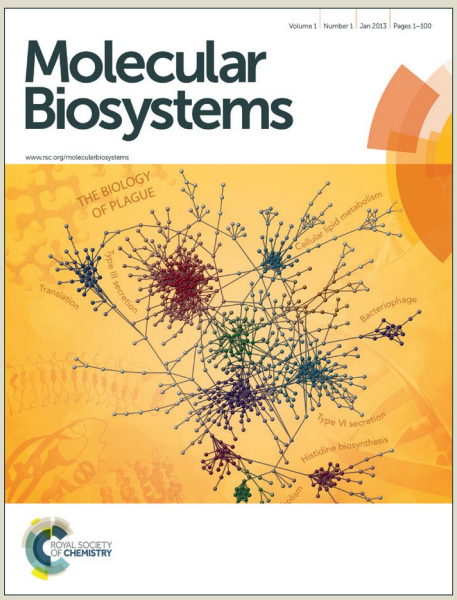
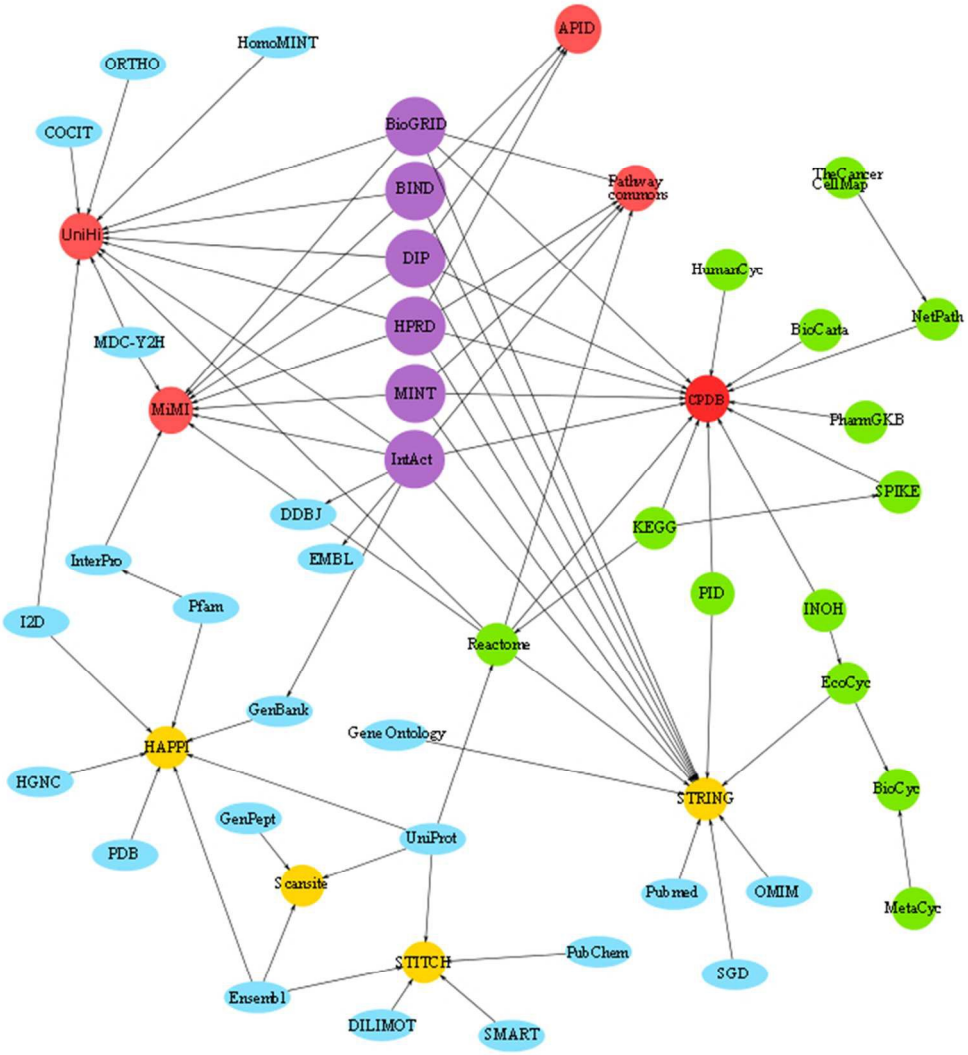
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



Molecular BioSystems

ARTICLE

Detecting reliable non interacting proteins (NIPs) significantly enhancing computational prediction of protein-protein interactions using machine learning methods

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

A.Srivastava[a*], G.Mazzocco[b,c*], A.Kel[d], L.S.Wyrwicz[a] and D.Plewczyński[b]

Abstract:

Protein–protein interactions (PPIs) play a vital role in most biological processes. Hence their comprehension can promote a better understanding of the mechanisms underlying living systems. However, besides the cost and the time limitation involved in the detection of experimentally validated PPIs, the noise in the data is still an important issue to overcome. In the last decade several *in silico* PPI prediction methods using both structural and genomic information were developed for this purpose. Here we introduce a unique validation approach aimed to collect reliable non interacting proteins (NIPs). Thereafter the most relevant protein/protein-pair related features were selected. Finally, the prepared dataset was used for PPI classification, leveraging the prediction capabilities of well-established machine learning methods. Our best classification procedure displayed specificity and sensitivity values of 96.33% and 98.02% respectively, surpassing the prediction cababilities of other methods, including those trained on gold standard datasets. We showed that the PPI/NIP predictive performances can be considerably improved by focusing on data preparation.

1. Introduction

Biological processes are typically regulated by the interactions of proteins with either other proteins or genetic material. Protein–protein interaction (PPI) is virtually involved in every process of a living system such as DNA replication, transcription, translation, cellular secretion, cell cycle regulation, signal transduction, metabolic regulation, etc. Hence the study of PPIs is crucial for the understanding of biological processes and has prompted the development of various experimental methods targeted to the identification of new PPIs. While the amount of genomic information continues to grow exponentially, the functional annotation of both proteins and their interactions is updated at a slower pace. Conventionally, both *in vitro* and *in vivo* methods are used to detect protein interactions. Tandem Affinity Purification (TAP)[1], Affinity chromatography[2,3] , co-immunoprecipitation (Co-IP)[4], X-ray crystallography[5], Nuclear Magnetic Resonance (NMR)[6] and yeast two-hybrid system (Y2H) are among the techniques most commonly applied. These experimental techniques have contributed to the generation of databases containing large sets of protein-protein interaction pairs, such as the Database of Interacting Proteins (DIP)[7], MIPS mammalian protein-protein interaction database (MIPS)[8], Biomolecular Interaction Network Database (BIND)[9], IntAct molecular interaction database (IntAct)[10] and the Molecular INTeraction database (MINT)[11]. High throughput techniques are labor intensive, expensive and time-consuming, especially when PPIs of complete species are considered. In the last decade several computational methodologies have been applied to the prediction of PPIs. The initial strategies included comparative analysis such as phylogenetic profiling of fused homologs into a single chain obtained from different organisms (Pellegrini et al., 1999)[12] or other gene fusion methods such as Rosetta stone (Enright et al. 1999)[13]. In 1998 the conserved gene neighborhood analysis of nine bacterial and archaeal genomes performed by Dandekar et al.[14] proved that the products of conserved genes are likely to interact. In 2001, Wuchty et al.[15] proposed the domain co-occurrence scale-free interaction network. These methods rely on information about protein functional domains, genes and functional pathways found in related species. Furthermore proteins' physico-chemical properties can be used to generate statistical models and train machine learning algorithms. In 2001 Bock et al.[16] successfully trained support vector machine (SVM) using both primary structure information and physico-chemical properties of proteins included in the Database of Interacting Proteins (DIP). Gomez et al. in 2003[17] described an attraction-repulsion model, in which the interaction between a protein pair is represented as the sum of attractive and repulsive forces associated with domain or motif-sized features. Several machine learning classifiers including Support Vector Machines (SVM)[18–20], Artificial Neural Network (ANN)[21], Naïve bias [22,23], K-Nearest neighbors[24,25], Decision Tree[26–28] and Random Forest[26,29] have being used to predict PPIs. Despite the popularity of PPI prediction methods, there are some limitations. The predictive performances of these methods can be negatively affected by the meager availability of information about non-interacting proteins. Secondly, the use of features with limited biological significance

[a] *Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland*
[b] *Centre of New Technologies, University of Warsaw, Warsaw, Poland*
[c] *Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland*
[d] *GeneXplain GmbH, Am Exer 10b, D-38302, Wolfenbüttel, Germany*

* Both authors contributed equally

may increase the rate of false positives, lowering the effectiveness of the methods.

In order to leverage the prediction capability of some of the machine learning approaches mentioned above, we propose a procedure for the preparation of both positive and negative protein-protein interaction data. This methodology pays particular attention to the biological relevance of the protein properties taken under consideration, while it tries to minimize the bias.

The proposed procedure can be divided into three phases. The first phase consisted in the acquisition of human PPI data from different data sources, the validation of such data and their coherent integration. The dataset of non-interacting proteins (NIP) was carefully prepared using a triple-layer validation inspired by previous works[30–32]. The aim of the second phase was to annotate both PPIs and NIPs using biologically relevant features from a protein-protein interaction perspective. The 43 protein features obtained from this analysis were grouped into four different types: Probabilistic modeling of inferred domain–domain interaction (DDI), Network analysis of PPIs, gene co-expression and amino acid information. The proposed approach present three major novelties:

1. A secondary PPI database was used without applying pre-filtering procedures in order to limit the amount of instilled bias.
2. The negative dataset of NIPs was developed considering only combinations of proteins found within the PPI database. A triple-layer of validation based on biologically relevant observations was then applied to select reliable NIP candidates. The NIP dataset represent itself a valid resource for future work in the field of PPI prediction.
3. A specific combination of biologically meaningful features encompassing different aspects of protein interactions (e.g. protein co-expression, properties of the protein interactome, domain-domain interactions, etc.) was carefully chosen before applying feature selection.

The machine learning methods trained on this data showed higher performances than the best prediction methods reported in literature. A schematic representation of the methodological workflow is presented in Figure 1.

2. Methods

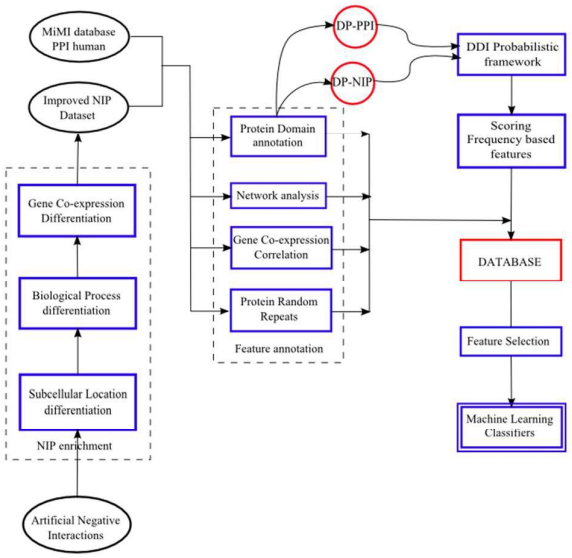
2.1 Protein-protein interaction data

More than 250 PPI primary and secondary databases (containing information from single or combined data sources respectively) are publicly available. We analyzed these databases and chose Michigan Molecular Interaction MiMI[33] database because of its versatility. In fact among all meta-databases it is the most non-discriminatory database, aiming to include all protein interaction data in one single comprehensive database[34]. The detailed relationship between experimentally validated PPI databases is depicted in Figure-A within supplementary materials. 157,863 binary interactions between human proteins were reported, including 80,965 PPIs involving 11,075 unique proteins. Several detection methods were used for the same protein interactions, resulting in data redundancy. The problem was solved including only unique PPIs in the positive dataset.

2.2. Non interacting protein data

Reliable NIP datasets[30] are difficult to obtain since predicting "absence of interaction" is not a trivial task to achieve. Moreover, the lack of positive interaction cannot be automatically considered as a negative interaction although this simplification has been sometimes used in previous works[28,35–37]. In order to deal with this problem, we applied a triple-layer validation to a raw dataset

Figure. 1. Schematic view of the methodological workflow



generated from all the possible 2-combinations of the 11,075 unique PPI proteins previously obtained. The self-interactions and known protein-protein interactions were removed and the validation procedure was then applied. While the first validation layer uses information about the subcellular localization of the proteins, the second relies upon knowledge about the proteins' involvement into specific biological processes. In both cases the Gene Ontology (GO) database was queried. Finally validation by gene co-expression correlation was applied. The common idea behind the refining procedure is that interacting protein pairs are more likely to be situated in the same biological contexts (e.g. subcellular localization, biological process or gene co-expression) and should be removed from the NIP dataset. This procedure can significantly reduce the number of false positives. The validation steps are described in the sections below.

2.3. Validation with Gene Ontology

The human gene association data included 374,356 GO annotations consisting of subcellular localization (SCL), molecular function (MF) and biological process (BP). Both BP and SCL annotations were cross-mapped with the PPI and NIP datasets. Protein pairs sharing any BP or SCL were removed from the NIP dataset. From the initial PPI dataset, 8,199 proteins included complete GO data, resulting in 21,021,863 NIPs without any shared SCL and 12,532,471 NIPs with one or more SCL in common. From the selected NIPs, 18,020,740 negative pairs were found to have dissimilar biological process ontologies. It should be noticed that a given protein may be located in multiple subcellular localizations and could be involved in several biological processes. A protein pair is considered to be non-

interacting only if their proteins don't share any common subcellular localizations nor biological processes. A Similar approach was adopted by Ben-Hur *et. al.*[30] and Xiao li *et al.* [31].

2.4. Validation with gene co-expression correlation

Gene co-expression data from COXPRESdb (version c4.1)[38] were downloaded and processed. The version used in this work contains 73,083 GeneChip experimental data. The directory included 20,280 files, each representing one gene. The database was parsed and the co-expression correlation values of 156,419,640 gene pairs were obtained. From the NIP dataset obtained after GO validation 16,254,117 protein pairs had gene co-expression information. In order to establish the difference between the distributions of the co-expression coefficients values of PPI and NIP datasets, random samples of equal size were obtained from PPI and NIP data respectively (refer the figure 2 where distribution in red is for PPIs and is in green for NIPs over Pearson's correlation coefficient values on x-axis and number of protein pair on y-axis). The Kolmogorov-Smirnoff (KS) test and the Welch two sample t-test (WT) were performed on these samples. The tests confirmed a statistical difference between the distributions (Figure 2) under a confidence value of 0.025. In figure 2, co-expression values for subset of PPI (in red) and equal subset of NIP (in blue). The obtained p-values were 2.43e-06 and 1.39e-04 respectively. The NIP pairs with a co-expression coefficient greater than 0.425 were removed from the NIP dataset because proteins with higher co-expression values are more likely to interact[39]. The co-expression coefficient cutoff mentioned above was chosen such that the NIPs/PPIs ratio was approximately 1/10. More detailed information can be found in table-A of supplementary material. Ultimately the validation procedure leaded to the selection of 13,523,822 NIPs. This data represent the negative dataset used in the machine learning procedure.

2.5. Probabilistic modeling of domain-domain Interaction

Information about confirmed protein domains were obtained from Pfam[40]. All the possible domain combinations were computed for each protein within a given pair. Scores based on the domain occurrences (e.g. frequency, probability standard score (*z-score*), enrichment scores etc.) were calculated for both PPIs and NIPs. A similar approach was previously proposed by Li 2006, Han et al. 2003 and Chatterjee et al. 2011[19,31,41] Since only a subset of the combinatorically-generated DDIs is truly interacting, a significant rate of false positives is foreseen. Nevertheless it would be logical to expect a higher rate of DP-DDI (domain Pair from PPI) with respect to DP-NIP (domain pair from NIP). In fact the probability of finding an interacting domain pair within an interacting protein pair is higher than finding it within a non-interacting one. The characteristics of DP-DDI and DP-NIP distributions are summarized in Table 1 in results section.

2.6. Network analysis

Protein interaction network analysis was also adopted in this study in order to assign protein functions[42]. A simple technique known as neighbor counting method predicts the function of unknown proteins, using the frequency of the closest neighbors' functions. This information can be statistically assessed[43]. The protein function prediction based on network mining is a novel approach in PPI studies[44,45]. In order to integrate the network properties in the

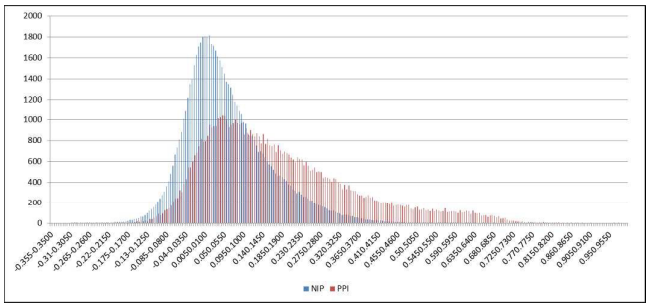


Figure 2. Frequency distribution NIPs Vs PPIs on co-expression coefficient

analysis, an interaction network was generated from PPI data. Instead of defining protein pairs as interacting or non-interacting entities, the network analysis defines them in terms of their degree of connectivity, contributing to provide an additional dimension to the data. The PPI network can be defined as a scale free network [46,47] where few nodes share many connections while the majority of the nodes have few connections. This makes the degree distribution of the PPI networks heavy-tailed[48,49]. A total number of 23 network properties were computed based on graph's betweenness, common edge, centrality eccentricity, neighbourhood degree, etc.

2.7. Feature assembly and selection

The selection of representative features is a crucial task in machine learning and has dramatic repercussions on the performances of such methods. Description of the feature extraction methods is given below.

2.7.1. DDI features

During the validation by DDIs described in section 2.5, a number of domain-related features were extracted. The features are formally introduced below.

Let a protein pair to be constituted by proteins a and b. Consider the set of domains $D_a = \{d_{a,i}: i = 1,2,\dots,n\}$ and $D_b = \{d_{b,j}: j = 1,2,\dots,m\}$ for proteins a and b respectively. The following properties are then defined.

a) ***Number of domains within each protein*** *(e.g. in the protein a)**:*** $|D_a|$ where $|D_a|$ is the cardinality of D_a.

b) ***Number of DDI combinations in a protein pair:*** $|D_{a,b}|$, where $D_{a,b} = \varphi(D_a, D_b) = \{\{d_{a,i}, d_{b,j}\} \mid d_{a,i} \in D_a \wedge d_{b,j} \in D_b\}$
$= \{\{d_{a,1}, d_{b,1}\}, \{d_{a,1}, d_{b,2}\}, \dots, \{d_{a,i}, d_{b,j}\}\}$.

c) ***Number of DDI combination in the whole dataset:*** Let P be the set of the domain sets of all the proteins within the PPI dataset. The total number of DDI pairs for all the $\binom{|P|}{2}$ possible 2-combinations of P within the PPI dataset can be defined as $|D_{PPI}|$, where

$$D_{PPI} = \cup_{x,y \in P} \varphi(D_x, D_y)$$

We can similarly define the total number of DDI pairs within NIP as

$$D_{NIP} = \cup_{x,y \in P \wedge x \neq y} \varphi(D_x, D_y)$$

d) ***Domain probability:*** Let consider a domain d_α, which appears in a certain number of domain pairs within the D_{PPI} set. Note that the domain definition is irrespective of the protein source. Then we can define a subset $D_{PPI,\alpha}$ containing all the D_{PPI} domain pairs that include the domain d_α. Formally:

$$D_{PPI,\alpha} = \{d \mid \forall\, d \in D_{PPI} : d_\alpha \in d\}$$

The occurrence probability of the domain d_α in D_{PPI} is therefore

$$\mathcal{P}_\alpha^{PPI} = \mathcal{P}(\mathfrak{D}^{PPI} = d_\alpha) = \frac{|D_{PPI,\alpha}|}{2|D_{PPI}|}$$

where $2|D_{PPI}|$ is the total number of domain occurrences and $\mathfrak{D}^{PPI}$ is the domain random variable for the PPI dataset. Similarly we can calculate $\mathcal{P}_\alpha^{NIP}$.

e) ***Domain pair probability:*** Let consider a domain pair $d_{\alpha,\beta}$ within the D_{PPI} set. Note that the domain pair definition is irrespective of the protein source. Then we can define a subset $D_{PPI,\alpha,\beta}$ as

$$D_{PPI,\alpha,\beta} = \{d_{\alpha,\beta} |\, d_{\alpha,\beta} \in D_{PPI}\}$$

containing all the D_{PPI} domain pairs $d_{\alpha,\beta}$. The occurrence probability of $d_{\alpha,\beta}$ in D_{PPI} is therefore

$$\mathcal{P}_{\alpha,\beta}^{PPI} = \mathcal{P}(\mathfrak{D}\mathcal{P}^{PPI} = d_{\alpha,\beta}) = \frac{|D_{PPI,\alpha,\beta}|}{|D_{PPI}|}$$

Where, $|D_{PPI}|$ is the total number of domain pairs and $\mathfrak{D}\mathcal{P}^{PPI}$ is a domain-pair random variable in the PPI dataset. Similarly we can calculate $\mathcal{P}_{\alpha,\beta}^{NIP}$. The probability of finding a highly interacting domain pairs is expected to be lower within the NIPs data.

Ideally,

$$\mathcal{P}_{\alpha,\beta}^{PPI} \propto \frac{1}{\mathcal{P}_{\alpha,\beta}^{NIP}}.$$

f) ***Enrichment score (ES):*** Ratio of the inferred domain pairs and expected number of pairs. The enrichment score for a domain pair ES_{PPI} within the PPI dataset can be defined as

$$ES_{PPI}(d_{\alpha,\beta}) = \frac{|D_{PPI,\alpha,\beta}|}{E_{PPI}[d_{\alpha,\beta}]}$$

Where the expected number of protein pairs is defined as:

$$E_{PPI}[d_{\alpha,\beta}] = \mathcal{P}_\alpha^{PPI} \cdot \mathcal{P}_\beta^{PPI} \cdot 2|D_{PPI}|$$

g) ***Standard probability score for domain pair:*** A numerical indicator similar to the z-score for domain pair $d_{\alpha,\beta}$ within PPI, can be defined as follow:

$$z_{PPI} = |D_{PPI,\alpha,\beta}| - \frac{E_{PPI}[d_{\alpha,\beta}]}{\sigma_{PPI}(d_{\alpha,\beta})}$$

Where

$$\sigma_{PPI}(d_{\alpha,\beta}) = \sqrt{2|D_{PPI}| \cdot [\mathcal{P}_\alpha^{PPI}(1 - \mathcal{P}_\alpha^{PPI}) + \mathcal{P}_\beta^{PPI}(1 - \mathcal{P}_\beta^{PPI})]}$$

The same metrics can be applied to NIP data.

2.7.2. Gene co-expression features

The following Co-expression coefficient features were attained from COXPRESdb[38].

a) Gene co-expression value: The Pearson's correlation coefficients between the expression values of gene pairs were considered.

b) Mutual Rank: Co-expressed gene networks in COXPRESdb are defined in terms of rank of correlation[52]. The correlation rank is not commutative for the proteins within a pair. Therefore the geometric average between two directional ranks (one for each protein) is used. This value is defined as Mutual Rank (MR).

2.7.3. Protein network features

A protein network was computed using the igraph package in R, representing proteins as vertices and edges as protein connections. The extracted network-based features are listed below.

a) Vertex properties: Network based properties such as betweenness, degree, closeness, eccentricity, neighbourhood degree, centrality closeness, and eigenvector centrality, were computed for each vertex (protein). These properties were extracted from either the 68,265 PPIs or the 8,519,279 NIPs.

b) Edge properties: The calculations based on comparative or combined values of any two vertices were also performed for the PPI and derived NIP protein pairs. The adjacency matrix of the 11,075 PPIs was built with the help of the R[53] package igraph. The eigenvalues of each vertex pair were computed. Similarity coefficients for each pair were calculated using three similarity metrics: (i) Jaccard similarity, (ii) Dice similarity and (iii) Inverse log-weighted similarity.

2.7.4. Protein disorder features

Protein disordered regions are known to be over represented in protein domains involved in binding[54]. This is especially visible in *hub proteins*, which interact with several different protein partners[55]. Hence, disorder related features were also taken into account. The percentage of disorder regions was calculated for each protein with the help of the tool ESpritz[56]. Applying the methods described above, 43 features were collected comprehensively. The complete list of these features is provided in Table B in the Supplementary material. Then two methods **Boruta**[57] and **Monte Carlo Feature Selection**[58] applied to identify the features which better describe the interactions between proteins. Network-based features, co-expression correlation coefficients (**CORs**) and the frequency of inferred DDI were among the most important attributes. On the contrary, the contribution of attributes based on disorder regions was negligible. This process allowed the selection of 17 protein features out of the initial 43. These data were further used to train the machine learning algorithms for PPI/NIP prediction. Please refer supplementary tables C-F for detailed information about the feature selection results mentioned above.

2.8 Classification methods

Machine learning methods were trained on multi-dimensional random samples of different sizes. The assortment of attributes was based on the feature selection outcome. The following classifiers were adopted: k-Nearest Neighbours (k-NN), Support Vector Machine (SVM) with Linear and Radial Base Function (RBF) kernels, Decision Tree (DT), Random forest (RF), Adapting Boosting (AdaBoost), Naïve Bayes (NB) and Linear Discriminant Analysis (LDA). A Multilayer Perceptron (MLP) implementation of artificial neural networks was also adopted. The samples were randomly obtained from both positive and negative datasets. Balanced samples consisting of 4000, 8000, 20000, 50000, and 100000 protein pairs were used in the training phase. In order to test the performances of the classifier under conditions closer to the real nature of the original data, two imbalanced samples were also considered. In these cases the PPI/NIP ratios were 4000/20000 and 4000/40000 respectively.

2.9 Tools

The relational database management Sqlite v.3.0 was used for data storage and management. The codebase infrastructure was implemented in Python v.2.7.8 Modules for data preparation and validation were implemented in Perl v.5.16.3, Python v.2.7.8 and R v.3.0.1. Statistical and machine learning analyses were performed in R (igraph library) and Python (including NumPy, Pandas, scikit-learn and scikit-neuralnetwork libraries). The classification parameters are included in the supplementary material.

3. Results and discussion

The enrichment process of artificially created non interacting protein pairs generated 13,523,822 NIPs of good quality, which were used as training set for computational prediction methods. Apart from Gene Ontology differentiation on the basis of exclusivity of location and biological process, two more properties namely gene Co-expression and domain-domain interaction probability were also successful to establish the difference between positive and negative interactions. The probabilistic approach developed, helped to reveal that inferred protein domain pairs in positive and negative datasets are dissimilar in terms of distribution and occurrence. The DDI features estimated in section 2.7.1 were formulated to assess the difference between interacting and not interacting domain-pairs belonging to the PPI and NIP datasets respectively. Approximately the 40% of DP-PPI were found to be absent in DP-NIP, despite the forty times larger DP-NIP. The dissimilarity between domain pair distributions from PPI and NIP was confirmed using Kolmogorov Smirnoff test and Welch t-test.

Table 1. Standard probability standard score (*z*) across DDI.

	DP-PPI	DP-NIP	Common DDIs in DP-PPI	Common DDIs in DP-NIP	Exclusive DP-PPI
Mean	0.530	0.050	0.644	-0.024	0.185
Median	0.092	0.011	0.102	-0.010	0.081
Variance	3.019	0.116	3.916	0.585	0.116
Std. dev.	1.737	0.340	1.979	0.765	0.340
Range of z-score	-1.4 to 15.3	-13.3 to 6.12	-1.4 to 15.31	-13.3 to 6.12	-0.09 to 2.92

The statistical distribution of probability standard score (z) is presented in table 1 which compares different DDI distributions. Out of 100,923 DP-PPI (Domain pairs inferred from PPI), 61,451 were found in the DP-NIP dataset also. The distributions of their respective probability standard scores were obtained showing a moderately negative correlation (Pearson's coefficient = -0.24). This observation is coherent with the fact that interacting domain pairs are more likely to be found within in the positive dataset. Moreover the missing 39,472 DP-PPI from DP-NIP displayed a higher standard probability score (z) with respect to the expected probability score for the whole DP-PPI set (Table 2). The intuitive hypothesis is that both probability and the standard score (*z*) of DDI pairs should be higher within PPIs then NIPs, and vice-versa. The average z-score within DP-NIP was significantly lower than the same score measured in the DP-PPI. The table 2 shows precision performance with top machine learning methods with different sample size of balanced class having equal number of PPI and NIP. These data are in accordance with the hypothesis above and were used as features in the classification phase. A similar approach was adopted by Deng et. al. in 2002[50] using maximum likelihood method. Han et al. in 2003[41] designed a probabilistic framework that considers domain combinations instead of single domains as basic units of protein interactions[51]. To the best of our knowledge these observations were not reported in previous works.

In each experiment we evaluated the classifiers computing both average and standard deviation of the performance measured under 10-fold cross validation. These values are presented in table 2. Both the *Random Forest* and *Decision Tree* methods performed surprisingly well, whereas the other methods performed similarly to the best approaches reported in literature. We also investigated the dependence of the prediction performances from both sample size and sample size ratio. In order to evaluate the effect of the sample size, we compared the performances of the five best methods. In the balanced data, an improvement in precision was noticed for samples of size 20,000 and above (compared in table 4). On the other hand a minor improvement of the recall values was registered in small samples.

Results from different PPI prediction studies are not easy to compare, since the properties of the datasets used, the protein coverage and the degree of data-reliability may vary in every case. Moreover most of the methods, including the one here reported, rely on specific protein features which are usually available only for a subset of known proteins. Taking this fact into account, a careful comparison with published results was made. The PPI_SVM[19] method, which leverage domain information to predict PPI, obtained 76% recall and 95% precision on a reliable subset of 3000 protein pairs from yeast consisting of balanced class. PreSPI[35] which involves a probabilistic approach to predict the interaction probability of proteins from yeast, achieved very similar values: 95% precision and 77% recall. In this experiment the size of interacting and non-interacting test groups was 1590 and 1490, respectively. Liu et al.[36] used sequence-based method predicting interactions for yeast proteins from DIP reporting 87% precision and 90% recall. In a recent attempt for a selected gold standard dataset, the Ensemble method [28] obtained 94% precision, 89% recall for human and 89% precision, 91% recall for yeast proteins. The values of precision and recall of our classifiers are comparable with the top results reported in the literature, summarized in table 4, a comparison plot is also presented in figure 3. The performances of Random Forest classifier were particularly brilliant, showing a precision 96.4% and recall of

Table 2a. Precision performance value with standard deviation (in %) for different test sample size in balanced class

Sample Size	Decision Tree	SD (+/-)	Random forest	SD (+/-)	Linear SVM	SD (+/-)	Nearest Neighbor	SD (+/-)	Naïve Bayes	SD (+/-)
4,000	94.83	1.44	95.15	1.54	92.93	3.64	90.69	4.30	91.05	5.02
8,000	94.92	0.78	95.34	1.01	92.64	3.74	90.21	4.53	90.83	4.75
20,000	96.37	1.05	96.40	0.80	93.15	4.41	91.12	4.45	91.63	4.53
50,000	96.56	0.21	96.61	0.23	93.38	4.29	91.23	4.41	91.66	4.36
100,000	96.25	0.24	96.47	0.31	93.34	4.14	91.25	4.24	91.71	4.37

Table 2b. Recall performance value with standard deviation (in %) for different test sample size in balanced class

Sample Size	Decision Tree	SD (+/-)	Random forest	SD (+/-)	Linear SVM	SD (+/-)	Nearest Neighbor	SD (+/-)	Naïve Bayes	SD (+/-)
4,000	97.84	1.15	97.34	1.39	85.58	14.38	82.26	11.27	79.35	14.92
8,000	97.14	1.19	97.67	1.07	86.44	13.80	82.71	10.98	79.86	14.68
20,000	97.62	0.61	98.02	0.65	87.93	12.57	84.52	10.04	81.11	14.62
50,000	98.14	0.32	98.47	0.43	89.05	11.89	85.52	9.60	81.94	14.35
100,000	98.74	0.17	98.91	0.18	89.62	11.66	86.17	9.42	82.42	14.47

Table 3. ML performance (in %) for sample with 50000 protein pairs in balanced class

Method used	Precision	SD (+/-) precision	Recall	SD (+/-) for Recall	AUC
Nearest Neighbors	91.23	4.83	85.52	11.23	88.40
Linear SVM	93.38	4.59	89.05	13.35	90.48
RBF SVM	92.78	4.12	86.72	12.26	90.17
Decision Tree	96.56	0.22	98.14	0.33	98.06
Random Forest	96.61	0.24	98.47	0.44	98.92
AdaBoost	92.22	4.93	87.66	11.32	99.77
Naive Bayes	91.66	4.76	81.94	17.51	80.81
LDA	91.17	4.65	80.22	12.02	88.24
MLP	87.28	0.02	85.13	03.21	95.55

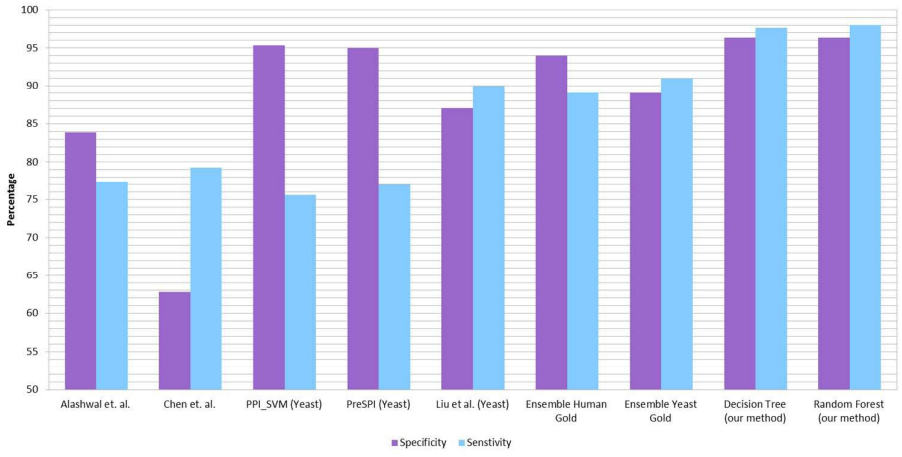


Figure 3: Histogram of comparison between protein - protein interaction prediction methods

Table 4: Comparison between the performances of previous methods and the here presented methods.

Previous methods	Specificity	Sensitivity
Alashwal et. al. [59]	83.90%	77.40%
Chen et. al. [60]	62.80%	79.30%
PPI_SVM (Yeast) [19]	95.35%	75.65%
PreSPI (Yeast) [35]	95.00%	77.00%
Liu et al. (Yeast) [36]	87.00%	90.00%
Ensemble[28]HumanGold	94.00%	89.00%
Ensemble[28] Yeast Gold	89.00%	91.00%
Our methods		
Nearest Neighbors	91.76%	84.52%
Linear SVM	93.53%	87.93%
RBF SVM	93.22%	85.51%
Decision Tree	**96.32%**	**97.62%**
Random Forest	**96.33%**	**98.02%**
AdaBoost	92.65%	86.89%
Naive Bayes	92.59%	81.11%
LDA	92.25%	79.51%
QDA	92.89%	76.83%

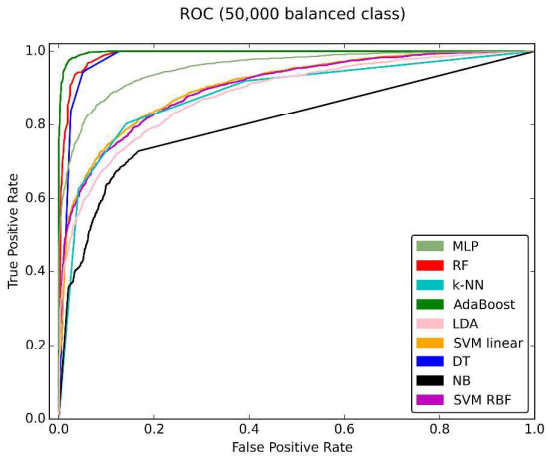


Figure 4 : ROC curves of the classification methods applied to the data sample of size 50,000 (balanced class). The respective AUC values are reported in table 3.

98.02 % using Random Forest. The risk of overfitting was handled applying 10-fold cross validation to large samples. The low standard deviation of the results (Table 3.) also allows to safely discarding that hypothesis. The Receiver Operating Characteristic (ROC) curves of the prediction methods are reported in Fig 3. The recall, precision and the ROC Area Under the Curve (AUC) values are present in table 3. We observed that decision-tree-based classifiers (decision three, random forest and AdaBoost with decision tree estimators) performed significantly better than the other methods. We investigated the effect that specific subset of features may play in this context. We noticed that DDI properties were capable of significantly improve the performances of decision-tree-based methods. A similar improvement was not observed in the other classifiers (see Figure 2 in supplementary materials).

4. Conclusion

In this study we introduced a novel protocol for the generation of reliable NIP data, which represent a crucial step in the enhancement of PPI prediction methods. We not only used most indiscriminate positive data, but our studies focused greatly upon the quality of negative data which is usually overlooked and many times wrongly used resulting in biased and inflated results. Biologically relevant protein features were selected. Each numerical feature was biologically reasonable and inspired the enrichment of positive as well as negative protein interaction data. The features derived from protein domains interaction, protein interaction network and gene expression allowed us to improve the quality of PPI prediction detection in a cumulative manner. The dataset was then used to train and test several machine learning classifiers. Some of the standard classifiers trained on the mentioned dataset, outperformed all the previously existing methods without losing robustness of the results. Algorithms based on decision trees gave excellent performances on these data, while maintaining low variance. The results obtained are undoubtedly associated with the diligent preparation of the NIPs data. The biological plausibility of the data was in fact taken into careful consideration during every step of the data preparation. Moreover, our definition of protein interactions relied on the combination of independent biological aspects and resulted to be beneficial in the machine learning training phase. We believe that our methodology represents which currently represent an issue in the prediction of protein-protein interaction. Besides the classification perspective, knowledge of NIP could be useful for the identification of least interacting proteins within a pathway. Being potentially indispensable factors in a pathological process, such proteins may represent valid targets for drug design. Despite exhibiting good performances, our method still has scope of improvement. Physicochemical properties like solvent accessibility and hydrophobicity could be included in order to improve the classification results. The established methodology could be extended to proteome other organisms.

ACKNOWLEDGEMENTS

AS was supported by MPD international program, Foundation for Polish Science, grant number MPD/2009/5/styp11. LSW was supported by National Science Centre HARMONIA 3, grant number. DEC-2012/06/M/NZ2/00112. DP, GM were supported by the Polish National Science Centre, grant numbers 2014/15/B/ST6/05082 and 2013/09/B/NZ2/00121.

References

1 A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer and G. Superti-Furga, *Nature*, 2002, **415**, 141–7.

2 M. Uhlén, *Biotechniques*, 2008, **44**, 649–54.

3 T. Berggard, S. Linse and P. James, *Proteomics*, 2007, **7**, 2833–2842.

4 A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer and G. Superti-Furga, *Nature*, 2002, **415**, 141–7.

5 P. Y. Chou and G. D. Fasman, *Biochemistry*, 1974, **13**, 222–45.

6 M. R. O'Connell, R. Gamsjaeger and J. P. Mackay, *Proteomics*, 2009, **9**, 5224–32.

7 I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S.-M. Kim and D. Eisenberg, *Nucleic Acids Res.*, 2002, **30**, 303–5.

8 H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, M. Münsterkötter, P. Pagel, N. Strack, V. Stümpflen, J. Warfsmann and a Ruepp, *Nucleic Acids Res.*, 2004, **32**, D41–4.

9 G. D. Bader and C. W. Hogue, *Bioinformatics*, 2000, **16**, 465–77.

10 H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman and R. Apweiler, *Nucleic Acids Res.*, 2004, **32**, D452–5.

11 A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli and G. Cesareni, *Nucleic Acids Res.*, 2007, **35**, D572–4.

12 E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg, *Science*, 1999, **285**, 751–3.

13 a J. Enright, I. Iliopoulos, N. C. Kyrpides and C. a Ouzounis, *Nature*, 1999, **402**, 86–90.

14 T. Dandekar, B. Snel, M. Huynen and P. Bork, *Trends Biochem. Sci.*, 1998, **23**, 324–8.

15 S. Wuchty, *Mol. Biol. Evol.*, 2001, **18**, 1694–702.

16 J. R. Bock and D. A. Gough, *Bioinformatics*, 2001, **17**, 455–460.

17 S. M. Gomez, W. S. Noble and A. Rzhetsky, *Bioinformatics*, 2003, **19**, 1875–81.

18 M. Rashid, S. Ramasamy and G. P. S. Raghava, *Curr. Protein Pept. Sci.*, 2010, 11, 589–600.

19 P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri and D. Plewczynski, *Cell. Mol. Biol. Lett.*, 2011, **16**, 264–78.

20 S. Dohkan, A. Koike and T. Takagi, *In Silico Biol.*, 2006, **6**, 515–29.

21 P. Fariselli, F. Pazos, A. Valencia and R. Casadio, *Eur. J. Biochem.*, 2002, **269**, 1356–61.

22 X. Lin and X. Chen, *Proteomics*, 2013, **13**, 261–8.

23 H. S. Najafabadi and R. Salavati, *Genome Biol.*, 2008, **9**, R87.

24 F. Browne, H. Wang, H. Zheng and F. Azuaje, in *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, IEEE, 2007, pp. 1365–1369.

25 L. Li, L. Jing and D. Huang, in *2009 International Conference on Natural Language Processing and Knowledge Engineering*, IEEE, 2009, pp. 1–7.

26 X.-W. Chen and M. Liu, *Bioinformatics*, 2005, **21**, 4394–400.

27 G. T. Valente, M. L. Acencio, C. Martins and N. Lemke, *PLoS One*, 2013, **8**, e65587.

28 I. Saha, J. Zubek, T. Klingström, S. Forsberg, J. Wikander, M. Kierczak, U. Maulik and D. Plewczynski, *Mol. Biosyst.*, 2014, **10**, 820–30.

29 Y. Qi, J. Klein-Seetharaman and Z. Bar-Joseph, *Pac. Symp. Biocomput.*, 2005, 531–42.

30 A. Ben-Hur and W. S. Noble, *BMC Bioinformatics*, 2006, **7 Suppl 1**, S2.

31 X.-L. Li, S.-H. Tan and S.-K. Ng, *Int. J. Data Min. Bioinform.*, 2006, **1**, 138–149.

32 Y. Park and E. M. Marcotte, *Bioinformatics*, 2011, **27**, 3024–8.

33 M. Jayapandian, A. Chapman, V. G. Tarcea, C. Yu, A. Elkiss, A. Ianni, B. Liu, A. Nandi, C. Santos, P. Andrews, B. Athey, D. States and H. V Jagadish, *Nucleic Acids Res.*, 2007, **35**, D566–71.

34 T. Klingström and D. Plewczynski, *Brief. Bioinform.*, 2011, **12**, 702–13.

35 D.-S. Han, H.-S. Kim, W.-H. Jang, S.-D. Lee and J.-K. Suh, *Nucleic Acids Res.*, 2004, **32**, 6312–20.

36 H. Liu, *Third Int. Symp. Optim. Syst. Biol.*, 2009, 198–206.

37 R. Singh, D. Park, J. Xu, R. Hosur and B. Berger, *Nucleic Acids Res.*, 2010, **38**, W508–15.

38 T. Obayashi, Y. Okamura, S. Ito, S. Tadaka, I. N. Motoike and K. Kinoshita, *Nucleic Acids Res.*, 2013, **41**, D1014–20.

39 R. Jansen, D. Greenbaum and M. Gerstein, *Genome Res.*, 2002, **12**, 37–46.

40 M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn, *Nucleic Acids Res.*, 2012, **40**, D290–301.

41 D. Han, H.-S. Kim, J. Seo and W. Jang, *Genome Inform.*, 2003, **14**, 250–259.

42 D. S. Goldberg and F. P. Roth, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 4372–6.

43 K. S. Ahmed, N. H. Saloma and Y. M. Kadah, *Theor. Biol. Med. Model.*, 2011, **8**, 11.

44 J. Chen, W. Hsu, M. L. Lee and S.-K. Ng, *Bioinformatics*, 2006, **22**, 1998–2004.

45 P.-Y. Chen, C. M. Deane and G. Reinert, *PLoS Comput. Biol.*, 2008, **4**, e1000118.

46 A.-L. Barabási, *Science*, 2009, **325**, 412–3.

47 A. Barabasi and R. Albert, *Science*, 1999, **286**, 509–12.

48 A. Wagner, *Mol. Biol. Evol.*, 2001, **18**, 1283–92.

49 H. Jeong, S. P. Mason, A. L. Barabási and Z. N. Oltvai, *Nature*, 2001, **411**, 41–2.

50 M. Deng, S. Mehta, F. Sun and T. Chen, *Genome Res.*, 2002, **12**, 1540–8.

51 D.-S. H. D.-S. Han, H.-S. K. H.-S. Kim, W.-H. J. W.-H. Jang and S.-D. L. S.-D. Lee, *Proceedings. Fourth IEEE Symp. Bioinforma. Bioeng.*, 2004.

52 T. Obayashi and K. Kinoshita, *DNA Res.*, 2009, **16**, 249–60.

53 W. Han, J. Lee and J. X. Yu, *PVLDB*, 2010, **3**, 449–459.

54 R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed and P. E. Wright, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 11504–9.

55 Z. Dosztányi, J. Chen, A. K. Dunker, I. Simon, P. Tompa and Z. Dosztanyi, *J. Proteome Res.*, 2006, **5**, 2985–2995.

56 I. Walsh, A. J. M. Martin, T. Di Domenico and S. C. E. Tosatto, *Bioinformatics*, 2012, **28**, 503–9.

57 M. B. Kursa and W. R. Rudnicki, *J. Stat. Softw.*, 2010, 1–13.

58 M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki and J. Komorowski, *Bioinformatics*, 2008, **24**, 110–7.

59 H. Alashwal, S. Deris and R. M. Othman, *World Acad. Sci. Eng. Technol.*, 2009, **51**, 785–790.

60 X.-W. Chen and M. Liu, *EURASIP J. Adv. Signal Process.*, 2006, **2006**, 1–9.