Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi
Pamukkale University Journal of Engineering Sciences

Kabul Edilmiş Araştırma Makalesi (Düzenlenmemiş Sürüm)

Accepted Research Article (Uncorrected Version)

Makale Başlığı / Title Taxonomic diversity-based domain interaction prediction

Taksonomik çeşitlilik tabanlı protein altünite etkileşim tahmini

Yazarlar / Authors Erdem TÜRK[1*], Barış Ethem SÜZEK[2]

Referans No / Reference No PAJES-18828

DOI 10.5505/pajes.2018.18828

[illegible]
Pamukkale University Journal of Engineering Sciences

[illegible]

[illegible]

Erdem T[illegible], Barış [illegible]

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

Received / [illegible]: 19.04.2018, Accepted / Kabul Tarihi: 05.11.2018
doi: 10.5505/pajes.2018.18828
Research Article / [illegible]

* Corresponding author / [illegible]

[illegible]

[illegible] ... bir araya getiren bir meta-[illegible]

Anahtar kelimeler: [illegible] birlikte evrimi, Protein fonksiyon analizi

Abstract

Identification of protein domain-domain interactions (DDIs) is an [illegible] MirrorTree is a DDI prediction method that is based on the principle [illegible] 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 [illegible] is better than ME, its specificity remained below ME. In conclusion, TAXDIP, with its performance, is a viable alternative to existing [illegible] 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.

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 [illegible]. 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 [illegible]. Furthermore, machine learning based algorithms (RDFF) [5], phylogenetic profiling [6] and probabilistic network models [7, 8] [illegible]. 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 coevolution 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-Â Á Ó Å Ä Í Å Ô È Ï Ä Ó õ Ð Ò Å Ä É Ã Ô É Ö Å 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-Ì Å Ö Å Ì Ô Á Ø Ï Î Ï Í É Ã Ò Á Î Ë Ó õ 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 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] É Ó Á Ä Á Ô Á Â Á Ó Å Ï Æ $ $) õ Ó Á Í 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 experimental, 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 Â Ï Ô È 0 Æ Á Í Ä ïroíméSÉep î (ñ.Ó - 3 ! Ó Æ

   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].

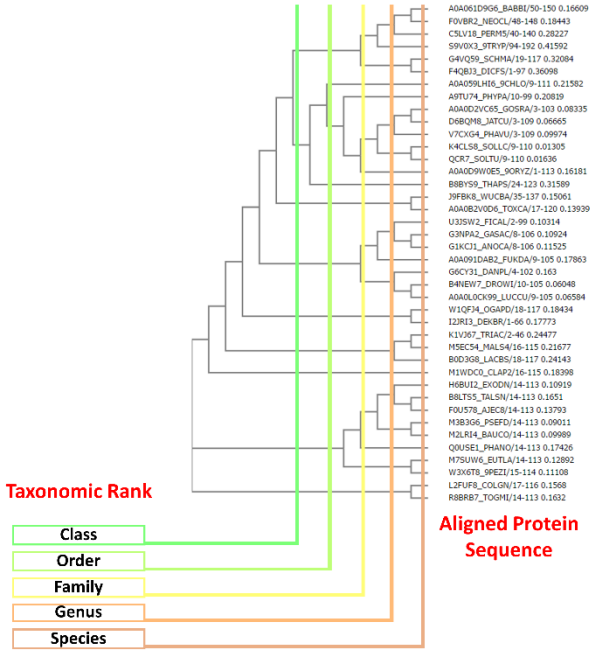

Figure 1: The impact of selection at different taxonomic ranks (species to class) for one domain illustrated on an MSA-based phylogenetic tree.

3.3 Computation of correlation coefficients

The co-evolution of domain pairs is evaluated by based 0 Å Á Ò Ó Ï Î õ Ó Ã Ï Ò Ò Å [38] Á ê ê ì Î the Ã Ä Ï Ï Å Í Æ Æ É É Í Õ É õ Å Î Ô Ó É Í É Ì Á Ò É Ô Ù Í Á Ô Ò É Ã Å Ó j 3 Å Å σ 8 ς Q 8 4 È Å 0 Å Á Ò Ó Ï Î õ Ó Ã Ï Ò Ò Å Ì Á Ô É Ï Î coefficient (Equation 1) is given as:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (1)$$

where n represents the size of similarity matrices, X_i 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 taxonomic rank of the family for two Pfam domains PF00005 (ATP-binding domain of ABC

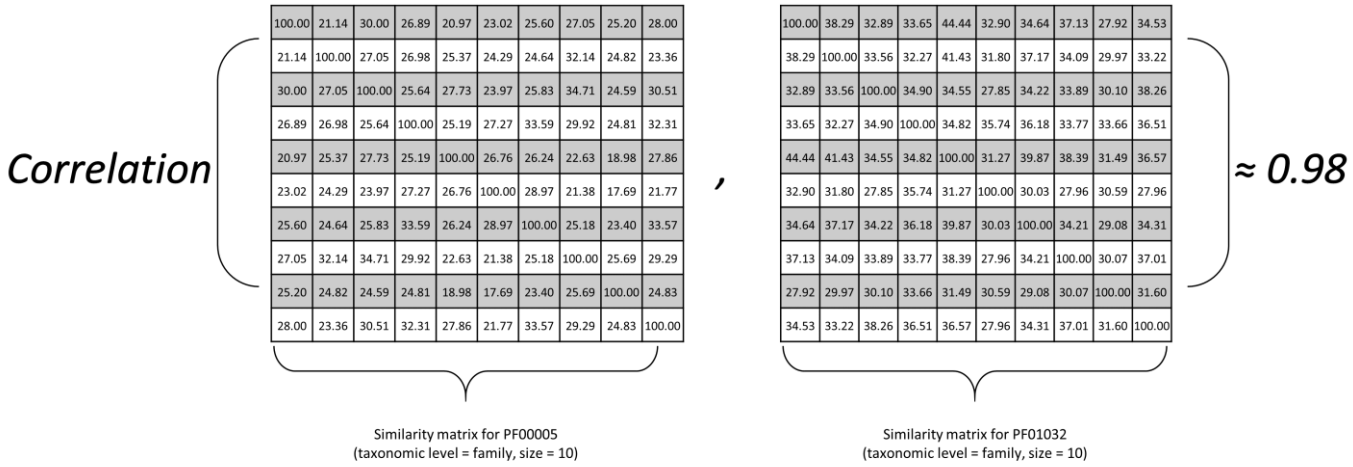

Figure 2: The similarity matrices computed at a family taxonomic rank level for two Pfam domains; PF00005 (ATP-binding domain of ABC transporters) and PF01302 (Cytoskeleton-associated Protein Glycine-rich domain), and their correlation. As per MirrorTree method, a high correlation coefficient of 0.98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

transporters) and PF01302 (Cytoskeleton-associated Protein Glycine-rich domain).

3.4 Proposed Taxonomic Diversity -based Domain Interacti on 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. # Ï Í Ð Õ Ô Å Ô È Å 0 Å Á Ò Ó Ï Î õ Ó Á Ì Ó Q Å Ì Á Ô É Pearson 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] 8 - # # Ò Á Î Ç Å Ó Â Å Ô where +1, 0 and -1 correspond to the perfect prediction, random prediction, and disagreement between actual and predicted class respectively.

$$\ddot{U}\ \ddot{o}\ \ddot{Q}_{\text{سس}} = \frac{\text{"Y Ü}_{\text{عي}}\text{"Y Ü طي "O Ü}_{\text{عي}}\text{"O Ü}}{\sqrt{\text{"Y Ü}_{\text{طي}}\text{"O Ü "Y Ü}_{\text{طي}}\text{"O Ü "Y Ü}_{\text{طي}}\text{"O Ü "Y Ü}_{\text{طي}}\text{"O Ü}}} \quad (2)$$

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.

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

"Y ì óÜ'Ωε í "Ω'ϑ"Ω ϑ"Ω Ü "Yسس"Y Üj Ü

"Y ì óÜ'Ωε ώ ϑ"Ω ϑ"Ω Ü "Yسس"Y Üj Ü

"O ώ ά Üí ΄Ω í "Ω'ϑ"Ω ϑ"Ω Ü "Yسس"O Üj Ü

"O ώ ά Üí ΄Ω'Ω ώ ϑ"Ω ϑ"Ω Ü "Yسس"O Üj Ü

Ü ì Ω ώ "Ωسس"Y Ωε "Y Ü طي "O Ü

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, the 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.

4 ! 8 $) 0 õ Ó 4 0 2 Á Î Ä 4 . 2 Á Ò Å Ã Ï Í Ð Õ Ô Å Ä Á Ó 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- Ì É Æ Å Â É Á Ó Ô Ï × Á Ò Ä Ó Ǒ Î Ï É Î Ô Å Ò Á Ã 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), 4 ! 8 $) 0 õ Ó Ð Å Ò Æ Ï 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 - Á Ô Ô È Å × õ Ó Ã Ï Ò Ò Ã Ï Å Æ Æ É Ã É Å Î Ô ȷ - # # Ɋ Æ Ï Ò Á Ì Ì Ô È Ò Å Å Table 4 provides the benchmark results of these methods.

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

Method	TPR	TNR	MCC
	χ ρ 8 π π Þ	Á Î Ä	φ σ 8 π π Þ
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 Ó Å Ô É Ó Õ Ó Å Ä Ô Ï É Ä Å Î Ô É Æ Ù Ï Ö Å Ò 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

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

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 Ô Ï Â Å Ô Ô Å Ò Õ Î Ä Å Ò Ó Ô Á Î Ä Ð Ò Ï Ô Å É Î Ó õ Ò 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 Ð Ò Ï Â Ì Å Í Ï Æ - É Ò s Ô ë Ò Ä Ô Å evolutionary Ï Ä õ 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 2 $ & & ignificantly