Pamukkale Univ Muh Bilim Derg, 25(2), 215-222, 2019

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Taxonomic diversity-based domain interaction prediction

Taksonomik çeşitlilik tabanlı protein altünite etkileşim tahmini

Erdem TÜRK[1*], *Barış Ethem SÜZEK*[2]

[1,2]Department of Computer Engineering, Engineering Faculty, Mugla Sitki Kocman University, Mugla, Turkey.
erdemturk@mu.edu.tr, barissuzek@mu.edu.tr

Received/Geliş Tarihi: 19.04.2018, Accepted/Kabul Tarihi: 05.11.2018
doi: 10.5505/pajes.2018.18828
* Corresponding author/Yazışılan Yazar
Research Article/Araştırma Makalesi

Abstract

Identification of protein domain-domain interactions (DDIs) is an essential step in understanding proteins' functional and structural roles. MirrorTree is a DDI prediction method that is based on the principle of interacting proteins' co-evolution. However, this method is sensitive to taxonomic diversity and evolutionary span within the two protein homolog sets compared to predict DDI. In this work, we propose a new MirrorTree-based DDI prediction method, namely Taxonomic Diversity-based Domain Interaction Prediction (TAXDIP). TAXDIP improves the MirrorTree method by adding a sampling step that favors representation of higher-level taxonomic ranks (e.g. family over species) in two protein homolog sets prior to their comparison. This additional step ensures increased evolutionary span within protein homolog sets. TAXDIP is first assessed using a set containing 6,514 positive (interacting) domain pairs and a negative (non-interacting) set of equal size containing randomly generated domain pairs with no known interactions. TAXDIP achieved 71.0% sensitivity and 63.0% specificity on this set. Next, a benchmark-set containing 500 interacting and 500 non-interacting domain pairs is used to compare the performance of TAXDIP against DDI prediction methods ME and RDFF. TAXDIP showed better sensitivity and specificity than RDFF. While TAXDIP's sensitivity is better than ME, its specificity remained below ME. In conclusion, TAXDIP, with its performance, is a viable alternative to existing prediction methods. Furthermore, given TAXDIP's true predictions are overlapping with, and furthermore, complementing other DDI prediction methods, TAXDIP has a strong position in becoming part of a meta-DDI prediction method that combines multiple methods to build a consensus prediction.

Keywords: Protein domain-domain interactions, Protein co-evolution, Protein functional analysis

Öz

Protein altünite-altünite etkileşimlerinin (AAE) belirlenmesi, proteinlerin fonksiyonel ve yapısal rollerinin anlaşılmasında önemli bir adımdır. MirrorTree, etkileşen proteinlerin birlikte-evrimi prensibine dayanan, bir AAE tahmin yöntemidir. Ancak bu yöntem, AAE tahmin etmek için karşılaştırılan iki protein homolog kümesindeki taksonomik çeşitliliğe ve evrimsel açıklığa duyarlıdır. Bu çalışmada Taksonomik Çeşitliliğe Dayalı Protein Altünite Etkileşimi Tahmini (TAXDIP) olarak adlandırılan MirrorTree tabanlı yeni bir protein AAE tahmin yöntemi önermekteyiz. TAXDIP, iki protein homolog kümesini karşılaştırmadan önce, bunlarda daha yüksek düzeydeki taksonomik sıraların (ör. Tür yerine Aile) temsil edilmesini destekleyen bir örnekleme adımı ekleyerek, protein homolog kümeleri içindeki evrimsel kapsamın artmasını sağlar. TAXDIP öncelikle deneysel olarak doğrulanmış 6.514 pozitif (etkileşimli) altünite çiftini ve aynı sayıda, bilinen etkileşimleri olmayan, rastgele oluşturulmuş negatif (etkileşmeyen) altünite çiftini içeren bir küme kullanılarak değerlendirildi. TAXDIP bu kümede %71,0 duyarlılık ve %63,0 özgüllük elde etti. Daha sonra, TAXDIP'in performansının ME ve RDFF adlı AAE tahmin yöntemiyle karşılaştırılması için, 500 etkileşimli ve 500 etkileşmeyen altünite çiftini içeren, bir kıyaslama kümesi kullanıldı. TAXDIP RDFF'den daha iyi duyarlılık ve özgüllük gösterdi. TAXDIP'in duyarlılığı ME'den daha iyi olsa da, özgüllüğü ME'nin altında kaldı. Sonuç olarak, TAXDIP göstermiş olduğu performansla mevcut tahmin yöntemlerine uygun bir alternatiftir. Ayrıca, TAXDIP'in diğer tahmin yöntemleriyle örtüşen ve dahası onları tamamlayan doğru AAE tahminleri, onu birçok yöntemi bir araya getiren bir meta-AAE tahmin yönteminin parçası olma konusunda güçlü bir konuma getirmektedir.

Anahtar kelimeler: Protein altünite-altünite etkileşimleri, Protein birlikte evrimi, Protein fonksiyon analizi

1 Introduction

Proteins physically interact through domains located on them and these interactions are critical for their molecular functions and involvement in biological processes. Hence, identification of domain-domain interactions (DDIs) is a step towards the identification of protein-protein interactions (PPIs) that helps in understanding proteins' functional and/or structural roles. Several in vivo (e.g., yeast 2-hybrid) and in vitro (e.g., X-ray crystallography, protein microarrays) PPI identification, techniques exist, they require costly and time-consuming experiments. Compared to in vitro and in vivo techniques, computational (in silico) methods are fast and inexpensive alternatives.

Computational DDI and PPI prediction methods with various algorithmic approaches are available. Spinzak and Margalit developed the Association Method (AM) to predict DDIs using the frequency of observed protein interactions [1]. Alternatively, Deng and colleagues [2] proposed a method based on an expectation maximization (EM) algorithm and employs a maximum likelihood estimation (MLE) to find DDIs from protein interactions. Jothi et al. proposed a co-evolution based method, namely, Relative Co-evolution of Domain Pairs (RCDP) that uses sequence co-evolution to predict the domain pairs which are most likely to interact for a given PPI [3]. Gonzalez and Liao [4] employed a support vector machine (SVM) based computational method on interaction profile hidden Markov models (ipHMM) in order to predict DDIs. Furthermore, machine learning based algorithms (RDFF) [5], phylogenetic profiling [6] and probabilistic network models [7],[8] are also applied to predict DDI's. In addition to these individual prediction methods in the literature, there are meta-DDI/PPI prediction methods. For instance, Integrated DDI and protein interaction analysis system (IDDI) [9] combines several prediction methods in order to build a consensus prediction score.

The basis of the MirrorTree method is co-evolution; interacting proteins evolve simultaneously and likely to have similar phylogenetic trees. Briefly, the steps for MirrorTree-based

methods are 1) finding orthologs of two domains (or proteins) that are candidates for interaction, 2) computing multiple sequence alignments for orthologs from organisms common to both domains, 3) computing two similarity matrices (one for each domain) using pairwise distances between orthologs, and finally 4) compute correlation coefficient between these similarity matrices to assess co-evolution and predict DDI (or PPI). The MirrorTree-based methods are used extensively to predict DDI and PPIs [3],[10]-[16]. Craig and Liao [10] used support vector machines calculated from phylogenetic trees in order to predict PPIs. Gertz et al. [11] used the Metropolis Monte Carlo optimization algorithm to detect possible matches between two distance matrices. Goh and colleagues used co-evolution for quantifying the behavior of the evolutionary histories of protein ligands and their receptors [12],[13]. Jothi et al. [14] introduced a Monte Carlo search-based method to detect interacting protein pairs. Pazos and Valencia [15],[16] employed a MirrorTree based approach for finding physical interactions between protein pairs and identifying the most probable sequence regions for these interactions. Although the MirrorTree-based methods' predictive power position them as competent approaches for DDI prediction, they are sensitive to taxonomic diversity and evolutionary span of constructed matrices [17].

In this paper, we propose an improved MirrorTree-based DDI prediction method, namely Taxonomic Diversity-based Domain Interaction Prediction (TAXDIP), that uses representative homologs for taxonomic ranks higher than species (e.g., genus, family, etc.) and consequently ensuring higher evolutionary span within similarity matrices prior to computation of correlation coefficient. TAXDIP is based on MirrorTree method but provides a solution to taxonomic diversity and evolutionary span problem of this method by adding a new taxonomy-rank based sampling before computing similarity matrices. This extra step favor representation of higher-level taxonomic ranks (e.g. family) over lower-level taxonomic ranks' (e.g. species) in computing similarity matrices and, therefore, provides increased evolutionary span.

The rest of the paper is organized as follows. Section 2 describes the datasets used. Section 3 describes our method, namely Taxonomic Diversity-based Domain Interaction Prediction (TAXDIP), and steps towards its development including data preparation and experiments conducted to identify its parameters. Section 4 reports on the performance of our method and how it compares against available computational DDI prediction methods. Finally, Section 5 presents our conclusions.

2 Materials

2.1 Taxonomy dataset

The NCBI Taxonomy database [18] is one of the central sources that provides the species names and taxonomic lineage data for all known organisms. The NCBI Taxonomy database is downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/). Species, genus, family, order and class level taxa for about 1.49 million possible organisms are computed by traversing their lineages. These organisms represent 1.16 million species, 78.45 thousand genera, 8.30 thousand families, 381 orders and 304 classes.

2.2 Protein domain and multiple sequence alignment dataset

Pfam [19] is a protein family database which contains family annotations and multiple sequence alignments generated using Hidden Markov models. The Pfam release 30.0 is used and contains a total of 16,306 families that covers 17.7 million protein sequences. The multiple sequence alignments (MSA) for protein domains are downloaded using the Pfam website (http://pfam.xfam.org). The average number of members in a Pfam family for the release used is 2,667.

2.3 Protein domain-domain interaction dataset

DOMINE [20] is a database of DDI's among Pfam domains. The DDIs are either extracted from Protein Data Bank (PDB) entries or predicted by at least one of the several computational methods including ME [21], RCDP [3], P-value [8], Interdom [22], DPEA [23], PE [24], GPE [25], DIPD [26], RDFF [5], K-GIDDI [27], Insite [28], and DomainGA [29] and DIMA [6]. The latest version of DOMINE has 26,219 unique DDIs involving 5,410 unique active Pfam domains. The same DDI may be predicted by multiple methods. Table 1 provides the number of DDIs predicted by individual computational methods in DOMINE. A subset of DOMINE that contains 6,634 DDIs extracted from high-resolution 3D structures using iPfam [30] and 3did [31], is referred to as DOMINE gold set. The main reason for our use of DOMINE gold set is testing TAXDIP algorithm with experimentally verified true DDI's and avoiding, potentially, false DDI predictions made by other in silico methods.

Table 1: The number of domain-domain interactions in DOMINE predicted by individual computational methods.

Method	Number of DDIs
ME	2391
RCDP	960
P-value	596
Interdom	2768
DPEA	1812
PE	2588
GPE	1563
DIPD	2157
RDFF	2475
K-GIDDI	386
Insite	2408
DomainGA	459
DIMA	8012

3 Method

3.1 Creation of positive and negative sets

To assess the performance of our DDI prediction method, positive (interacting) and negative (non-interacting) domain pair sets are created. The positive set is based on the DOMINE gold set (see section 2.3) that contains 6,634 domain pairs. A total of 6,514 DDIs positive domain pairs are obtained after removal of 120 domain pairs due to missing data; in 81 pairs at least one of the Pfam domains are no longer in Pfam database and in 39 pairs Pfam multiple sequence alignments are not available. A negative set of equal size (6,514 DDIs) is generated using randomly selected Pfam domain pairs that have no known or predicted interactions reported in the DOMINE database.

3.2 Generation of similarity matrices

The similarity matrices for Pfam domain pairs in positive and negative sets are computed using the following steps:

i. For each Pfam domain pair, multiple sequence alignments are retrieved from Pfam web site,

ii. For each taxonomic rank from species to class (the taxonomic ranks above the class level are not examined since the number of taxa common to both domains becomes insignificant, e.g., 5 or less.),

   a. For the selected taxonomic rank, find the list of taxa common to the proteins found in both Pfam domain's MSAs from Step (i),

   b. Sample the intersecting taxa list from Step (a) using different sample sizes (10 to 100) and select one representative protein for each taxon. Figure 1 illustrates the impact of selection at different taxonomic ranks for one domain on an MSA-based phylogenetic tree,

iii. Compute similarity matrices based on pairwise distances between representative proteins from Step (b) using T-Coffee suite [32].

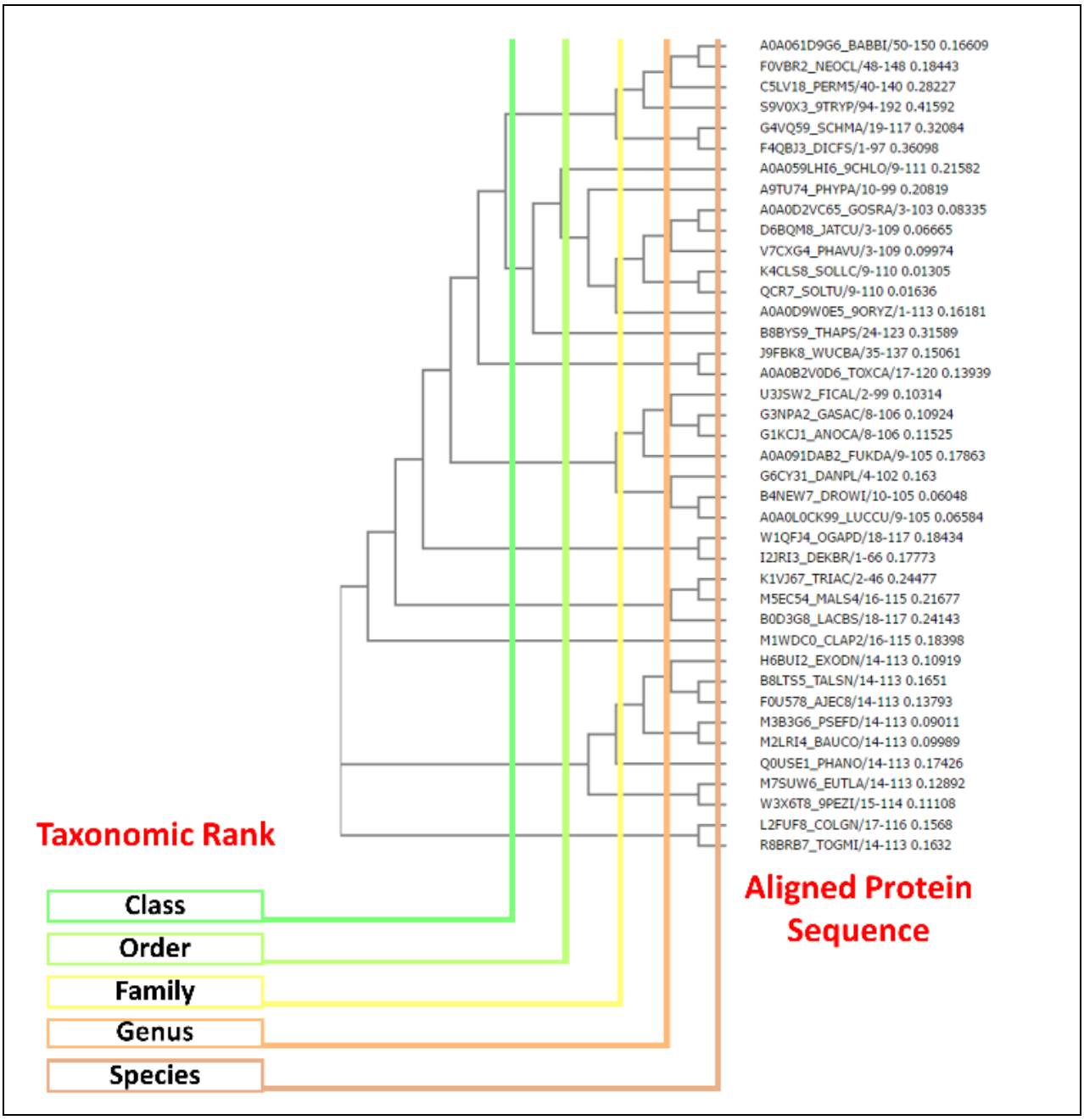


Figure 1: The impact of selection at different taxonomic ranks (species to class) for one domain illustrated on an MSA-based phylogenetic tree.

3.3 Computation of correlation coefficients

The co-evolution of domain pairs is evaluated based on Pearson's correlation coefficient [33] between the domains' similarity matrices (See 3.2). Pearson's correlation coefficient (Equation 1) is given as:

$$r_{XY} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(X_{ij}-\bar{X})(Y_{ij}-\bar{Y})}{\sqrt{\sum_{j=j+1}^{n-1}\sum_{j=j+1}^{n}(X_{ij}-\bar{X})^2\sum_{i=1}^{n-1}\sum_{j=j+1}^{n}(Y_{ij}-\bar{Y})^2}} \quad (1)$$

where n represents the size of similarity matrices, X_{ij} and Y_{ij} are the distances between proteins computed from MSAs. The value of r_{XY} ranges between -1.0 and +1.0. According to MirrorTree method, higher correlation coefficients indicate a higher level of co-evolution and, thus, an interaction between the domains. Figure 2 illustrates the similarity matrices and the correlation computed at the taxonomic rank of family for two Pfam domains PF00005 (ATP-binding domain of ABC transporters) and PF01302 (Cytoskeleton-associated Protein Glycine-rich domain).

3.4 Proposed taxonomic diversity-based domain interaction prediction (TAXDIP) algorithm

In the light of findings as reported in the Results section, TAXDIP algorithm is formulated as follows;

i. Given a Pfam domain pair, the MSA of sequences used to generate domains are retrieved,

ii. For each taxonomic rank from family to species (higher to lower evolutionary span),
   a. Find the taxa list common to the proteins containing the respective domains. If the intersection size is small (5 or less) proceed with the lower taxonomic rank,
   b. Sample the common taxa list (40 is used as sample size) in Step (a) and select one representative protein for each taxon,
   c. Generate two similarity matrices using proteins for both domains,
   d. Compute the Pearson's correlation coefficient between the similarity matrices.

iii. Use the threshold of 0.5 on the Pearson correlation coefficient computed on the highest taxonomic rank in Step (ii) to decide on DDI. The 0.5 threshold is based on the previous co-evolution based RCDP [3] method.

3.5 Performance evaluation

The performance of DDI predictions is computed based on a confusion matrix (Figure 3). There are four different outcomes for the DDI prediction. The number of correct DDI predictions on the positive and negative sets correspond to True and False Positives (TP and FP), respectively. Similarly, the number of incorrect DDI predictions correspond to True and False Negatives (TN and FN).

Table 2 shows the general structure of a confusion matrix and the common metrics that can be calculated from it. For comparison with other DDI prediction methods, mutually used metrics true positive rate (or sensitivity) and true negative rate (or specificity) are used.

The Matthews correlation coefficient (MCC) is another performance metric, given by Equation 2, which measures the quality of binary classifications [34]. MCC ranges between ±1.0 where +1, 0 and -1 correspond to the perfect prediction, random prediction, and disagreement between actual and predicted class respectively.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (2)$$

In addition, the area under the receiver operating characteristic (ROC) curve [35] (AUC); a commonly used method to reduce ROC performance to a single scalar value and compare classifier performances [36], is computed.

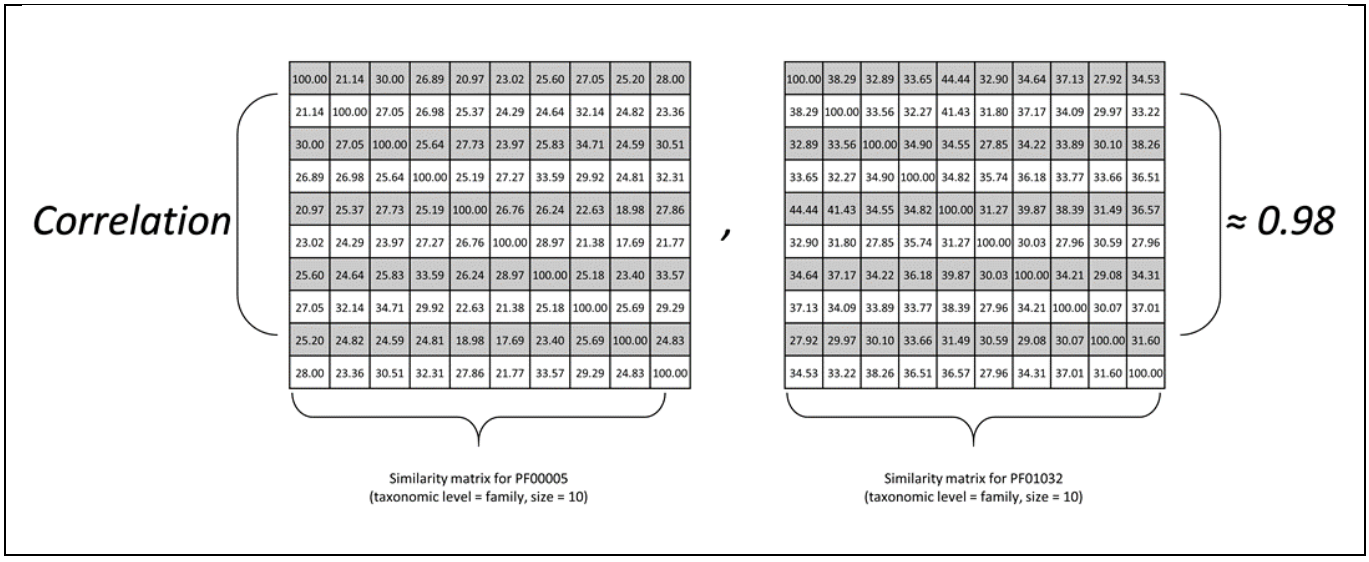
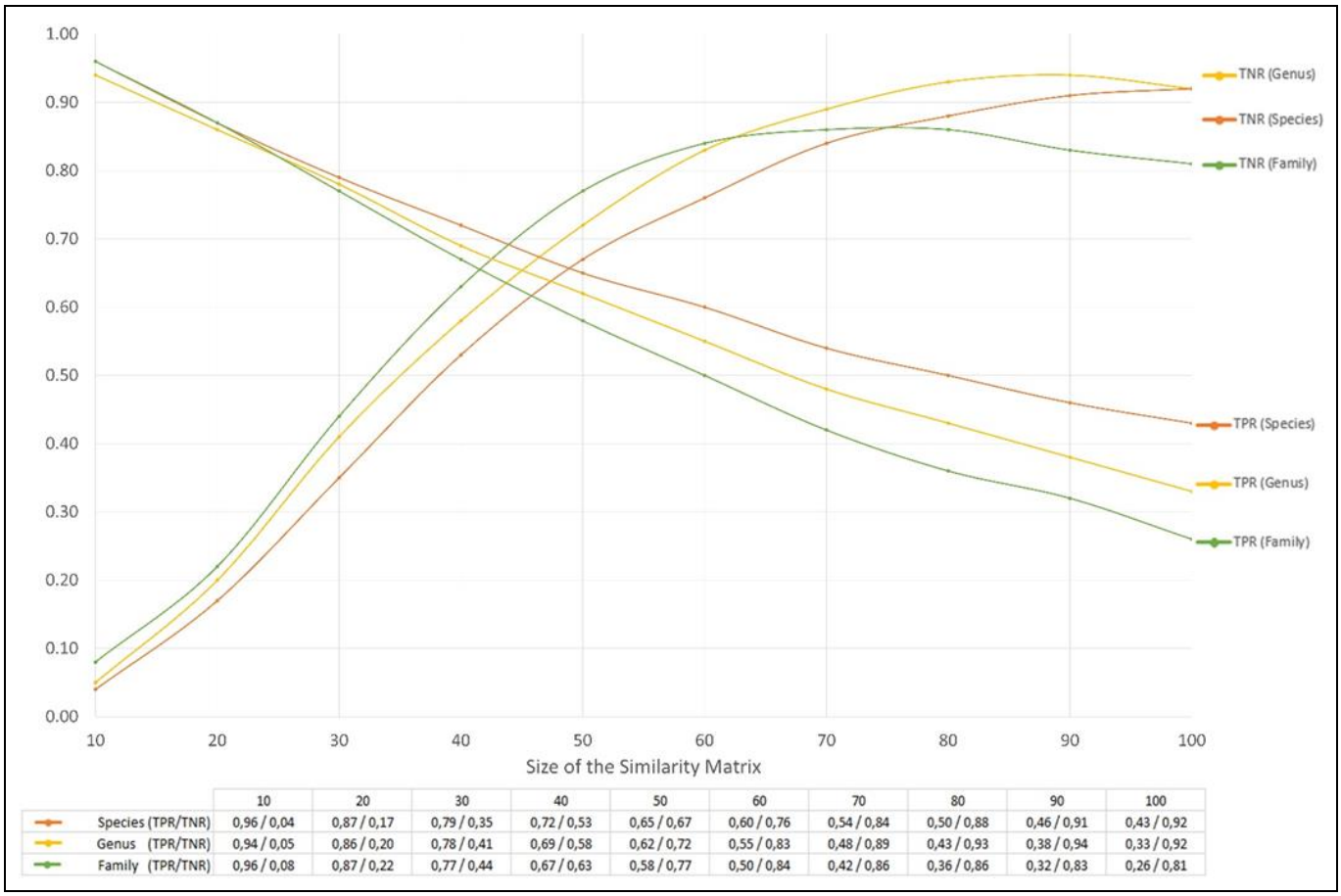


Figure 2: The similarity matrices computed at a family taxonomic rank level for two Pfam domains; PF00005 (ATP-binding domain of ABC transporters) and PF01302 (Cytoskeleton-associated Protein Glycine-rich domain), and their correlation. As per MirrorTree method, a high correlation coefficient of 0.98 is considered as an indication of co-evolution and interaction between these domains.



	10	20	30	40	50	60	70	80	90	100
Species (TPR/TNR)	0,96 / 0,04	0,87 / 0,17	0,79 / 0,35	0,72 / 0,53	0,65 / 0,67	0,60 / 0,76	0,54 / 0,84	0,50 / 0,88	0,46 / 0,91	0,43 / 0,92
Genus (TPR/TNR)	0,94 / 0,05	0,86 / 0,20	0,78 / 0,41	0,69 / 0,58	0,62 / 0,72	0,55 / 0,83	0,48 / 0,89	0,43 / 0,93	0,38 / 0,94	0,33 / 0,92
Family (TPR/TNR)	0,96 / 0,08	0,87 / 0,22	0,77 / 0,44	0,67 / 0,63	0,58 / 0,77	0,50 / 0,84	0,42 / 0,86	0,36 / 0,86	0,32 / 0,83	0,26 / 0,81

Figure 3: True positive and true negative rates for different taxonomic levels and sample sizes using 0.5 as the correlation coefficient threshold.

Table 2: Confusion matrix and common performance metrics computed based on this matrix.

Predicted class\Actual Class	Positive	Negative
Positive	True Positives (TP)	False Positives (FP)
Negative	False Negatives (FN)	True Negatives (TN)
Total	P	N

$$True\ Positive\ Rate\ (TPR) = TP/P$$
$$True\ Negative\ Rate\ (TNR) = TN/N$$
$$False\ Positive\ Rate\ (FPR) = FP/N$$
$$False\ Negative\ Rate\ (FNR) = FN/P$$
$$Precision = TP/(TP + FP)$$
$$Accuracy = (TP + TN)/(P + N)$$
$$F\ score = 2 \times Precision \times TPR/(Precision + TPR)$$

4 Results

4.1 Evaluation of taxonomic rank and sample size parameters on Mirrortree-based domain-domain interaction prediction

The effect of taxonomic rank and sample size (number of taxa sampled from common taxa list at selected taxonomic rank) on MirrorTree method is assessed using true positive rate (TPR) and true negative rate (TNR) metrics. Different correlation coefficient thresholds from 0.4 to 0.9 are tested. The best performance is obtained using a correlation coefficient of 0.5 and this corroborates with earlier MirrorTree method RCDP [3].

Figure 3 illustrates the true positive and negative rates for different taxonomy ranks and sample sizes using the correlation coefficient of 0.5. The MirrorTree method achieves best TPR/TNR ratios for sample sizes in 40-50 bracket. The best AUC value of 0.759 is for the taxonomy rank family and the sample size 40.

The motivation behind the proposed TAXDIP method (see section 3.4) is based on findings from the above results. In summary:

- A sample size corresponding to 10% of overall common taxa list size (mostly in 40-50 bracket) results in optimal DDI predictions
- The taxonomic diversity can change by domain. For instance, in some cases, the great majority of domain family members are from the same species. Furthermore, diversity becomes next to none above the taxonomic rank of family. Hence, although DDI predictions are often more accurate when higher taxonomic ranks (e.g. family) are used, using such ranks may not be possible due to the limited evolutionary span of the domain family.

4.2 Comparison of TAXDIP with other methods

Table 3 provides the number of DDIs in DOMINE gold set predicted by individual computational methods. ME [21] has the highest number of DDI predictions in DOMINE gold set as shown in Table 2 and better performance results on iPfam [30] and 3did [31] databases.

TAXDIP's TPR and TNR are computed as 71.00% and 63.00% based on DOMINE gold set. A wide range of TPR values (DPEA:23.63%, PE:29.63%, DIPD:29.76%, RCDP:52.13%, DIMA:54.00%, ME:55.00%, RDFF:79.78%) and TNR (ME:55.55%, RDFF:64.38%, DIMA:99.84%) have been reported in their respective manuscripts, RDFF [5] has the best reported TPR and TNR values(79.78%/64.38%).

Considering the real-life situation where the negatives far outnumber the positive interactions, a pseudo-real life situation is simulated by subsampling positive and negative DDI sets with a 1-to-10 ratio (100 positive and 1000 negative DDIs) from the complete DDI set and by assessing the performance of TAXDIP on this subset. These simulations are repeated 1000 times. TAXDIP shows rather consistent performance regardless of negative-bias, with mean TPR of 71.00% and TNR of 63.00% for all the simulations. Hence, the real-life bias towards "no interaction", is expected to have an insignificant impact on the overall performance of TAXDIP.

Table 3: The number of domain-domain interactions in DOMINE gold set predicted by individual computational methods.

Method	Number of DDIs
ME	1326
RCDP	144
P-value	63
Interdom	399
DPEA	247
PE	328
GPE	362
DIPD	588
RDFF	148
K-GIDDI	68
Insite	147
DomainGA	459
DIMA	106

Using the benchmark set with a total of 1,000 domain pairs (500 interacting, 500 non-interacting), TAXDIP's performance is compared with two existing DDI prediction methods considering two criteria. The first criterion is the reported performance results of these methods in their respective manuscripts [5],[21] and the second one is their coverage within or awareness of DOMINE gold set (Table 3). ME [21] is selected as it has the highest overlap with DOMINE gold set, and RDFF is with the best-reported TPR/TNR (79.78%/64.38%). The benchmark set contains DDIs common to sets used by TAXDIP, RDFF and ME. The DDI sets for RDFF and ME are obtained using their manuscripts. TPR, TNR, and Matthew's correlation coefficient (MCC) for all three methods' results are calculated. Table 4 provides the benchmark results of these methods.

Table 4: Benchmark results of TAXDIP, ME, and RDFF.

Method	TPR	TNR	MCC
TAXDIP	75.40%	52.51%	0.2867
ME	66.20%	90.60%	0.5878
RDFF	21.00%	49.00%	-0.3116

The DDI prediction algorithms typically used in combination to establish a level of confidence on predictions. Thus, to access how TAXDIP corroborates with other methods, the DOMINE gold set is used to identify overlap between TAXDIP's DDI predictions and other prediction methods contributed to DOMINE database, namely ME[21], RCDP [3], P-value [8], Interdom [22], DPEA [23], PE [24], GPE [25], DIPD [26], RDFF [5], K-GIDDI [27], Insite [28] and DomainGA [29] and DIMA [6]. Table 5 provides the percentage of DDIs predicted by both TAXDIP and another method among all the DOMINE gold set DDI predictions made this other method. A percentage of 100% indicates all the DDIs predicted by this other method is also predicted by TAXDIP while a percentage of 0% indicates the predictions made by TAXDIP complements these other methods predictions by 100%.

Table 5: The percentage of DDIs predicted by both TAXDIP and other prediction methods among all DDIs predicted by the respective method within DOMINE gold set.

Method	Percentage
ME	74.66%
RCDP	74.30%
P-value	66.66%
Interdom	71.00%
DPEA	71.42%
PE	75.30%
GPE	67.96%
DIPD	74.00%
RDFF	73.98%
K-GIDDI	79.41%
Insite	51.70%
DomainGA	67.65%
DIMA	80.19%

5 Discussion

Identification of interactions between proteins is a critical step to better understand proteins' roles in biological systems. Using computational methods to predict DDIs and, consequently, PPIs is a cost-effective and rapid way to complement experimental studies, especially, to identify and prioritize interacting protein candidates for experimental validation.

In this work, we proposed a new algorithm called TAXDIP that is based on the MirrorTree method. TAXDIP fixes the known problem of MirrorTree method's sensitivity to evolutionary span as previously reported [17] by introducing an effective taxonomic rank-based sampling prior to generation of similarity matrices and computation of correlation coefficients.

Based on the *reported* performance results of existing DDI prediction methods, TAXDIP predicts DDIs with better sensitivity/specificity (71.00%/63.00%) than almost any other method. The only exception to this is RDFF [5] that has a pretty small coverage within DOMINE gold set (148 DDIs). TAXDIP predictions show overlap with other prediction methods.

According to the benchmark set, we used to compare TAXDIP against RDFF and ME, TAXDIP outperformed RDFF with better sensitivity, specificity, and MCC score. The RDFF's significantly lower performance with respect to its manuscript is explainable by the choice of true set's difference from and limited coverage in DOMINE gold set (~2.3%). Although TAXDIP performed better than ME in terms of sensitivity, demonstrated weaker specificity. This is attributable to TAXDIP's preference towards the prediction of positive DDIs and ME's preference towards the prediction of negative DDIs.

In conclusion, TAXDIP, with its performance, is a viable alternative to existing prediction methods. Furthermore, given TAXDIP's true predictions, not only overlapping with but also complementing other DDI methods in DOMINE gold set, TAXDIP has a strong position in becoming part of a meta-DDI predictor, such as IDDI [9], that combines multiple methods to build a consensus prediction.

6 Availability

The source code of the scripts to compute the TAXDIP method is available to academic users 'as is' on request. Supplementary data associated with this article is available from http://eng1.mu.edu.tr/~eturk09/TAXDIP/.

7 References

[1] Sprinzak E, Altuvia Y, Margalit H. "Characterization and prediction of protein-protein interactions within and between complexes". *Proceedings of the National Academy of Sciences of the United States of America*, 103(40), 14718-14723, 2006.

[2] Deng M, Mehta S, Sun F, Chen T. "Inferring domain-domain interactions from protein-protein interactions". *Genome Research*, 12(10), 1540-8, 2002.

[3] Jothi R, Cherukuri PF, Tasneem A, Przytycka TM. "Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions". *Journal of Molecular Biology*, 362(4), 861-875, 2006.

[4] Gonzalez AJ, Liao L. "Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines". *BMC Bioinformatics*, 11(537), 2-14, 2010.

[5] Chen XW, Liu M. "Prediction of protein-protein interactions using random decision forest framework". *Bioinformatics*, 21(24), 4394-4400, 2005.

[6] Pagel P, Wong P, Frishman D. "A domain interaction map based on phylogenetic profiling". *Journal of Molecular Biology*, 344(5), 1331-46, 2004.

[7] Gomez SM, Rzhetsky A. "Towards the prediction of complete protein-protein interaction networks". *Pacific Symposium on Biocomputing Pacific Symposium on* Biocomputing, Kauai, Hawaii, 3-7 January 2002.

[8] Nye TM, Berzuini C, Gilks WR, Babu MM, Teichmann SA. "Statistical analysis of domains in interacting protein pairs". *Bioinformatics*, 21(7), 993-1001, 2005.

[9] Kim Y, Min B, Yi GS. "IDDI: Integrated domain-domain interaction and protein interaction analysis system". *Proteome Science*, 10, 1-9, 2012.

[10] Craig RA, Liao L. "Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices". *BMC Bioinformatics*, 8(6), 1-12, 2007.

[11] Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, Cokus S, et al. "Inferring protein interactions from phylogenetic distance matrices". *Bioinformatics*, 19(16), 2039-2045, 2003.

[12] Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. "Co-evolution of proteins with their interaction partners". *Journal of Molecular Biology*, 299(2), 283-293, 2000.

[13] Goh CS, Cohen FE. "Co-evolutionary analysis reveals insights into protein-protein interactions". *Journal of Molecular Biology*, 324(1), 177-92, 2002.

[14] Jothi R, Kann MG, Przytycka TM. "Predicting protein-protein interaction by searching evolutionary tree automorphism space". *Bioinformatics*, 21(1), 241-250, 2005.

[15] Pazos F, Valencia A. "Similarity of phylogenetic trees as indicator of protein-protein interaction". *Protein Engineering*, 14(9), 609-614, 2001.

[16] Pazos F, Valencia A. "In silico two-hybrid system for the selection of physically interacting protein pairs". *Proteins*, 47(2), 219-227, 2002.

[17] Zhou H, Jakobsson E. "Predicting protein-protein interaction by the mirrortree method: possibilities and limitations". *PLoS One*, 8(12), 81100, 2013.

[18] Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. "Database resources of the national center for biotechnology information". *Nucleic Acids Research*, 37, 5-15, 2009.

[19] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. "The Pfam protein families database: towards a more sustainable future". *Nucleic Acids Research*, 44(1), 279-285, 2016.

[20] Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. "DOMINE: a comprehensive collection of known and predicted domain-domain interactions". *Nucleic Acids Research*, 39, 730-735, 2011.

[21] Lee H, Deng M, Sun F, Chen T. "An integrated approach to the prediction of domain-domain interactions". *BMC Bioinformatics*, 25(7), 1-15, 2006.

[22] Ng SK, Zhang Z, Tan SH, Lin K. "InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes". *Nucleic Acids Research*, 31(1), 251-4, 2003.

[23] Riley R, Lee C, Sabatti C, Eisenberg D. "Inferring protein domain interactions from databases of interacting proteins". *Genome Biology*, 6(10), 1-17, 2005.

[24] Guimaraes KS, Jothi R, Zotenko E, Przytycka TM. "Predicting domain-domain interactions using a parsimony approach". *Genome Biology*, 7(11), 1-14, 2006.

[25] Guimaraes KS, Przytycka TM. "Interrogating domain-domain interactions with parsimony based approaches". *BMC Bioinformatics*, 9(171), 1-14, 2008.

[26] Zhao XM, Chen L, Aihara K. "A discriminative approach for identifying domain-domain interactions from protein-protein interactions". *Proteins*, 78(5), 1243-53, 2010.

[27] Liu M, Chen XW, Jothi R. "Knowledge-guided inference of domain-domain interactions from incomplete protein-protein interaction networks". *Bioinformatics*, 25(19), 2492-9, 2009.

[28] Wang H, Segal E, Ben-Hur A, Li QR, Vidal M, Koller D. "InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale". *Genome Biology*, 8(9), 1-18, 2007.

[29] Singhal M, Resat H. "A domain-based approach to predict protein-protein interactions". *BMC Bioinformatics*, 13(8), 1-19, 2007.

[30] Finn RD, Miller BL, Clements J, Bateman A. "iPfam: a database of protein family and domain interactions found in the Protein Data Bank". *Nucleic Acids Research*. 42(Database issue), 364-373, 2014.

[31] Mosca R, Ceol A, Stein A, Olivella R, Aloy P. "3did: a catalog of domain-based interactions of known three-dimensional structure". *Nucleic Acids Research*, 42(Database issue), 374-379, 2014.

[32] Notredame C, Higgins DG, Heringa J. "T-Coffee: A novel method for fast and accurate multiple sequence alignment". *Journal of Molecular Biology*, 302(1), 205-217, 2000.
[33] Pearson K. "Note on Regression and Inheritance in the Case of Two Parents". *Proceedings of the Royal Society of London*, 58, 240-242, 1895.
[34] Matthews BW. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochimica et Biophysica Acta*, 405(2), 442-451, 1975.
[35] Fawcett T. "An introduction to ROC analysis". *Pattern Recognition Letters*, 27(8), 861-874, 2006.
[36] Bradley AP. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". *Pattern Recognition*. 30(7), 1145-1159, 1997.