Supplemental Material for

Genome-scale prediction of moonlighting proteins using a diverse protein association information

Ishita K. Khan and Daisuke Kihara
Contact: dkihara@purdue.edu

Feature Selection Procedure

Detail discussion of feature selection process in the protein-protein interaction (PPI) feature domain is given here. PPI data was extracted from the STRING database (Szklarczyk D et al., 2014). For each protein in the dataset of moonlighting and non-moonlighting proteins (MP and non-MP), we extracted PPI interactions that had sufficient confidence score (> 0.4) in STRING. 124 moonlighting proteins (46.3%) and 61 non-moonlighting proteins (37.7%) in the dataset had such PPI interactions in STRING. Next, we checked the functional divergence of interacting proteins. Interacting proteins for each MP or non-MP were clustered using GO term-based functional similarity. To quantify the functional similarity of two proteins, we used the funsim score (Schlicker et al., 2006). The funsim score of two GO term sets, GO^A and GO^B of a respective size of N and M, is calculated from an all-by-all similarity matrix s_{ij}.

$$s_{ij} = sim(GO_i^A, GO_j^B)_{\forall i\in\{1..N\},\forall j\in\{1..M\}} \qquad \text{(Eqn. 1)}$$

For $sim(GO_i^A, GO_i^B)$, we used the relevance similarity score (Schlicker et al., 2006), SS^{Rel} for two GO terms, c_1 and c_2 is defined as:

$$SS^{\mathrm{Rel}}(c_1,c_2) = \max_{c\in S(c_1,c_2)}\left(\frac{2\log p(c)}{\log p(c_1)+\log p(c_2)}\left(1-p(c)\right)\right) \qquad \text{(Eqn. 2)}$$

Here p(c) is the probability of a GO term c, which is defined as the fraction of the occurrence of c in the GO database (Gene Ontology Consortium, 2013). The root of the ontology has a probability of 1.0. $s(c1,c2)$ is the set of common ancestors of the GO terms c_1 and c_2. The first term considers the relative depth of the common ancestor c to the depth of the two terms c_1 and c_2 while the second term takes into account how rare it is to identify the common ancestor c by chance (i.e. the depth of the c in the GO hierarchy). Since the relevance similarity score is defined only for GO pairs of the same category, a matrix is computed separately for the three categories, Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). Then, the GO_{score} of the matrix of each GO category is computed as follows:

$$GO_{score} = \max\left(\frac{1}{N}\sum_{i=1}^{N}\max_{1\le i\le M} s_{ij}, \frac{1}{M}\sum_{i=1}^{M}\max_{1\le i\le N} s_{ij}\right) \qquad \text{(Eqn. 3)}$$

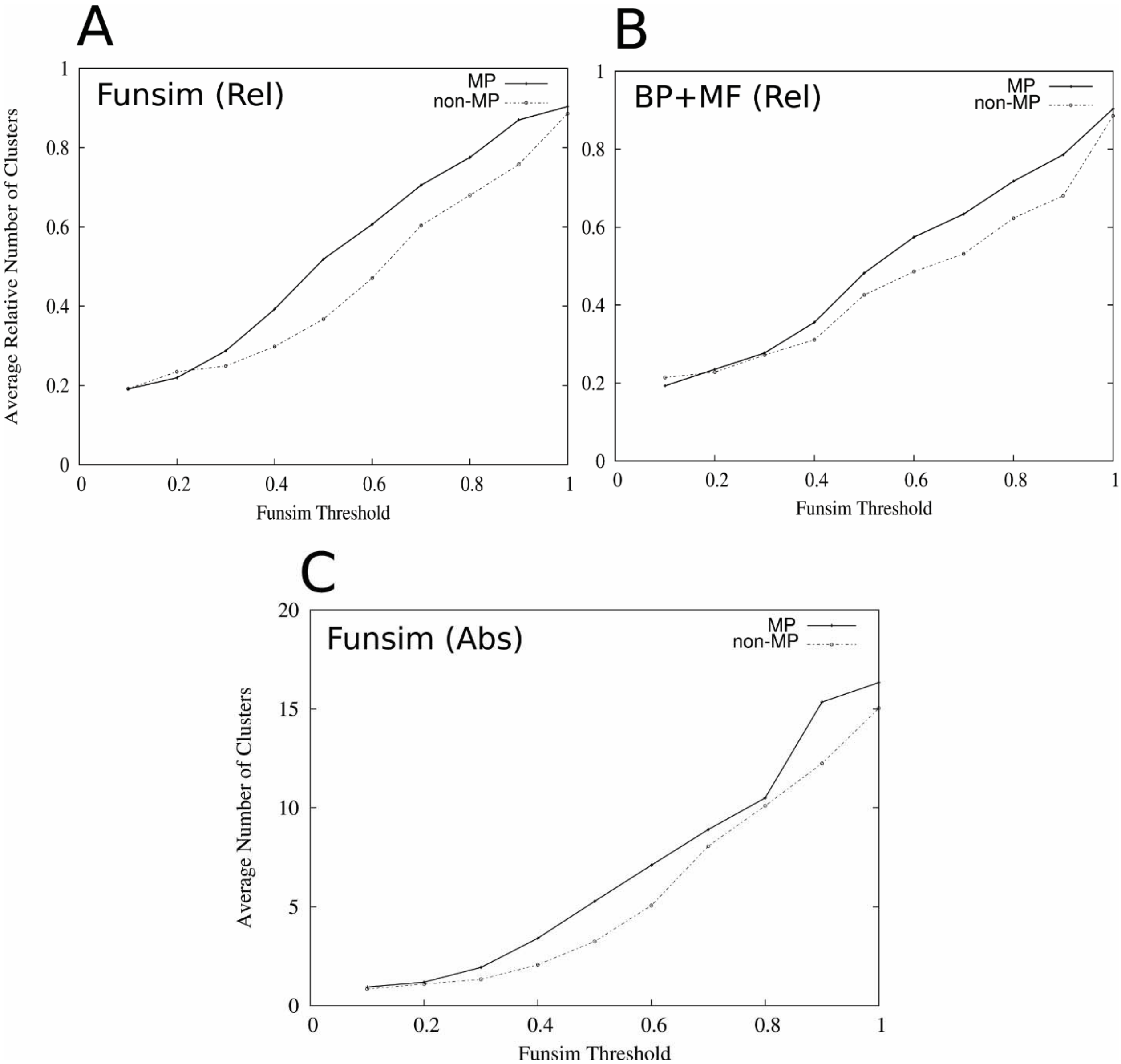
GOscore will be any of the three category scores (MFscore, BPscore, CCscore). Finally, the funsim score is computed as

$$funsim = \frac{1}{3}\left[\left(\frac{MFscore}{\max(MFscore)}\right)^2 + \left(\frac{BPscore}{\max(BPscore)}\right)^2 + \left(\frac{CCscore}{\max(CCscore)}\right)^2\right] \qquad \text{(Eqn. 4)}$$

where $max(GOscore) = 1$ (maximum possible GOscore) and the range of the funSim score is [0, 1].

Using this framework of GO-based functional similarity (Eqn. 4) between two proteins, we clustered the interacting proteins of each of the MPs and non-MPs in the dataset and created a clustering profile (Fig. S1). A clustering profile shows the number of clusters formed by using ten different cutoff values (from 0.1 to 1.0 with an interval of 0.1). For PPI network, we selected three different GO category combinations (Fig. S1). Using these three clustering profiles (Fig. S1A, S1B, S1C), we selected the number of protein clusters (y-axis) at 5 score thresholds each

(0.1, 0.3, 0.5, 0.7, and 0.9 at the x-axis). This procedure constructs 15 features in total for each MPs and non-MPs in the PPI feature domain.



Supplementary Figure S1 – Clustering profiles of interacting proteins of moonlighting and non-moonlighting proteins. Physically interacting proteins for a MP or a non-MP were clustered using 5 cutoff values of a functional similarity score. Single linkage clustering was used. (A-B) the average number of clusters of interacting proteins relative to the number of interacting proteins. The funsim score with all three GO categories was used for A, and the funsim score with BP & MF GO term only in Eqn. 4 was used for B. C) the funsim score with all three GO categories was used. Note that the y-axis is the average number of clusters per interacting proteins in the PPI network, which is different from the value used in (A).

Performance of MPFit with random forest for GO and all omics-based feature combinations

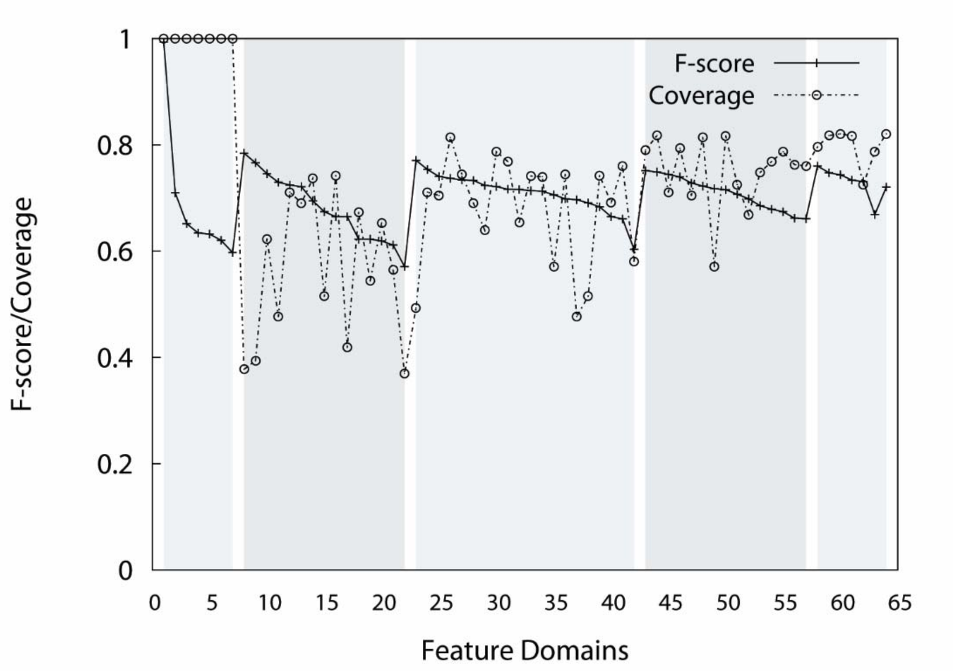


Figure S2. Performance of MPFit with Random Forest. Results of 5-fold cross validation of MPFit with random forest classifier for the GO based features, and all possible feature combinations of the six omics-based features. Feature legends – GO: Gene Ontology, PPI: Protein-Protein Interactions, Phylo: Phylogenetic profile, GE: Gene Expression, DOR: DisOrdered Regions, GI: Genetic Interactions, NET: 3 graph properties – betweenness, degree centrality, closeness centrality. F-score computed as 2-class weighted average over MP/non-MP class. Coverage was computed as the mean protein coverage of MP/non-MP classes. For combinations with the same number of features, the results are sorted by their F-scores.

Numbers 1-64 shown on the x-axis represent the following feature combinations:

1:GO , 2:GE , 3:DOR , 4:Phylo , 5:GI , 6:PPI , 7:NET
8:Phylo+GI , 9:Phylo+NET , 10:Phylo+GE , 11:PPI+Phylo , 12:Phylo+DOR , 13:PPI+GE , 14:GE+DOR ,15:PPI+GI , 16:PPI+DOR , 17:PPI+NET , 18:DOR+NET , 19:GE+GI , 20:GI+DOR , 21:GE+NET , 22:GI+NET
23:Phylo+GI+NET , 24:PPI+Phylo+GE , 25:PPI+GE+GI , 26:PPI+GE+DOR , 27:GE+DOR+NET ,28:PPI+GE+NET , 29:Phylo+GE+GI , 30:Phylo+GE+DOR , 31:PPI+Phylo+DOR 32:Phylo+GE+NET , 33:Phylo+GI+DOR , 34:GE+GI+DOR , 35:PPI+Phylo+GI , 36:Phylo+DOR+NET , 37:PPI+Phylo+NET , 38:PPI+GI+NET , 39:PPI+DOR+NET , 40:GI+DOR+NET , 41:PPI+GI+DOR , 42:GE+GI+NET
43:Phylo+GE+GI+DOR , 44:PPI+Phylo+GE+DOR , 45:PPI+Phylo+GE+NET , 46:Phylo+GE+DOR+NET , 47:PPI+GE+GI+NET , 48:PPI+GE+DOR+NET , 49:PPI+Phylo+GI+NET 50:PPI+GE+GI+DOR , 51:PPI+Phylo+GE+GI , 52:Phylo+GE+GI+NET , 53:GE+GI+DOR+NET , 54:PPI+Phylo+DOR+NET , 55:PPI+Phylo+GI+DOR , 56:Phylo+GI+DOR+NET , 57:PPI+GI+DOR+NET
58:Phylo+GE+GI+DOR+NET , 59:PPI+Phylo+GE+DOR+NET , 60:PPI+Phylo+GE+GI+DOR , 61:PPI+GE+GI+DOR+NET , 62:PPI+Phylo+GE+GI+NET , 63:PPI+Phylo+GI+DOR+NET , 64:PPI+Phylo+GE+GI+DOR+NET

Note that the coverage generally increases as the number of used features increases because missing features were imputed for a protein that have at least one feature among a particular combination considered.

Performance of MPFit with random forest for GO and all possible omics-based feature combinations without missing feature imputation.

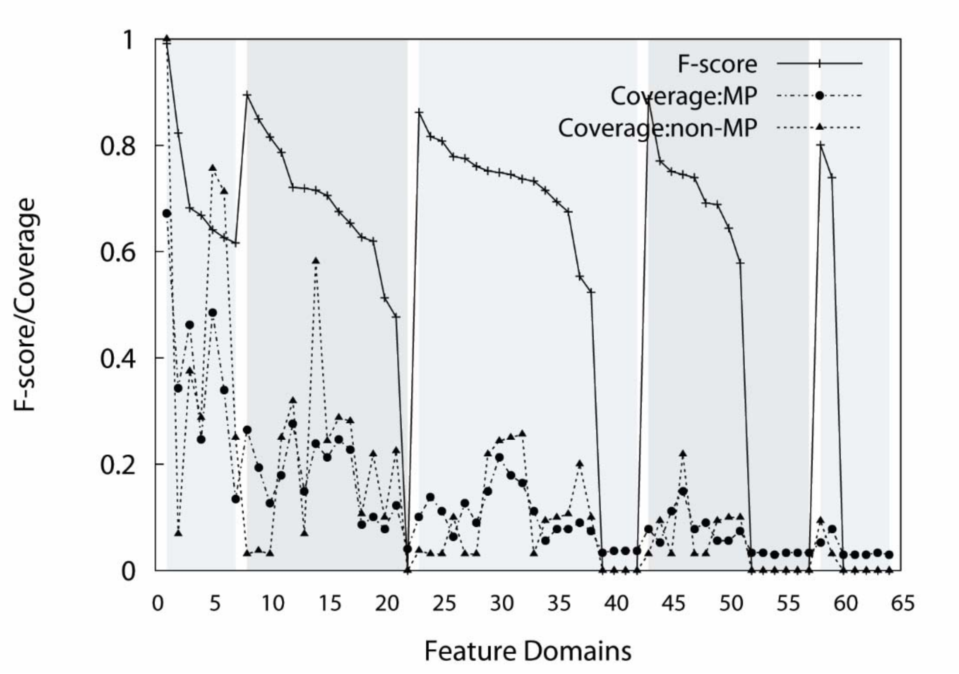


Figure S3. Performance of MPFit with random forest without missing feature imputation. Results of a five-fold cross validation were reported. Coverage is reported separately for the MP class (circles) and non-MP class (triangles).

The feature combinations on the x-axis are the same as Fig. S2:

1:GO , 2:Phylo , 3:PPI , 4:NET , 5:DOR , 6:GE , 7:GI
8:PPI+Phylo , 9:Phylo+DOR , 10:Phylo+NET , 11:DOR+NET , 12:PPI+DOR , 13:Phylo+GE , 14:GE+DOR , 15:GE+NET , 16:PPI+NET , 17:PPI+GE , 18:PPI+GI , 19:GI+DOR , 20:GI+NET , 21:GE+GI , 22:Phylo+GI
23:Phylo+GE+DOR , 24:PPI+Phylo+DOR , 25:Phylo+GE+NET , 26:PPI+GI+DOR , 27:PPI+Phylo+NET , 28:Phylo+DOR+NET , 29:GE+DOR+NET , 30:PPI+GE+NET , 31:PPI+DOR+NET , 32:PPI+GE+DOR , 33:PPI+Phylo+GE , 34:GI+DOR+NET , 35:PPI+GI+NET , 36:PPI+GE+GI , 37:GE+GI+DOR , 38:GE+GI+NET 39:Phylo+GI+DOR , 40:PPI+Phylo+GI , 41:Phylo+GI+NET , 42:Phylo+GE+GI
43:Phylo+GE+DOR+NET , 44:GE+GI+DOR+NET , 45:PPI+Phylo+GE+NET , 46:PPI+GE+DOR+NET , 47:PPI+Phylo+GE+DOR , 48:PPI+Phylo+DOR+NET , 49:PPI+GI+DOR+NET , 50:PPI+GE+GI+DOR , 51:PPI+GE+GI+NET , 52:PPI+Phylo+GI+DOR , 53:Phylo+GI+DOR+NET , 54:Phylo+GE+GI+NET , 55:PPI+Phylo+GI+NET , 56:Phylo+GE+GI+DOR , 57:PPI+Phylo+GE+GI
58:PPI+GE+GI+DOR+NET , 59:PPI+Phylo+GE+DOR+NET , 60:PPI+Phylo+GE+GI+DOR , 61:Phylo+GE+GI+DOR+NET , 62:PPI+Phylo+GE+GI+NET , 63:PPI+Phylo+GI+DOR+NET , 64:PPI+Phylo+GE+GI+DOR+NET

Note that the coverages are low because no imputation was performed.

Random forest classifier with a probabilistic imputation

We also examined a different way of missing feature imputation. In the alternative approach, unlike filling missing features by voting using temporarily assigned feature values as described in Methods (termed “explicit imputation”), the splitting probabilities in random forest that were learned from the training data were used for imputation. The concrete pipeline of this so-called “probabilistic imputation” is as follows: first, we train the random forest with only those proteins that have non-missing features in a certain feature combination. In each branch of each decision tree in the random forest, a fraction is learned (and stored) from the training data that indicates what portion of the proteins in the training set was split with that branch. Then we run down each protein P_i in the test data through each tree in the trained random forest. Whenever P_i falls into a tree node that splits based on a feature which is missing in P_i, we split P_i using the branch probabilities associated with that node that we learned from the training data. Finally, a majority vote is taken for P_i counting the number of trees that classifies P_i in MP/non-MP class. Two slightly different ways of the probabilistic imputation were implemented. The first method takes a weighted majority vote of the trees that classifies a test protein P_i as MP/non-MP, where a weight for one tree T_i is the fraction that is learned from the training data for the leaf branch of T_i that leads to a MP/non-MP class for P_i (Random Forest Probabilistic Imputation, Weighted, RF-PI-W). The second method simply takes a non-weighted majority vote for the test data point P_i (RF-PI-NW, Random Forest Probabilistic Imputation, Not Weighted).

Fig. S4 shows that the explicit imputation overall outperforms the two probabilistic imputation methods. Indeed, the explicit imputation showed higher F-score for all the feature combinations except for two cases: The DOR+NET combination had a higher F-score with RF-PI-NW (difference is 0.0156) and DOR had a higher F-score with RF-PI-W than the explicit imputation (difference 0.0139). Comparing the two probabilistic imputation methods, the non-weighted version (RF-PI-NW) showed a higher F-Score than its weighted counterpart (RF-PI-W) in 38 out of 64 (59.38%) feature combinations.

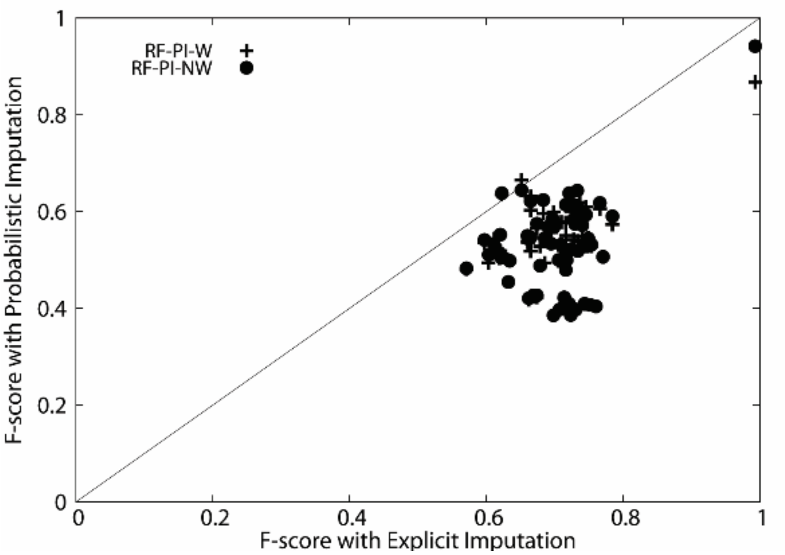
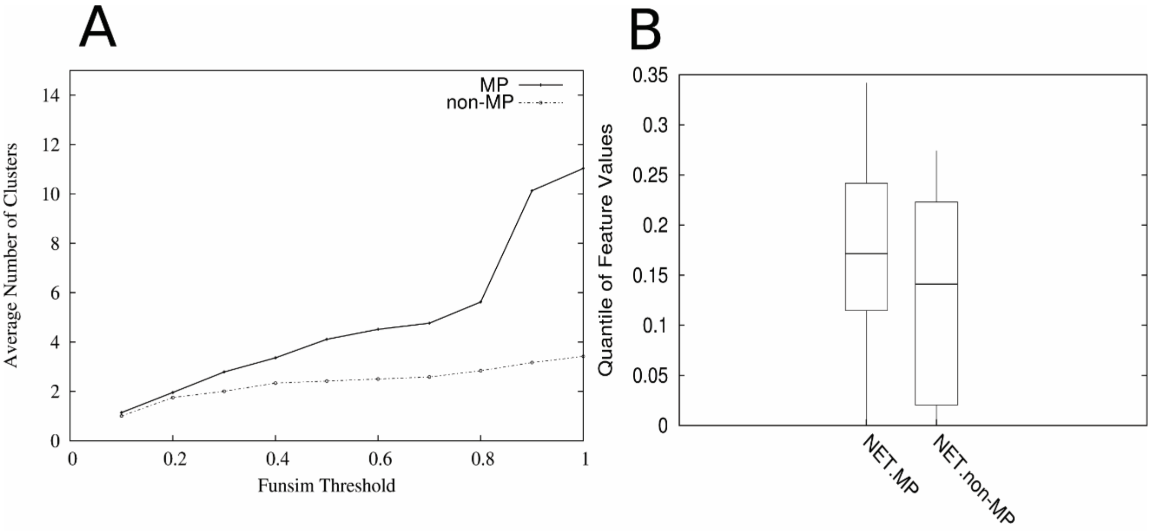


Figure S4. – Performance comparison of explicit and probabilistic imputation. The former is described in Methods. Values shown are the weighted class average F-score over fivefold cross validation. RF-PI-W: Random Forest Probabilistic Imputation, Weighted; RF-PI-NW: Random Forest Probabilistic Imputation, Not Weighted. See text for details.

The reason why the explicit imputation worked better than the probabilistic imputation would be because the latter performs training on only a small the portion of the dataset that have no-missing features for a certain feature combination. For example, for a combination of all six omics features, PPI+Phylo+GE+GI+DOR+NET, there are only eight proteins with no missing features that could be used for training the probabilistic imputation. This lack of sufficient training data resulted in poor F-scores for MPFit with probabilistic imputation (0.409 for both RF-PI-NW and RF-PI-W), which contrasted with the good performance exhibited by MPFit with explicit imputation (F-score: 0.721)

Feature Distribution for Phylogenetic Profile and NET



Supplementary Figure S5 – Feature distribution of moonlighting and non-moonlighting proteins in phylogenetic profile and NET domain. A. Phylogenetically interacting proteins for a MP or a non-MP were clustered using 5 cutoff values of a functional similarity score. Single linkage clustering was used. The funsim score with all three GO categories was used for and the average number of clusters of interacting is shown in the y-axis. B. Quantile plot showing the between-ness centrality (one of the three NET features) of MP and non-MP proteins. PPI interaction network was used to extract the NET features.

Reference List

Gene Ontology Consortium. (2013) Gene Ontology annotations and resources. *Nucleic Acids Research*, 41, D530-D535.

Schlicker,A. et al. (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7.

Szklarczyk D et al. (2014) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, [Epub ahead of print].