} \tag{2}$$

III. Results

Comparison between the PIPE-Site interaction prediction tool and PiSITEs began by comparing the distribution of site sizes across the original datasets. Fig. 1 illustrates the distributions of the length of all sites, as defined by PIPE-Sites and PiSITEs. While some sites of length greater than 120 AAs were reported (not included in Fig. 1), 99.99% of the data falls below this site size.

The mean and median were computed for each distribution; interestingly, a median value of 27 was computed for both sets. The [1,19] gap of missing values in the PIPE-Sites distribution is a result PIPE algorithm which considers a binding interface to be a continuous sequence of a minimum of

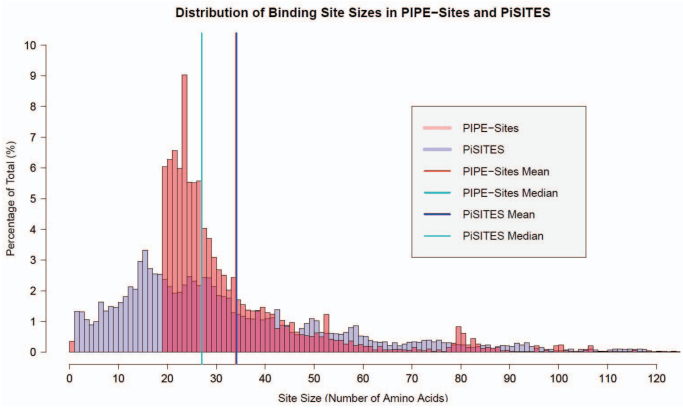

Fig. 1. Comparison of the distribution of interface site sizes in the complete PiSITE and PIPE-Site datasets. The PIPE-Site dataset considered its three predicted interaction sites for each protein in every interaction (n = 212,781) while PiSITEs had an average of 4.03 interactors per protein (n = 12,434). The PiSITE distribution had a mean of 34.1 and a median of 27. The PIPE-Sites distribution had a mean of 34.0 and a median of 27. PIPE-Sites which predicted sites of size >50% of the total protein length were excluded.

20 AAs. Although PIPE-Sites can report up to three interaction regions, when only one or two suitable regions are identified the remaining regions are considered to have zero length here. These empty sites, however, were not considered in the analysis and were excluded from the intersection dataset. PiSITEs binding interfaces are non-contiguous, discrete positions along a sequence, permitting site sizes of any positive value greater than one AA.

The IntersectionPPI data was used to determine whether significant overlap existed between methods considering the DoA metric. Under the null hypothesis that no significant overlap existed, 1,000 independent RTs were performed to generate a distribution of mean DoA values to compare against the actual observed DoA (Fig. 2). Since the original mean distinctly exceeds the distribution, we reject the null hypothesis ($p < 0.001$) and establish that significant overlap exists between the two methods.

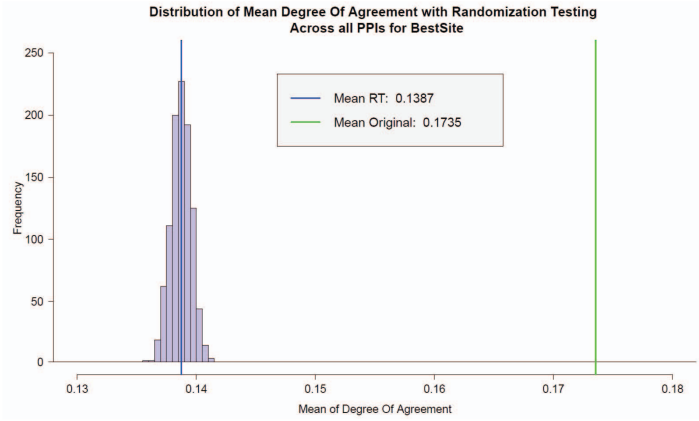

Fig. 2. Distribution of averaged DoA for all PPIs from 1,000 independent randomization tests. The mean DoA for all PPIs in original data is plotted in green.

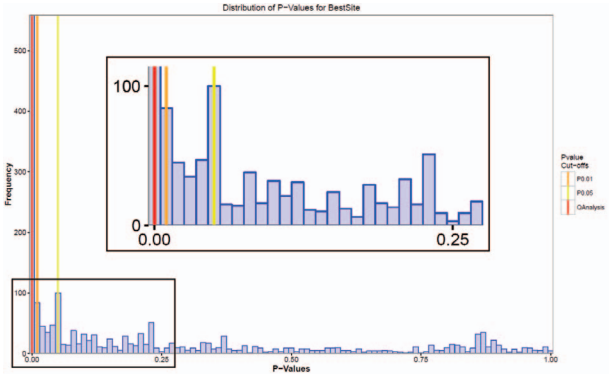

Fig. 3. Distribution of p-values for the BestSite (minimum p-value from the hypergeometric test applied to the three PIPE-Sites). The red line (QAnalysis) is $\alpha = 0.002$, obtained when applying a threshold of $q \leq 0.01$; the orange and yellow line are $\alpha = 0.01$ and $\alpha = 0.05$, respectively. The [0.99-1.0] bin was omitted to better resolve the figure. The boxed panel highlights the p-value range from [0 - 0.25].

The identification of PPIs of greatest interest was accomplished using the hypergeometric test to examine each protein pair for significantly high agreement between the sequence- and structure-based PPI site definitions. Since multiple hypotheses were being tested (n = 3,109), Q-value analysis was applied to limit the FDR and determine an appropriate p-value threshold. We selected to operate at a FDR ≤ 0.05 resulting in α= 0.002. This identified 491 significant PPIs of which <20 may be false positives. These PPIs were then examined further using Uniprot to determine the nature and biological function.

Proteins with diverse biological functions were identified, however 30% were found to be involved in regulation of transcription based on Gene Ontology (GO) terms and another 10% were small GTPases, members of the Ras-superfamily. The DNA Fragmentation Factor subunit α and β (DFFA/DFFB) interaction was identified as having one of the highest degrees of significance among these PPIs (Fig. 4).

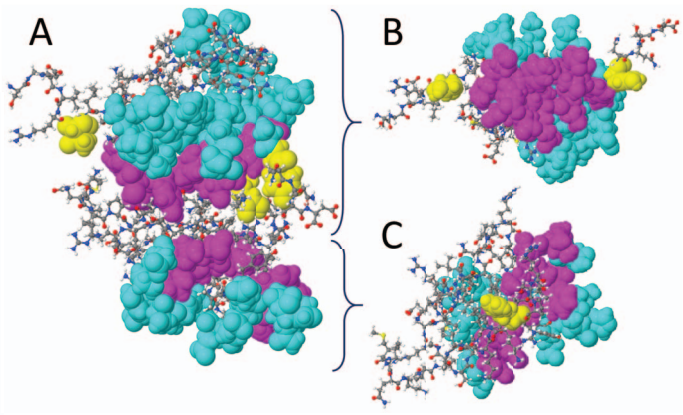

Fig. 4. Molecular view of the DFFA/DFFB PPI. Cyan molecules represent the predicted PIPE-Site region, yellow molecules represent the PiSITEs AAs, and the magenta molecules are the AAs in agreement with both methods. Panel A depicts the interacting molecules, Panel B and Panel C represents DFFB and DFFA, respectively, when rotated along the horizontal axis to better resolve the interaction region. Generated using the Jmol open-source software.

IV. Discussion

The primary objective of this work was to determine how two methods adopting unique paradigms for PPI site determination compare. The two are inherently complimentary; the weaknesses in one approach being the strength of the other. Evaluating the agreement between the two is therefore valuable for the development of novel PPI prediction techniques.

A. Interaction Site Sizes are Similarly Distributed

Considering the original set of all PPIs in both methods, we sought to determine if the site sizes were similarly defined. We found the two methods to be distributed with similar means and identical median values (Fig. 1). Considering that PiSITE only includes those AA in direct contact, while PIPE-Sites identifies the wider region required to support the interaction, the high agreement in overall region size was somewhat surprising. This result implies that for uncharacterized proteins, the predicted site size from PIPE-Sites is likely to be in agreement with reality.

B. Statistically Significant Degree of Agreement Between Methods

We found the DoA between both methods to be significant ($p < 0.001$) indicating that the sites defined by each approach produce comparable interaction interfaces. This agreement further supports the complimentarity between methods with one determining those residues physically participating in a PPI while the other additionally identifies those residues necessary to support that interaction.

C. Significant PPIs Highlight Complimentarity of Methods

Establishing the agreement between methods we investigated the 491 PPIs with significantly high ($p < 0.002$) overlap. GO terms identified approximately one third as being involved in transcriptional regulation and another subset to be small GTPases, members of the Ras-superfamily, including RHO6, RHOA, ARHGEF25, RAC1, RAB11A, and RAB11B. Functionally important for cell proliferation and cytoskeletal dynamics, mutations in members of the Ras signalling pathway are commonly found in cancers. The DNA fragmentation factor subunit α and β (DFFA/DFFB) PPI was also identified and selected for its importance in the inhibition of apoptosis (Fig. 4). The majority of the physically interacting residues are captured by the PIPE-Site in addition to highlighting those AAs believed to be necessary to support the interaction. Those residues independent of the interaction (ball and stick) are unlikely to be functionally important to this PPI and therefore absent from both methods.

V. Conclusion

This study represents the first comparison of PPI sites determined through sequence- and structure-based methods. Despite the difference in approaches, a high degree of agreement was observed between these two methods over a set of 3,109 PPIs amenable to analysis using both techniques. This agreement was observed in both the total size of the interaction region and also in the specific AA identified as supporting the interaction. This study serves to validate both approaches and suggests their effective combination in future studies.

Acknowledgment

The authors would like to thank Dr. Andrew Schoenrock, Dr. Frank Dehne, and the members of the Carleton University Bioinformatics Research Group for insightful discussion in support of this work.

References

[1] L. Skrabanek *et al.*, "Computational prediction of protein–protein interactions," *Molecular biotechnology*, vol. 38, no. 1, pp. 1–17, 2008.

[2] Y. Park, "Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences," *BMC bioinformatics*, vol. 10, no. 1, p. 1, 2009.

[3] A. Schoenrock *et al.*, "Mp-pipe: a massively parallel protein-protein interaction prediction engine," in *Proceedings of the international conference on Supercomputing*, 2011, pp. 327–337.

[4] J. D. Watson, R. A. Laskowski, and J. M. Thornton, "Predicting protein function from sequence and structural data," *Current opinion in structural biology*, vol. 15, no. 3, pp. 275–284, 2005.

[5] S. Pitre *et al.*, "Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC bioinformatics*, vol. 7, no. 1, p. 1, 2006.

[6] A. Amos-Binks *et al.*, "Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences," *BMC bioinformatics*, vol. 12, no. 1, p. 225, 2011.

[7] M. Higurashi, T. Ishida, and K. Kinoshita, "Pisite: a database of protein interaction sites using multiple binding states in the pdb," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D360–D364, 2009.

[8] S. Pitre *et al.*, "Short co-occurring polypeptide regions can predict global protein interaction maps," *Scientific reports*, vol. 2, 2012.

[9] R. Apweiler *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D115–D119, 2004.

[10] B. Chapman and J. Chang, "Biopython: Python tools for computational biology," *SIGBIO Newsl.*, vol. 20, no. 2, pp. 15–19, Aug. 2000. [Online]. Available: http://doi.acm.org/10.1145/360262.360268

[11] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[12] T. Sørensen, "{A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}," *Biol. Skr.*, vol. 5, pp. 1–34, 1948.

[13] E. Kosman and K. J. Leonard, "Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species," *Molecular Ecology*, vol. 14, no. 2, pp. 415–424, 2005. [Online]. Available: http://dx.doi.org/10.1111/j.1365-294X.2005.02416.x

[14] K. H. Zou *et al.*, "Statistical validation of image segmentation quality based on a spatial overlap index 1: Scientific reports," *Academic radiology*, vol. 11, no. 2, pp. 178–189, 2004.

[15] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[16] J. D. Storey, "The positive false discovery rate: a bayesian interpretation and the q-value," *Annals of statistics*, pp. 2013–2035, 2003.