Boldi *et al. BMC Bioinformatics* 2018, **19**(Suppl 14):417
https://doi.org/10.1186/s12859-018-2385-x

BMC Bioinformatics

RESEARCH Open Access

Evaluating the impact of topological protein features on the negative examples selection

Paolo Boldi[†], Marco Frasca[*†] and Dario Malchiodi

From 5th International Work-Conference on Bioinformatics and Biomedical Engineering
Granada, Spain. 26-28 April 2017

Abstract

Background: Supervised machine learning methods when applied to the problem of automated protein-function prediction (*AFP*) require the availability of both positive examples (i.e., proteins which are known to possess a given protein function) and negative examples (corresponding to proteins not associated with that function). Unfortunately, publicly available proteome and genome data sources such as the Gene Ontology rarely store the functions *not* possessed by a protein. Thus the *negative selection*, consisting in identifying informative negative examples, is currently a central and challenging problem in *AFP*. Several heuristics have been proposed through the years to solve this problem; nevertheless, despite their effectiveness, to the best of our knowledge no previous existing work studied which protein features are more relevant to this task, that is, which protein features help more in discriminating reliable and unreliable negatives.

Results: The present work analyses the impact of several features on the selection of negative proteins for the Gene Ontology (GO) terms. The analysis is network-based: it exploits the fact that proteins can be naturally structured in a network, considering the pairwise relationships coming from several sources of data, such as protein-protein and genetic interactions. Overall, the proposed protein features, including local and global graph centrality measures and protein multifunctionality, can be *term-aware* (i.e., depending on the GO term) and *term-unaware* (i.e., invariant across the GO terms). We validated the informativeness of each feature utilizing a temporal holdout in three different experiments on yeast, mouse and human proteomes: (i) feature selection to detect which protein features are more helpful for the negative selection; (ii) protein function prediction to verify whether the features considered are also useful to predict GO terms; (iii) negative selection by applying two different negative selection algorithms on proteins represented through the proposed features.

Conclusions: Term-aware features (with some exceptions) resulted more informative for problem (i), together with node *betweenness*, which is the most relevant among term-unaware features. The node *positive neighborhood* instead is the most predictive feature for the *AFP* problem, while experiment (iii) showed that the proposed features allow negative selection algorithms to select effectively negative instances in the temporal holdout setting, with better results when nonlinear combinations of features are also exploited.

Keywords: Negative example selection, Protein function prediction, Biological networks, Protein features

*Correspondence: frasca@di.unimi.it
[†]Paolo Boldi and Marco Frasca contributed equally to this work.
Department of Computer Science, Università degli Studi di Milano, Via Comelico 39, 20135 Milano, Italy

Background

The publicly available databases devoted to record protein functions (for instance, the Functional Catalogue [1] and the Gene Ontology [2]) typically contain entries associating a protein with the biological functions the protein is known to possess. On the other hand, these repositories rarely consider *not possessed* functions. Thus, if a protein is not associated with a function, this could be simply due to a lack of information. Indeed, in such cases it is not possible to exclude that future studies could in principle associate that protein with that function.

Among the available protein function taxonomies, this work considers the Gene Ontology (GO), a hierarchy composed of three branches, biological process (BP), molecular function (MF), and cellular component (CC), each structured as a direct acyclic graph [2]. The functions described in this ontology (referred to as *GO terms*) are often (positively) annotated solely to a small number of proteins. Therefore, remaining proteins either do *not possess* the function, or correspond to not yet discovered positive annotations.

This observation leads to a central and critical issue in the problem of automated protein-function prediction (*AFP*), consisting in discovering novel associations of proteins with biological functions through computational methodologies. Indeed, the automated prediction process is typically based on *supervised*/*semi-supervised* machine learning techniques, requiring both positive and negative associations of proteins with functions (technically referred to as *positive* and *negative examples*, respectively) in order to infer accurate predictors. In this context, selecting the negative examples is a central issue for *AFP* [3–5]. The methods proposed in the literature to tackle the negative selection problem typically rely on bagging (bootstrap aggregating) techniques, based on the repeated inference of binary classifiers discriminating positive examples from reliable subsets of non-positive examples. These subsets are obtained through random subsampling on non-positive examples [6], either being guided by specific positive-negative similarity measures [7–9], or simply subsampling the items under the assumption that the probability to get a false positive be sufficiently small [10]. In addition, some heuristics have been proposed specifically for the *AFP* context, negatively associating a term with all proteins positive for sibling and/or ancestral GO terms [11], or computing the empirical conditional probability of a term given the annotations for other terms in the three GO branches, considering all nodes [4] or only the hierarchy leaves [12].

To our knowledge no researches have tried to investigate the possible relations between suitable 'protein features' and the fact that a protein can be considered as a reliable negative example. That is, before applying any algorithm to learn negative examples, it is of paramount importance studying which 'protein representation' is more informative for the problem itself. In this context, most information sources about the relationships between proteins are naturally represented through protein networks, where each node represents a protein and an edge the relationship between two proteins [13]; additionally, most approaches proposed for *AFP* are network-based [14–20]. Thus, the purpose here is twofold: extracting meaningful protein features from protein networks, and assessing their ability to improve the identification of good negative examples.

By extending the study presented in [21], this paper proposes a set of 14 features, ranging from protein multifunctional properties, to local and global graph centrality measures, including weighted degree, betweenness, and clustering centrality. Such features have been divided in the *term-aware* and *term-unaware* subsets, referring respectively to features varying with the GO term under study and to features independent of GO terms. With a dedicated experiment, the significance of each feature for selecting negatives has been assessed by adopting a state-of-the-art feature selection algorithm, along with a temporal holdout setting, necessary to determine the category of proteins not reliable as negative examples (that is, those that received a novel annotation in the holdout period). Through the paper this category is denoted by C_{np} (the **c**ategory of **n**egative proteins that become **p**ositive). As further validation, in another experiment the proposed features have been provided as input to two procedures for learning negatives, evaluating their ability in detecting proteins not in C_{np}. In the above mentioned analyses we also tested 3 probabilistic features computed by *3Prop*, a state-of-the-art method to extract features from biological networks; the results of *3Prop* have already been tested against the *AFP* problem [22], but their use within the negative selection has not been investigated yet. Finally, another experiment has been set up to predict the GO protein functions, to get more insight about the information encoded in the 14 proposed features. Overall, our paper extends the research done in [21] by adding 8 novel features, by performing feature selection on temporal holdout data, by applying linear and nonlinear state-of-the-art methods to learn negative examples, and by constructing extended and updated datasets for three organisms (yeast, mouse, and human).

Our studies showed that the set of features informative for identifying negative examples depends on both the organism and the GO branch considered. As a trend common to different settings, term-aware features tend to be selected more frequently, especially *Positive neighborhood* and *Mean of positive neighborhood*. Term-unaware features, however, play an important role, with some differences among organisms: *Neighborhood*

mean and *Weighted clustering coefficient* are more frequently selected in yeast, whereas *Betweenness* is largely more informative in mouse and human. The most predictive feature for the *AFP* problem is *Positive neighborhood*: indeed, when representing proteins by eliminating just this feature, the highest decrease in performance is observed. When providing the proposed protein representation as input to negative selection algorithms, our 14 features allow linear methods to achieve the lowest number of false negatives (that is, proteins in C_{np} classified as reliable negatives), which on the contrary increases when adding *3Prop* features to the representation, or when representing proteins just using *3Prop*. Finally, when using nonlinear methods to learn negatives, the number of false negatives largely decreases, and it is nearly the same when adopting the proposed features and *3Prop*; this phenomenon is likely due to novel information coming from nonlinear interactions of the *3Prop* features that linear methods are not able to exploit.

The paper is organized as follows: a first section describes the adopted methodology, including data description, the proposed features, and the setting of the different experiments carried out. The second section reports the obtained results and the related discussion, while some concluding remarks close the paper.

Methods

This section aims at describing the data sources leveraged in order to construct protein networks and protein functional annotations, the protein features extracted, and their experimental validation. Three different experiments have been performed to validate the adopted features:

- assessing feature relevance,
- predicting protein functions,
- selecting reliable negative proteins.

Each of the above mentioned steps is described in detail in the following sections.

Data

The input networks have been retrieved from the STRING database, version 10.0 [23], for the following organisms: *S.cerevisiae* (yeast), *Mus musculus* (mouse) and *Homo sapiens* (human). The STRING network already merges several sources of data, including protein homology relationships from different species, thus resulting in a highly informative network. Connections in such a network are endowed with a "combined score" that represents how reliable that relation should be considered; as suggested by STRING curators, connections with a combined score lower than 700 (combined scores range from 1 to 999) were filtered out. The network topological characteristics are reported in Table 1. All input networks have one large and some smaller connected components. The total number of nodes does not include nodes that became isolated after edge thresholding. Networks have been normalized as described in the next section.

Functional annotations for STRING proteins have been downloaded from the Gene Ontology, by considering two different temporal releases: the UniProt GOA releases 69 (9 May 2017) and 40 (25 November 2014) for yeast, releases 155 (6 June 2017) and 125 (25 November 2014) for mouse, and releases 168 (9 May 2017) and 139 (25 November 2014) for human. The two releases form a 'temporal holdout': the older release is used for the training phase, and the later release allows to evaluate the quality of predictions. In both releases, solely experimentally validated annotations have been considered. The relevance assessment of node/protein features to detect reliable negatives was focused on proteins which received at least a new annotation during the temporal holdout period (for a given GO term); we denote by C_{np} this category of proteins. Then we selected the GO terms with at least 20 proteins in C_{np}, obtaining the terms summarized in Table 2. The proposed features were also tested in terms of their capability in predicting the protein functions, by selecting GO terms with 20–200 annotations in the later release, in order to have a minimum of information to train a classifier, and to exclude terms with a large number of annotations, because they are too generic [13, 24, 25]. The total number of obtained GO terms is shown in Table 3.

Preliminaries

Protein networks are represented as an undirected graph $G\langle V, \boldsymbol{W}\rangle$, with $V = \{1, \ldots, n\}$ denoting the set of nodes/proteins and $\boldsymbol{W}$ being a $n \times n$ matrix whose entries

Table 1 Description of data networks

Organism	Nodes	Average degree	Components	Component size	Diameter	Weighted diameter
Yeast	5586	38.4740	41	5483, 2–7	12	3.0481
Mouse	13921	59.9990	190	13417, 2–10	13	3.1857
Human	15154	47.5572	89	14951, 2–10	11	3.1552

Column `Components` denotes the number of connected components in the network, whereas `Component size` denotes the corresponding number of nodes. `Diameter` is the number of edges on the longest path between two nodes, without considering edge weights

Table 2 Number of GO terms in the three GO branches for which $|C_{np}| \geq 20$. $|C_{np}|$ denotes the cardinality of C_{np} (i.e., the number of negative proteins that become positive in the temporal period)

Organism	CC	MF	BP
Yeast	5	9	29
Mouse	62	75	512
Human	71	105	363

$W_{ij} \in [0,1]$ encode some notion of intra-protein functional similarity (with $W_{ij} = 0$ when the corresponding nodes are not connected). The matrix $\boldsymbol{W}$ is obtained from the STRING connections $\hat{\boldsymbol{W}}$ after the following normalization, which preserves the connection symmetry:

$$\boldsymbol{W} = \boldsymbol{D}^{-1/2}\hat{\boldsymbol{W}}\boldsymbol{D}^{-1/2}$$

where $\boldsymbol{D}$ is a diagonal matrix with non-null elements $d_{ii} = \sum_j \hat{W}_{ij}$. The temporal holdout validation scheme relies on two additional matrices $\boldsymbol{Y}, \overline{\boldsymbol{Y}} \in \{0,1\}^{n\times m}$ containing the annotations of proteins to m GO terms $\{1,\dots m\}$: each matrix refers to a different temporal release of the ontology (assuming $\boldsymbol{Y}$ as the older one). If we denote the r-th column and the i-th row of a matrix $\boldsymbol{X}$ by $\boldsymbol{X}_{\cdot r}$ and $\boldsymbol{X}_{i\cdot}$ respectively, then $\boldsymbol{Y}_{\cdot k}$ and $\overline{\boldsymbol{Y}}_{\cdot k}$ describe the annotations for the GO term k to the proteins in V at the beginning and at the end of the houldout period. Moreover, $N_i := \{j \in V | W_{ij} \neq 0\}$ denotes the neighborhood of node $i \in V$, and for a given GO term k, $N_i^+ := \{j \in N_i | Y_{jk} = 1\}$ denotes its positive neighborhood, that is the subset of the neighborhood composed only of nodes positively annotated for k (Here, as in many of the following notations, the index k of the GO term is left implicit).

We recall that fixed a term k, $C_{np} \subseteq V$ is the set of proteins that received a new annotation in the holdout period, that is, $C_{np} = \{i \in V | Y_{ik} = 0 \wedge \overline{Y}_{ik} = 1\}$.

As mentioned in the previous section, the main aim of this paper is extracting features from nodes in G which effectively discriminate proteins belonging to C_{np} from proteins negatively annotated in both releases, as shown in the next section.

Extracting proteins features

The protein features studied in this work are selected in order to consider on the one hand information about the network topology, including both local and global 'standard' node centrality measures, on the other hand information about protein annotations. The resulting set of protein features is shown in Table 4.

A first group of features depends only on the structure of the network G: some of them are purely *local*, in the sense that they exploit a limited local neighborhood around the protein of interest ($f1$–$f4$); other features are more *global* in nature ($f6$–$f9$), and correspond broadly to some of the most common parameter-free centrality measures in network analysis. A second group of features, besides using the network structure, takes also into consideration the annotations ($f5$) or refers to the term-aware variant of some of the features of the first two groups ($f10$–$f14$). The considered features are summarized in Table 4 and described here below.

Table 3 Number GO terms with 20–200 annotated proteins in the more recent release

Organism	CC	MF	BP
Yeast	9	18	41
Mouse	18	32	178
Human	41	64	153

Table 4 The considered features for node $i \in V$ and GO term k

Symbol	Name	Definition
$f1$	Neighborhood mean	$\frac{1}{\lvert N_i\rvert}\sum_{j\in N_i} W_{ij}$
$f2$	Neighborhood variance	$\frac{1}{\lvert N_i\rvert-1}\sum_{j\in N_i}(W_{ij} - f1(i))^2$
$f3$	Weighted degree	$\sum_{j\in N_i} W_{ij}$
$f4$	Weighted clustering coefficient	$\sum_{jj'\in N_i, j'\in N_j} \frac{W_{ij}+W_{ij'}+W_{jj'}}{3} \Big/ \sum_{jj'\in N_i} \frac{W_{ij}+W_{ij'}}{2}$
$f5$	Number of annotations	$\sum_{h=1}^{m} Y_{ih}$
$f6$	Closeness centrality [33]	$\frac{1}{\sum_{j\in C_i} d_{ij}}$
$f7$	Lin's index [34]	$\frac{\lvert C_i\rvert^2}{\sum_{j\in C_i} d_{ij}}$
$f8$	Harmonic centrality [32]	$\sum_{j\in C_i} \frac{1}{d_{ij}}$
$f9$	Betweenness [35, 36]	$\sum_{s,t\in C_i, s\neq i, t\neq i, s\neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$
$f10$	Positive neighborhood	$\sum_{j\in N_i^+} W_{ij}$
$f11$	Mean of positive neighborhood	$\frac{1}{\lvert N_i^+\rvert}\sum_{j\in N_i^+} W_{ij}$
$f12$	Positive closeness centrality	$\frac{1}{\sum_{j\in C_i^+} d_{ij}}$
$f13$	Positive Lin's index	$\frac{\lvert C_i^+\rvert^2}{\sum_{j\in C_i^+} d_{ij}}$
$f14$	Positive harmonic centrality	$\sum_{j\in C_i^+} \frac{1}{d_{ij}}$
$f15$	1-step Random Walk	$\boldsymbol{P}_{i\cdot}\boldsymbol{y}$
$f16$	2-step Random Walk	$\boldsymbol{P}^2_{i\cdot}\boldsymbol{y}$
$f17$	3-step Random Walk	$\boldsymbol{P}^3_{i\cdot}\boldsymbol{y}$

C_i denotes the connected component of i, C_i^+ the positive nodes in C_i, d_{st} the shortest-path distance from s to t (using $\boldsymbol{W}$ as weight matrix), σ_{st} the number of shortest paths from s to t, and $\sigma_{st}(u)$ the number of such paths that include u as internal node. $\boldsymbol{P}$ and $\boldsymbol{y}$ are defined in (1) and (2)

$f1$ *Neighborhood mean*: mean of connection weights in the protein neighborhood.

$f2$ *Neighborhood variance*: variance of connection weights in the protein neighborhood.

$f3$ *Weighted degree*: sum of connection weights in the protein neighborhood.

$f4$ *Weighted clustering coefficient*: weighted proportion of triplets centered in the protein of interest that turn out to be closed (i.e. triangles).

$f5$ *Number of annotations*: number of GO terms for which the protein is annotated in the older release.

$f6$ *Closeness centrality*: reciprocal of the sum of shortest-path distances from the protein to all the other proteins in the same connected component.

$f7$ *Lin's index*: an adjusted version of closeness, obtained multiplying it by the square of the size of the component.

$f8$ *Harmonic centrality*: sum of the reciprocal of all the shortest-path distances from the protein to all the other reachable proteins.

$f9$ *Betweenness*: sum of the fractions of shortest paths that pass through the given protein.

$f10$ *Positive neighborhood*: sum of connection weights in the protein positive neighborhood.

$f11$ *Mean of positive neighborhood*: mean of connection weights in the protein positive neighborhood.

$f12$ *Positive closeness centrality:* reciprocal of the sum of shortest-path distances from the protein to all the positive proteins in the same connected component.

$f13$ *Positive Lin's index*: an adjusted version of positive closeness, obtained multiplying it by the square of the number of positive proteins in the same connected component.

$f14$ *Positive harmonic centrality*: sum of the reciprocal of all the shortest-path distances from the protein to all the positive reachable proteins.

The first two features refer to the first moments of the distribution of connection weights in the neighborhood of a node. The third feature provides information about the node connectivity, and moreover has been suggested in the literature as a proxy for gene multifunctionality [26, 27]. Jointly considering the first and third feature conveys information about the number of connections, one of the main measures for the connectivity of nodes in graphs along with the weighted degree [28].

Measure $f4$ is the weighted-aware version of the local clustering coefficient [29] of the node under consideration: for each triplet centered in the node, we compute its average weight (the average weight of the two or three edges involved). The ratio between the total weight of closed triples and total weight of all triples gives the local clustering coefficient; this quantity coincides with the standard (local) clustering coefficient when all the weights coincide. It is a variant of the weighted version proposed in [30] that takes into full account all the three weights appearing in the closed triples.

Feature $f5$ is related to the ability of a protein to play different roles: in its computation, the current GO term has been excluded in order to not introduce bias.

Features $f6$–$f8$ are among the most classical geometric centrality measures. As many authors observe [31, 32], closeness centrality $f6$ [33] (essentially, up to a constant, the reciprocal of the average distance between the node under consideration and the other nodes in its component) provides biased results in presence of disconnected components with largely different sizes; Lin's index $f7$ [34] and harmonic centrality $f8$ [32] both try to mitigate this big-in-Japan effect in different ways (one by explicitly taking the size of the connected component into account, and the other by looking implicitly at the distance from all nodes, using harmonic average instead of arithmetic average—where infinite distances give a null contribution). Another quite classical centrality measure is betweenness $f8$, originally defined by Anthonisse for edges [35] and then adapted by Freeman to nodes [36]; this index measures robustness rather than centrality (it is related to the probability that shortest-path routing fails when the node is deleted).

Feature $f10$ instead exploits both the number of positive neighbours and the corresponding weight magnitudes, and it plays the role of a *guilt-by-association* score [16]. Together with $f10$, feature $f11$ describes the number of connections toward positive nodes. Overall, features $f1$–$f9$ are *term-unaware*, in the sense that they do not need the annotation vector $\boldsymbol{Y}_{\cdot k}$ to be computed (for a given term k). A special case is represented by the feature $f5$, which uses an information not directly related to the annotations for the current GO term, but encompassing GO terms; hence, we did not include it in the group of term-aware features. Conversely, features $f10$ and $f11$ are the term-aware versions of $f1$ and $f3$, respectively, and similarly $f12$–$f14$ are the term-aware equivalent of $f6$–$f8$.

In order to have comparable ranges, features have been normalized so as to sum up to one across proteins, that is $\sum_{i=1}^{n} fk(i) = 1$, for each $k \in \{1, 2, \ldots 14\}$.

It is worth pointing out that the centrality measures considered here do not cover all the indices examined in the literature [32], and in particular they do not include any spectral index: measures such as PageRank or Katz's index were avoided in order to exclude the proliferation of parameters whose tuning would increase the chance of overfitting. Other spectral indices (such as Seeley's index) do not apply to

disconnected networks. For similar reasons, the consideration of scale-aware measures [37] was left as future work.

To further enrich our analysis, the state-of-the-art *3Prop* features have also been considered: originally proposed to extract features from biological networks, this algorithm describes a protein $i \in V$ with three features p_i^j, $j = 1, 2, 3$, each representing the probability that a random walk (respectively of length 1, 2 and 3) which starts from a positive (annotated) node ends in i [22]. Namely, fixed a GO term k, recalling the previously introduced diagonal matrix $\boldsymbol{D}$, and setting

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{W}, \qquad \boldsymbol{y} = (y_1, \ldots, y_n), \tag{1}$$

with components

$$y_i = \begin{cases} \frac{1}{\sum_h Y_{hk}} & \text{if } i \text{ is annotated for } k, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

it holds that $p_i^j = \boldsymbol{P}_{i.}^j \boldsymbol{y}$ for $j \in \{1, 2, 3\}$ (where of course $\boldsymbol{P}^2 = \boldsymbol{P}\boldsymbol{P}$, and $\boldsymbol{P}^3 = \boldsymbol{P}^2\boldsymbol{P}$). Features p_i^1, p_i^2 and p_i^3 ($f15$–$f17$ in Table 4) are thereby included in the group of term-aware features.

Assessing feature relevance

To evaluate the efficacy of node features $f1$–$f17$ described in Table 4 in detecting reliable negative proteins (i.e., those neither annotated in the first release nor in the second one), a binary classification problem was established. In this problem, proteins are represented through the extracted features, and their label is provided by the class C_{np} for GO terms in Table 2. The aim is selecting the features which mostly improve the classification performance. As classifier we adopted the *CART* algorithm [38], combined with the *Sequential floating forward Search* (SFFS) method [39] to determine the optimal subset of features. We employed the SFFS algorithm to capture the combined effect of multiple features; due to the potentially large number of add/remove steps until convergence, SFFS requires an efficient classifier, such as *CART*, which in addition is able to exploit feature interactions. To prevent selection bias and overfitting, data were partitioned into three non-overlapping subsets (following the setting proposed in [25]). On each subset a triple-loop of 3-fold cross-validation (CV) has been executed, using training data to select the classifier model (through the inner CV loop). Such model has been used on the corresponding test fold to validate the current subset of features.

In order to deal with the scarcity of positive instances, the F_1 measure was selected as performance criterion to be maximized both in the inner and the outer CV loops. Finally, we ranked features through the proportion of times they have been selected in all the experiments over the three data subsets.

Predicting protein functions

The proposed features have also been tested in classifying node/proteins to the Gene Ontology terms, to assess their capability in capturing network structures useful for the automated protein-function prediction. For each term previously described and summarized in Table 3, a binary classification problem was set up, with proteins represented in turn through features $f1$–$f14$, $f15$–$f17$, and $f1$–$f17$. In order not to have any bias toward a specific classifier, two state-of-the-art methods were used to solve the binary classification problems where instances were represented through feature vectors: *linear support vector machines* (SVM) with class weights [40] and *Random Forests* (RF) [41]. The performance has been evaluated using a 3-fold outer loop CV, and a 3-fold inner loop CV to select the parameters C and *mtry*, respectively for SVM and RF models. To counterbalance the large presence of negative examples and to avoid learning trivial models, the class weights of the SVM for term k have been set to 1 and $\frac{n-\sum_{i=1}^n Y_{ik}}{\sum_{i=1}^n Y_{ik}}$ for the negative and the positive class, respectively, as suggested in [42]. The F_1 measure has been adopted both to select the model and to measure the classification performance (averaged across folds), since this measure is more informative when positive instances are rare. Furthermore, results are also reported in terms of *Precision* (proportion of annotated proteins among those classified as positive) and *Recall* (proportion of annotated proteins that were positively classified).

Selecting reliable negative proteins

In order to study further the subsets of features that help detecting reliable negatives, the features described in Table 4 were supplied as input to negative selection algorithms, to investigate the relevance of the following different combinations of features:

- $f15$–$f17$,
- $f1$–$f14$ - top q,
- $f1$–$f14$ - mean,
- $f1$–$f14$,
- $f1$–$f17$ - top q,
- $f1$–$f17$ - mean,
- $f1$–$f17$,

where 'top q' denotes the selection of the top $q = 5$ features in the corresponding ranking in Fig. 1 and Additional file 1, and 'mean' denotes the selection of the features having a frequence larger than the mean frequency value (black dashed horizontal line in the figures). The choice $q = 5$ derives from the observation of frequency distributions: in some cases just three features

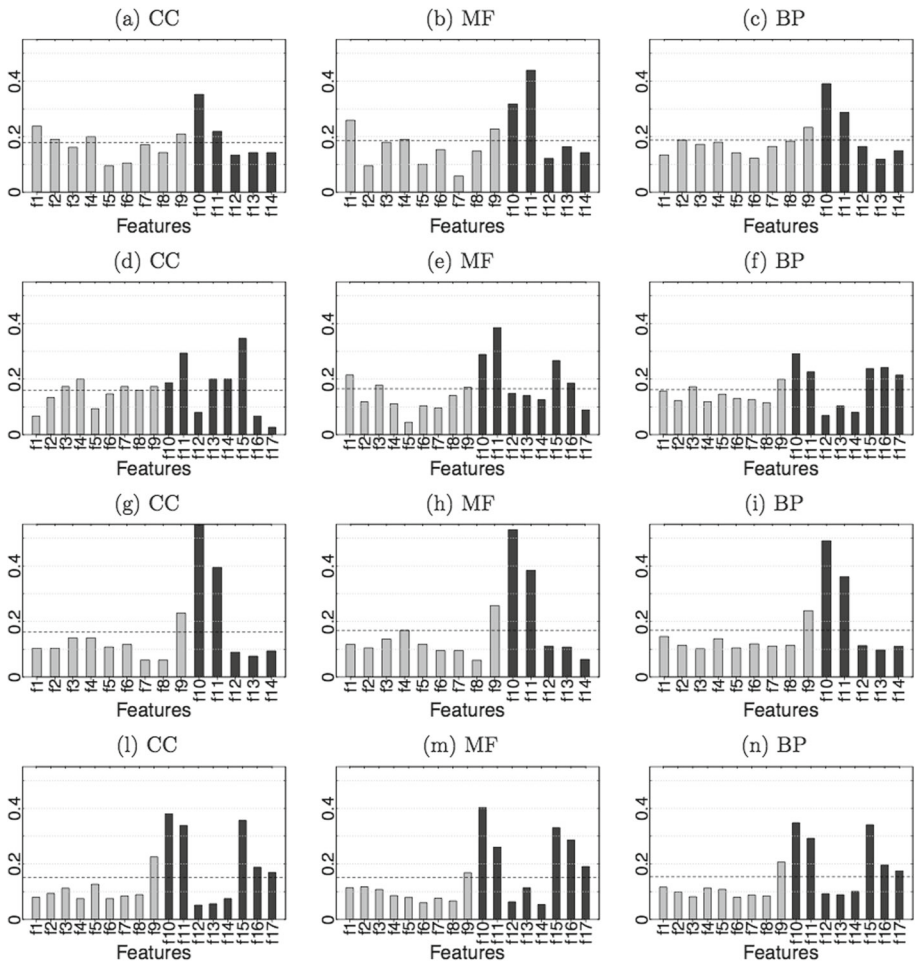

Fig. 1 Proportion of times features are selected by the *SFFS* algorithm on *yeast* (first two rows) and *human* (last two rows) data. Grey and black bars are for term-unaware and term-aware protein features. The black horizontal dashed line corresponds to the mean value of the bars. For each organism, the two rows refer to the use of features $f1$–$f14$ and $f1$–$f17$, respectively. **a**, **d**, **g**, **l** correspond to CC terms, **b**, **e**, **h**, **m** to MF terms, and **c**, **f**, **i**, **n** to BP terms

are above the mean, while in other ones the distribution has a lower variance with seven features overcoming the mean frequency value. The value of q has been tuned as a compromise between these two different conditions.

The selection of reliable negatives was performed through protein ranking, both exploiting the decision function values for SVMs and leveraging the probability to belong to a given class in RFs. The negative proteins selected as reliable are those bottom-ranked by the models. Following the temporal holdout setting, the models were trained on the older annotation release, by fixing a budget of negatives to be selected, subsequently computing the number of false negatives averaged across terms (those reported in Table 3) using the annotations in the newer release. The budget was set as the x% of the total number of proteins, with $x \in \{1, 5, 10, 15, 20, 25, 30\}$.

Results and discussion

Assessing feature relevance

To better evaluate the informativeness of graph centrality measures in classifying proteins in C_{np}, feature selection has been performed by representing proteins both using solely centrality features $f1$–$f14$ and using all features $f1$–$f17$. Figure 1 depicts the obtained frequencies for *yeast* and *human* organism (the results for *mouse* are shown in Additional file 1). In most experiments *Positive neighborhood* is the most informative feature, in both settings adopted. Also *Mean of positive neighborhood* and *1-step Random Walk* are frequently selected, being the top feature respectively on MF branch for yeast data (Fig. 1b, 1e), and on CC branch for yeast and mouse data (Fig. 1d, Additional file 1(d)). Term-aware features (black bars) tend to be predominant over those that are term-unaware, with an exception represented by the *Betweenness centrality*,

often more informative than some term-aware centralities, and nearly the top selected one in mouse (Additional file 1(d-f)). Overall, results for human and mouse show more similar trends than yeast: this fact is probably due to more similar topological structures of the corresponding protein networks (see Table 1), and to the fact that human and mouse are phylogenetically closer to each other than yeast.

Notably, betweenness appears to be much more informative than other geometric centrality measures (e.g., closeness); this outcome is in line with the general observation that betweenness is scarcely correlated with most of the remaining centrality indices, and it may be associated to the fact that the considered networks have a relatively small diameter.

Number of annotations seems to carry a significant signal on mouse, mainly for CC and BP branches (Additional file 1(d, f)), whereas non negligible frequence enhancing are seen for *Weighted clustering coefficient* (yeast – CC), and for *Neighborhood mean* (yeast – MF).

The two settings $f1$–$f14$ and $f1$–$f17$ provide to some extent analogous results, with some differences that however seem not to be related to an underlying physical topology: for instance, *Neighborhood mean* has a significantly higher frequency when discarding *3Prop* features on yeast CC data (Fig. 1a, d), but on yeast BP (Fig. 1c, f) and mouse MF (Additional file 1(b, e)) the opposite happens.

In summary, among term-unaware centralities only the *Betweenness centrality* is significantly enhanced in the majority of experiments, thus helping in discriminating reliable and unreliable negative examples. On the other hand, *Positive closeness, Positive Lin's* and *Positive harmonic* centralities are likely to be useless for this task. Finally, the relevance of *3Prop* features in detecting reliable negatives is decreasing with the number of steps of the random walk.

Protein function prediction

The classification performances in terms of F_1 measure are summarized in Fig. 2, whereas *Precision* and *Recall* results are shown in Additional file 2(a-b) and (c-d), respectively. Interestingly, concerning yeast data, centrality measures allow both classifiers to achieve the best results, even better than those obtained when the *3Prop* features are added to the protein representation ($f1$–$f17$). Such results are confirmed also in terms of *Recall*. On mouse and human data, $f1$–$f14$ representation is still more informative than $f1$–$f17$ when employing RFs, performing similarly to *3Prop* representation. Conversely, SVMs achieve the best results when using all features, and this is likely due to the ability of RFs in capturing the combined effect of features, thus making some of them redundant; on the other hand, SVMs need a more complex protein representation to achieve nearly the same results. SVMs also tend to have a higher *Recall*, while RFs are more precise. This is probably due to the adoption of cost-sensitive SVM learning, which, by attributing a larger misclassification weight to positive instances, tends to increase both the number of true and false positives.

In addition, to give an insight about the impact of each feature on the automated protein-function prediction, each classification experiment was repeated removing in turn one feature and using all the remaining features to represent proteins. The results in terms of F_1 are summarized in Fig. 3: *3Prop* features have been excluded from this experiment because their effectiveness in predicting GO functions has already been assessed in [22]. Due to its complexity, we ran this procedure solely on yeast data. Analogous results based on precision and recall are shown in Additional files 3 and 4, respectively.

Clearly, the most informative measure is *Positive neighborhood* ($f10$), whose removal causes the largest decrease in F_1 values. The removal of *Number of annotations* ($f5$) leads to a significant decay for both classifiers, whereas when singularly eliminating the other features just negligible differences can be observed. These results clearly show that some features are redundant for this task, and the application of feature selection methodologies may lead to better results than those depicted in Fig. 2.

Selecting Negatives

Figure 4 reports the results of the negative selection for yeast and human organisms, whereas the corresponding results for mouse are shown in Additional file 5. A first interesting insight is that when adopting the RF selection method, the number of false negatives significantly decreases compared to the results obtained by the linear SVM selection, suggesting the need of using algorithms able to exploit interactions among features. Considering specific experiments, according to the results of SVM on CC terms and yeast data, the subset of features $f1$–$f17$ *mean* is slightly the most informative, whereas $f15$–$f17$ (*3Prop*) achieve the (largely) worst performance. On BP terms most feature sets perform similarly, whereas on MF data the $f1$–$f14$ set has the top performance. On human data, again $f1$–$f14$ and $f1$–$f14$ *mean* achieve the top performance, with close results, while on mouse data the combination including all features ($f1$–$f17$) is the top performing one.

Different behaviors are observed when RF model is adopted as negative selection procedure: in all the experiments, $f1$–$f14$ and $f1$–$f17$ feature sets are the top performing ones, likely due to the fact that eliminating even the features with low absolute frequencies wastes some useful combined effects that the RF model, for its nonlinear nature, is able to exploit. Indeed, the *top 5* and *mean* feature sets have been selected on the basis of

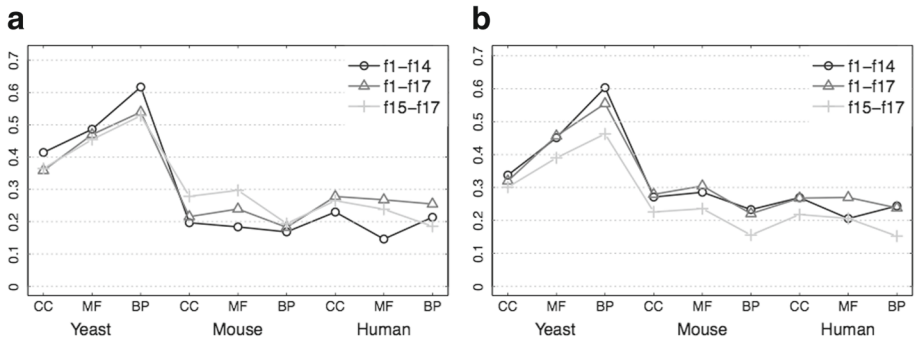

Fig. 2 Performance in terms on F_1 measure averaged across GO branches for linear SVM (**a**) and RF (**b**) classifiers

the absolute individual frequencies reported in Fig. 1 and Additional file 1, computed to provide a general trend of feature informativeness; nevertheless, for specific tasks also coupled frequencies, or more in general feature set frequencies, could help in selecting the optimal combination of features. Another interesting behaviour is related to the *3Prop* features: along with the above mentioned set of features, it represents the top performing set, as opposed to SVM results. Thus these three features, appropriately combined, may also provide information similar to that encoded in the features $f1$–$f17$. On the other side, this requires more complex selection algorithms, thus features $f1$–$f14$ are preferable when, for complexity reasons, simpler models must be adopted.

In summary, features $f1$–$f14$ seem to be more informative for selection algorithms not able to capture non-linear combined effects among feature subsets, whereas they perform similarly to the *3Prop* feature set when selection methods with higher classification capabilities are adopted. Nevertheless, by excluding features $f12$–$f14$, since they are rarely selected by the feature selection algorithm, the computation of the remaining features can nicely scale when input size increases, since features $f1$–$f9$ can be computed offline, being not term-specific, and features $f10$ and $f11$ can be computed efficiently. Conversely, the *3Prop* features need to simulate 6 random walks on the whole network, which also should be row-normalized, affecting thereby scalability.

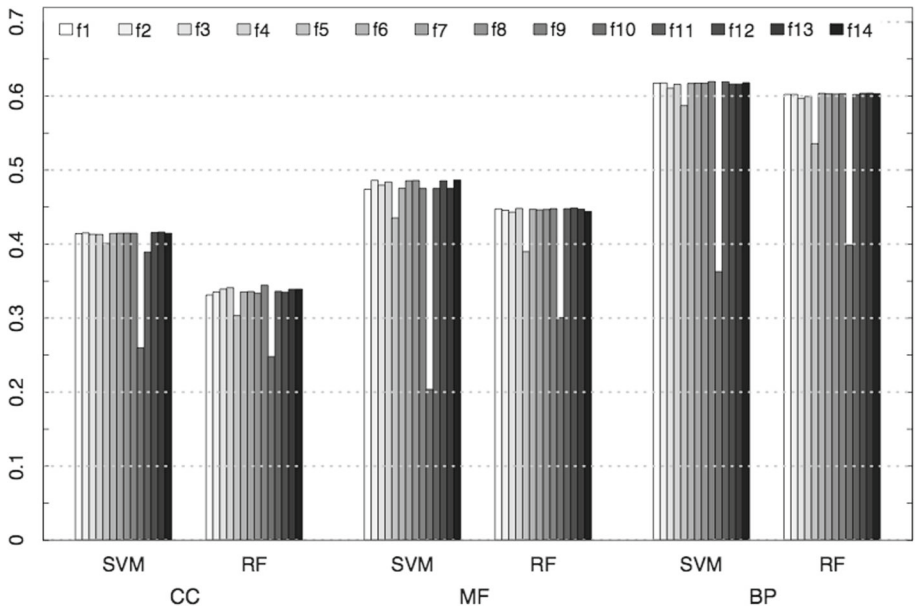

Fig. 3 Average F_1 across GO branch terms on yeast data when removing the corresponding feature

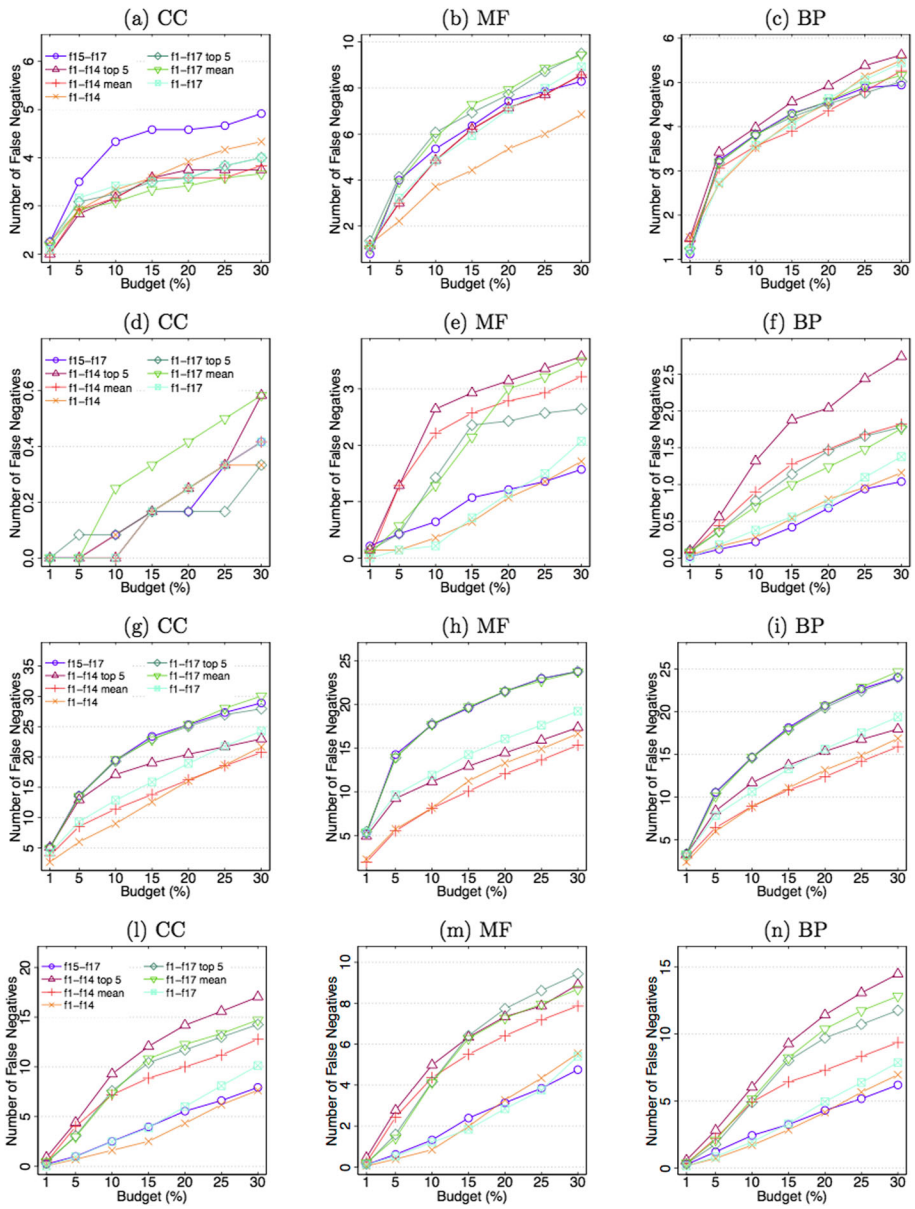

Fig. 4 Number of false negatives averaged across GO terms. Results in the first two rows are obtained on *yeast* data, whereas the last two rows refer to *human* data. First (resp. second) and third (resp. fourth) rows show the results of the SVM (resp. RF) selection algorithm. **a**, **d**, **g** correspond to CC terms, **b**, **e**, **h** to MF terms, and **c**, **f**, **i** to BP terms

Conclusions

Seventeen protein features in biological networks have been studied in this work to assess their 'usefulness' for selecting relevant negatives in the *AFP* context. State-of-the-art graph centrality measures, GO term-aware measures, and protein multifunctionality have been considered. Term-aware features resulted more informative for selecting reliable negative proteins through a state-of-the-art feature selection method in a temporal holdout setting, where the validation is carried out on the proteins that received novel annotations in the temporal holdout period. Among the remaining features, the node (protein) *betweenness* showed an interesting pattern, in particular on mouse data, where it is close to being the most relevant feature. The protein *positive neighborhood* instead is the most predictive feature for the *AFP* problem (that is, when the task to be predicted is the GO term itself). Finally, by supplying the proposed features as input to linear and nonlinear negative selection algorithms, we discovered that there is little or no redundancy among the features

when their linear combination is adopted, whereas their nonlinear interaction also provides novel discriminative abilities to negative selection algorithms.

Overall, apart for those mentioned above, a clear and regular trend did not arise, thus suggesting further analyses under different settings and/or adding (discarding) some features as future investigations.

Additional files

Additional file 1: Figure S1. Proportion of times each feature is selected by the *SFFS* algorithm on *mouse* data and CC (a-d), MF (b-e) and BP (c-f) terms. Same notations as in Fig. 1. (PNG 79 kb)

Additional file 2: Figure S2. Performance in terms on *Precision* (a-b) and *Recall* (c-d) measures averaged across GO branch terms when proteins are represented through $f1$–$f14$, *3Prop* ($f15$–$f17$), and $f1$–$f17$ features. Left and right columns correspond to SVM and RF results, respectively. (PNG 130 kb)

Additional file 3: Figure S3. Evaluation of the impact of features $f1$–$f14$ on the classification performance. Bars correspond to the *Precision* results averaged cross GO branch terms on yeast data when removing the related feature. (PNG 47 kb)

Additional file 4: Figure S4. Evaluation of the impact of features $f1$–$f14$ on the classification performance. Bars correspond to the *Recall* results averaged cross GO branch terms on yeast data when removing the related feature. (PNG 54 kb)

Additional file 5: Figure S5. Number of false negative averaged across GO terms on the *mouse* data. First (resp. second) row shows the results of the SVM (resp. RF) selection algorithm. (PNG 203 kb)

Funding
This work was supported by the grant *Machine learning algorithms to handle label imbalance in biomedical taxonomies*, PSR2017_DIP_010_MFRAS – Università degli Studi di Milano. Publication fees were covered by the grant RV_PRO_RIC16_DMALC_M – Università degli Studi di Milano.

Availability of data and materials
Data are available at http://frasca.di.unimi.it/topfeat.

About this supplement
This article has been published as part of *BMC Bioinformatics Volume 19 Supplement 14, 2018: Selected articles from the 5th International Work-Conference on Bioinformatics and Biomedical Engineering: bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-14.

Authors' contributions
All authors conceived the study. PB carried out the computation of graph-based features. MF participated in the computation of features and in the execution of negative selection and performed the classification experiments. DM participated in the execution of negative selection and drafted the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Published: 20 November 2018

References

1. Ruepp A, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research. 2004;32(18):5539–45.
2. Ashburner M, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000;25(1):25–9.
3. Radivojac P, et al. A large-scale evaluation of computational protein function prediction. Nature Methods. 2013;10(3):221–7.
4. Youngs N, Penfold-Brown D, Bonneau R, Shasha D. Negative Example Selection for Protein Function Prediction: The NoGO Database. PLOS Computational Biology. 2014 06;10(6):1–12. Available from: https://doi.org/10.1371%2Fjournal.pcbi.1003644.
5. Jiang Y, Oron TR, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology. 2016;17(1):184. Available from: https://doi.org/10.1186/s13059-016-1037-6.
6. Mordelet F, Vert JP. A Bagging SVM to Learn from Positive and Unlabeled Examples. Pattern Recogn Lett. 2014 Feb;37:201–9. Available from: https://doi.org/10.1016/j.patrec.2013.06.010.
7. Burghouts GJ, Schutte K, Bouma H, den Hollander RJM. Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos. Machine Vision and Applications. 2014;25(1):85–98.
8. Frasca M, Malchiodi D. Selection of Negative Examples for Node Label Prediction Through Fuzzy Clustering Techniques. In: Advances in Neural Networks: Computational Intelligence for ICT. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer International Publishing; 2016. p. 67–76.
9. Frasca M, Malchiodi D. Exploiting Negative Sample Selection for Prioritizing Candidate Disease Genes. Genomics and Computational Biology. 2017;3(3):47.
10. Gomez SM, Noble WS, Rzhetsky A. Learning to predict protein—protein interactions from protein sequences. Bioinformatics. 2003;19(15):1875–81.
11. Mostafavi S, Morris Q. Using the Gene Ontology Hierarchy when Predicting Gene Function. In: Proceedings of the twenty-fifth conference on Uncertainty in Artificial Intelligence. Arlington: AUAI Press; 2009. p. 419–27.
12. Youngs N, Penfold-Brown D, Drew K, Shasha D, Bonneau R. Parametric Bayesian Priors and Better Choice of Negative Examples Improve Protein Function Prediction. Bioinformatics. 2013;29(9):tt10–98.
13. Frasca M, et al. UNIPred: Unbalance-aware Network Integration and Prediction of protein functions. Journal of Computational Biology. 2015;22(12):1057–74.
14. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nature Biotechnology. 2003;21:697–700.
15. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature. 1999;402:83–6.
16. Oliver S. Guilt-by-association goes global. Nature. 2000;403:601–3.
17. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nature biotechnology. 2000 Dec;18(12):1257–61.
18. Li Y, Patra JC. Integration of multiple data sources to prioritize candidate genes using discounted rating systems. BMC Bioinformatics. 2010;11(Suppl I):S20. https://doi.org/10.1186/1471-2105-11-S1-S20.
19. Bogdanov P, Singh AK. Molecular Function Prediction Using Neighborhood Features. IEEE ACM Transactions on Computational Biology and Bioinformatics. 2011;7(2):208–17.
20. Frasca M, Bassis S, Valentini G. Learning node labels with multi-category Hopfield networks. Neural Computing and Applications. 2016;27(6):1677–92.
21. Frasca M, Lipreri F, Malchiodi D. Analysis of Informative Features for Negative Selection in Protein Function Prediction. In: Bioinformatics and Biomedical Engineering: 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26–28, 2017, Proceedings, Part II. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer International Publishing; 2017. p. 739–51.
22. Mostafavi S, Goldenberg A, Morris Q. Labeling Nodes Using Three Degrees of Propagation. PLoS ONE. 2012;7(12):e51947.
23. Szklarczyk D, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;43(D1):D447–D52. Available from: http://nar.oxfordjournals.org/content/43/D1/D447.abstract.

24. Mostafavi S, Morris Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics. 2010;26(14):1759–65.
25. Hulsman M, Dimitrakopoulos C, de Ridder J. Scale-space measures for graph topology link protein network architecture to function. Bioinformatics. 2014;30(12):i237.
26. Gillis J, Pavlidis P. The Impact of Multifunctional Genes on "Guilt by Association" Analysis. PLoS ONE. 2011 Feb;6(2):e17258+.
27. Frasca M. Automated gene function prediction through gene multifunctionality in biological networks. Neurocomputing. 2015;162: 48–56.
28. Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks. 2010;32(3): 245–51. Available from: http://www.sciencedirect.com/science/article/pii/S0378873310000183.
29. Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. Nature. 1998;393(6684):440–2.
30. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(11):3747–52.
31. Freeman L. Centrality in social networks: Conceptual clarification. Social Networks. 1979;1(3):215–39.
32. Boldi P, Vigna S. Axioms for Centrality. Internet Math. 2014;10(3-4):222–62.
33. Bavelas A. Communication patterns in task-oriented groups. J Acoust Soc Am. 1950;22(6):725–30.
34. Lin N. Foundations of Social Research. New York: McGraw-Hill; 1976.
35. Anthonisse JM. The rush in a directed graph: Mathematical Centre, Amsterdam; 1971. Mathematische Besliskunde No. BN 9/71.
36. Freeman LC. A set of measures of centrality based on betweenness. Sociometry. 1977;40(1):35–41.
37. Hulsman M, Dimitrakopoulos C, de Ridder J. Scale-space measures for graph topology link protein network architecture to function. Bioinformatics. 2014;30(12):237–45.
38. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks; 1984.
39. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters. 1994;15(11):1119–25.
40. Cortes C, Vapnik V. Support-Vector Networks. In: Machine Learning. AA Dordrecht: Kluwer Academic Publishers-Plenum Publishers; 1995. p. 273–97.
41. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.
42. Morik K, Brockhausen P, Joachims T. Combining Statistical Learning with a Knowledge-based Approach – a Case Study in Intensive Care Monitoring. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA: Bled, Slowenien. Morgan Kaufmann Publishers Inc.; 1999. p 268–77.