Chapter 5

MPFit: Computational Tool for Predicting Moonlighting Proteins

Ishita Khan, Joshua McGraw, and Daisuke Kihara

Abstract

An increasing number of proteins have been found which are capable of performing two or more distinct functions. These proteins, known as moonlighting proteins, have drawn much attention recently as they may play critical roles in disease pathways and development. However, because moonlighting proteins are often found serendipitously, our understanding of moonlighting proteins is still quite limited. In order to lay the foundation for systematic moonlighting proteins studies, we developed MPFit, a software package for predicting moonlighting proteins from their omics features including protein–protein and gene interaction networks. Here, we describe and demonstrate the algorithm of MPFit, the idea behind it, and provide instruction for using the software.

Keywords Moonlighting proteins, Protein function prediction, Dual function, Function annotation, Protein association, Feature imputation, Omics-data, Genome

1 Introduction

While annotating gene function in a genome, the possibility that a gene has two or more distinct functions is usually not explicitly considered. However, an increasing number of proteins have been demonstrated to have more than one biological function, termed moonlighting proteins [1–3]. As long as the additional functions do not interfere with its primary function, moonlighting proteins (MPs) can benefit a cell in several ways. The existence of multifunctional proteins can aid in energy conservation during cell growth and reproduction as well as resulting in a more compact genome. Understanding the variety of MPs will also have clinical benefits as studies have identified a number of MPs that play important roles in cellular activities and biochemical pathways that are involved in cancer, metabolic disorders, and other diseases [4–7]. It has been suggested that for this reason the presence of MPs is under positive selection. This selective pressure and cell-level benefits of MPs

Daisuke Kihara (ed.), *Protein Function Prediction: Methods and Protocols*, Methods in Molecular Biology, vol. 1611,
DOI 10.1007/978-1-4939-7015-5_5,

suggest that moonlighting proteins in diverse genomes might be a common phenomenon.

Considering our current insufficient knowledge of moonlighting proteins, it is a significant challenge for computational protein function annotation to deal with moonlighting proteins. Conventional sequence-based functional annotations methods that are based on the concept of homology or conserved motifs/domains have difficulty identifying additional functions due to the existence of cases where homolog of a MP does not possess the secondary function [8] or has a different secondary function [9, 10]. A study by Gomez et al. compared 11 methods and reported that PSI-BLAST performed relatively well in identifying moonlighting functions [11]. Investigation on our function prediction tools, PFP [12–14] and ESG [15] in comparison with PSI-BLAST [16] has shown that PFP, which mines function information from weakly similar sequences, had the best performance in predicting two distinct functions of MPs [17]. These two studies suggest that secondary functions may be found within distantly related sequences if not among close homologs; however, because MPs are usually found unexpectedly by experiments, the datasets used in the two studies were limited and require further investigation. Due to the limited number of known MPs, systematic studies of MPs are still in its early stage for obtaining a comprehensive picture of proteins' moonlighting functions [18].

Until recently, there have been several bioinformatics approaches proposed for the detection of MPs, but they either relied heavily on the existence of functional annotation of a protein [19, 20] or addressed individual aspects of moonlighting proteins' functional diversity, i.e., sequence similarity [11, 17], motifs/domains [21], structural disorder [22], or protein-protein interaction (PPI) patterns combined with existing gene ontology annotations [19, 20, 23].

In contrast to these previous works, we have recently developed a computational framework for genome-scale characterization of MPs using comprehensive functional and context-based information of proteins [18]. Biological contexts examined in the study include networks of protein–protein interaction (PPI), similarity in the phylogenetic profile, gene expression profile correlation, and genetic interaction. We found that in general MPs tend to have more functionally diverse proteins in their networks, which would be reasonable considering the multi-functional nature of MPs.

Based on this study, we have constructed a prediction model named MPFit (Moonlighting Protein prediction with missing Feature imputation) for identifying moonlighting proteins [24]. To address the diverse nature of moonlighting proteins, MPFit uses various features of proteins ranging from gene ontology (GO) when available [25], several omics data, namely protein–protein interaction (PPI), gene expression, phylogenetic profiles,

genetic interactions, and network-based graph properties (such as node between-ness, degree centrality, closeness-centrality), to protein structural properties such as the length of intrinsically disordered regions in the protein chain. For the omics features, interacting proteins to the target protein are clustered in terms of their functional similarity, and the numbers of clusters are used as features. In general, MPs have more clusters as they interact with proteins of diverse functions. For proteins that do not have certain features available in databases, we have additionally developed an imputation technique using random forest to predict missing features. These features are combined with machine learning classifiers to make moonlighting protein prediction.

MPFit was tested on a dataset of 268 known MPs from a manually curated database, MoonProt [26], which includes genomes such as human (45 proteins, 16.8%), *E. coli* (30 proteins, 11.19%), yeast (27 proteins, 10.1%), and mouse (11 proteins, 4.1%). The benchmark dataset also included 162 negative examples of MPs (termed as non-MP), which we computationally selected based on our previously established GO-based criteria [18]. The benchmark study on the dataset showed that MPFit predicted MPs with over 98% accuracy when proteins' GO terms were available. Using only non-GO-based features, MPFit maintained a high accuracy of over 75%. The latter result is important because it indicates that MPs can be identified by analyzing available omics data even without sufficient function annotations. Last, we have run MPFit with the best performing omics-based feature combinations on three genomes, *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans*, and *Homo sapiens* (human) and found that about 2–10% of the proteomes are potential MPs.

2 The MPFit Algorithm

An overview of the MPFit algorithm is illustrated in Fig. 1. The top panel (Fig. 1a) shows the four phases that MPFit undergoes: feature data collection and construction, feature extraction, missing feature imputation (when needed), and classification for a query protein into moonlighting protein (MP) or non-moonlighting protein (non-MP). A broad range of features are used, i.e., GO annotations, PPI network, gene expression profiles (GE), phylogenetic profiles (Phylo), genetic interactions (GI), disordered protein regions (DOR), and the protein's graph properties in the PPI network (NET) (Fig. 1a).

Figure 1b shows the feature extraction phase for omics data-based features (i.e., PPI, GE, Phylo, and GI) for a protein P_i, we first build a network N_i for P_i. Each node in N_i is a protein; edges in N_i connect proteins that physically interact with each other (in the case of PPI) or that have significant correlation between

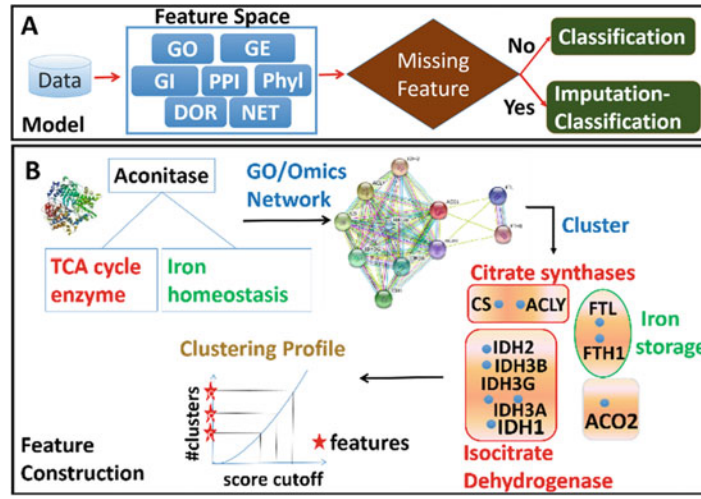

Fig. 1 Schematic diagram of MPFit. (**a**) Overall flowchart of the algorithm. (**b**) Feature construction. It shows the PPI feature extraction for human aconitase as an example

each other (for GE, Phylo, GI). Then proteins in the network are clustered with single linkage clustering in terms of their functional similarity based on their GO annotations. For clustering, several score thresholds are used and the number of clusters constructed at each threshold is recorded. Figure 1b illustrates the feature computation procedure for aconitase in human (*aco1*), an MP, for the PPI network. *Aco1* is an enzyme in the TCA cycle (the primary function) and is also involved in iron homeostasis (the second function). First, we extracted interacting partners for *aco1*, and then the PPI network was clustered based on the GO annotation similarity score of the interacting partners. The figure illustrates that four clusters of proteins were obtained for *aco1* using a certain GO similarity score threshold, where two of these clusters (circled in red) contain proteins related to the TCA cycle, while another cluster (green) was relevant to the second function. Such clustering was performed with five different similarity threshold scores (from 0.1 to 0.9 with an interval of 0.2), which resulted in a clustering profile shown in the bottom of Fig. 1b. Finally, we extracted the number of clusters at each score threshold as the PPI network features of *aco1*.

Proteins do not always have all of the feature data available that is used in MPFit. Thus MPFit uses a data imputation method based on the random-forest algorithm that fills in the missing features [27]. Figure 2 illustrates this procedure. Figure 2a shows the missing feature imputation process used when the MPFit algorithm was trained (i.e., parameters were optimized) using a training dataset and Fig. 2b shows how the imputation is applied in actual prediction.

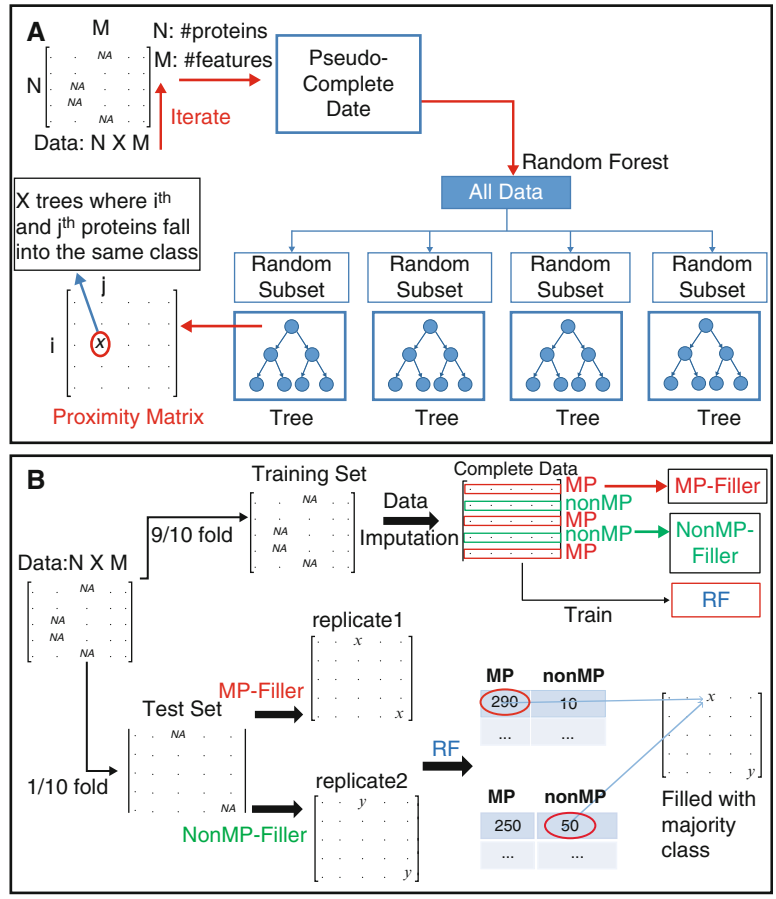

Fig. 2 Missing feature imputation in MPFit. (**a**) Feature imputation in the training stage. (**b**) Imputation in the testing stage

In Fig. 2a, the training dataset is represented top-left corner as a matrix where rows are proteins and columns are features. Missing features in the dataset are represented by NAs. The algorithm starts by replacing NAs with the column medians (i.e., median of the feature). Then a random forest was constructed using the feature set that are temporally filled by the previous step (pseudo-complete data in the figure). Random forest contains a number of decision trees, each of which is trained by a subset of the training set and feature combination, and thus in principle makes prediction in a different decision process. Using the random forest, each protein in the training set is predicted to be either a MP or non-MP, and the results are summarized in a so-called proximity matrix. The (i, j) element of the proximity matrix is the fraction of the trees in the random forest in which the proteins i and j fall in the same class. Now, the imputed value is updated to the weighted average of the non-missing features from other proteins, where weights are the proximities. When the missing features are determined, imputation

is iterated until the proximity matrixes converge or the procedure is iterated ten times. Finally, a random forest RF^{train} is computed with this imputed training data matrix.

To predict new proteins in a test dataset (Fig. 2b), the training dataset with missing values imputed is used to compute two filler-vectors (referred to as MP-filler and non-MP-filler), one for each of the MP and non-MP classes. The ith element of the filler vector MP-filler (non-MP-filler) is the mean of the imputed features at the ith column of the training matrix with the MP (non-MP) class label. The test dataset is represented as a matrix similar to the training data (rows are proteins; columns are features). For the test data row r^i_{test}, since the label (MP/non-MP) is not known, two replicates are made; the missing features in the first replicate are filled using the vector MP-filler and the same for the second replicate is filled using the non-MP-filler vector. Now these two completed test replicates are run down through the previously trained random forest RF^{train}. Each protein receives tree votes of MP and non-MP in RF^{train} from replicates 1 and 2, and the higher vote between the MP vote in replicate 1 and the non-MP vote in replicate 2 determines the final prediction of the protein. In Fig. 2b, the first protein received higher MP votes from replicate 1 (290 votes) over non-MP votes from replicate 2 (50 votes); thus, the protein is predicted to be MP. It is also possible in the MPFit package to fill the missing features of the protein with the voted filler, in this case the MP-filler-vector, and run a classifier other than RF on the filled dataset (such as the SVM or naive-Bayes algorithm) to make the final prediction. However, RF is recommended as it has shown the best performance according to our original work.

3 Using MPFit

3.1 Installing MPFit

MPFit is made freely available for use and can be obtained from http://kiharalab.org/MPprediction. Under the Source Code section, select "MPFit Source Code" to download and extract the zipped archive. MPFit is written in Perl and requires three R package dependencies to run: randomForest, e1071, and stats. The package also includes example input files and output files.

3.2 MPFit Stages

Here, we provide a step-by-step description of the logical stages that the MPFit package undergoes for making a MP or non-MP prediction for query proteins.

Stage 1. Input file preparation. Input file to edit is MPFit_generalized/Feature_Construction/Interacitons/input_uac.txt. Provide UniProt accession for each query protein, each on a new line. The UniProt ID is needed to retrieve feature data in the next stage.

Stage 2. Here the feature data of the queries is generated. MPFit uses following features: GO: Gene Ontology, PPI: Protein–Protein Interactions, Phylo: Phylogenetic profile, GE: Gene Expression, DOR: DisOrdered Regions, GI: Genetic Interactions, NET: 3 graph properties, i.e., between-ness, degree centrality, and closeness centrality of a node in PPI graph.

Data resources: The GO annotations for proteins were obtained from the UniProt database [28], and all GO annotations were extracted regardless of their GO evidence code (IEA or IDA). The STRING database [29] was used as the resource for the PPI network. Protein–protein physical association data were extracted from STRING for PPI. To construct the gene expression (GE) network, expression profiles were obtained from the COEXPRESdb database [30]. Gene pairs that had an absolute value of their Pearson correlation of expression levels within the top 2% among all the pairs in the database were connected as edges in the network. The Phylogenetic profile (Phylo) network was constructed using STRING. A protein pair was connected in the network if they had a sufficient score (>0.7 as recommended by STRING) at "neighborhood," "co-occurrence," or "gene-fusion" in the STRING database. For the genetic interaction (GI) network, we used the BIOGRID database [31]. Gene pairs were extracted that had the "experiment type" listed as "genetic" to be associated in the GI network. For the NET feature, three graph properties of proteins, namely, degree centrality, closeness centrality, and between-ness centrality, were computed from the PPI network. For the DOR feature, disordered region data are taken from the D2P2 database [32] and three properties were computed, namely, the number and the total length of disordered regions as well as the proportion of disordered regions in the entire protein sequence.

Run `$ get_interactions.pl` within each of the /GO (Gene Ontology), /PPI (Protein–protein interactions), /Phylo (Phylogenetic Profiles), and /GI (Genetic Interactions) directories of MPFit_generalized/Feature_Construction/Interactions. No actions are necessary for DOR (Disordered Protein Regions) and NET (Graph Properties in PPI) as the former is simply taken from the data resource and NET is computed from the PPI network later in Stage 3.

Stage 3. In this stage, features based on the data extracted in *Stage 2* are computed. The output of this stage is a feature file of the query proteins for each of the different feature spaces, with non-existing features represented as "NA," which will be imputed by the subsequent stage.

Run `$ compute_features <feature>` for each feature where <feature> is the type of feature to incorporate, i.e., GO, PPI, GI, Phylo, DOR, and NET. These are generated in MPFit_generalized/Feature_Construction/Construct/Features/<feature>.txt

Stage 4. Combine features and perform future imputation.

Change the directory to MPFit_generalized/MPFit_Model/ and run `$ combine_features.pl`

This will create feature files for the input dataset for all possible combinations of the seven omics-based features (PPI, GE, GI, Phylo, DOR, NET) in the MPFit_generalized/MPFit_Model/Data/Features/directory. If the user wishes to omit usage of certain features, they can delete corresponding feature combination from the MPFit_generalized/MPFit_Model/Data/comb.txt flat file.

Stage 5. Run MPFit using a chosen classifier and obtain a prediction result for query proteins.

```
Run $ call_MP_impute_classify.pl  <classifier>
```

This command will perform imputation of missing features with all possible feature combinations mentioned above and run MPFit. Output of this stage is a MP/non-MP prediction of the query proteins, which is provided at MPFit_generalized/MPFit_Model/Result/. Here, three possible classifiers can be given as input by the user: SVM (put 1 at <classifier>), naive Bayes (2), and random forest (3). Random forest is the recommended classifier to use as it showed improved performance over the others in our recent study [24]. Note that the MPFit prediction output is provided for all possible combinations of the seven omics-based features (PPI, GE, GI, Phylo, DOR, NET), and the GO feature. As we see in the next section, it is recommended to use a consensus of the predictions by two omics combinations to make a final MP/non-MP decision: Phylo+GE+GI+DOR+NET and PPI+Phylo+GE. According to the original paper [24], these two combinations performed well in the benchmark study in the case that the GO feature was not available for the query proteins. If the GO term feature is available, it is recommended to use a combination of GO and the two omics-based feature combination mentioned above: Phylo+GE+GI+DOR +NET or PPI+Phylo+GE. The resultant output file shows MP/non-MP predictions for the input proteins, and the number of input proteins predicted as MP/non-MP (Fig. 3).

3.3 MPFit Prediction Accuracy

In the original work of MPFit [24], we reported prediction performance by using various different combinations of features, and concluded that two feature combinations, Phylo+GE+GI+DOR+NET and PPI+Phylo+GE, had high coverage (i.e., proteins in the benchmark dataset that had corresponding features and can be predicted) and high *F*-scores (0.796, 0.760 for coverage and 0.711, 0.754, for *F*-score, respectively). Thus, for the genome-scale prediction performed in the work, the consensus of the two was used. We have also tested three machine learning methods for the classifier in MPFit, random forest, naive Bayes, and SVM and reported that

Prediction Results using Phylo+GE+GI+DOR+NET

```
"----------MPFit Output---------"
"P06745" "mp"
"P10809" "mp"
"P30041" "mp"
"Q43155" "mp"
"Q9CW03" "mp"
"Total Predicted MP"
5
"Total Predicted NONMP"
0
```

Prediction Results using PPI+Phylo+GE

```
"----------MPFit Output---------"
"P06745" "mp"
"P10809" "mp"
"P30041" "mp"
"Q43155" "nonmp"
"Q9CW03" "mp"
"Total Predicted MP"
4
"Total Predicted NONMP"
1
```

Fig. 3 Example of the output file. Prediction of five query proteins, P06745, P10809, P30041, Q43155, and Q9CW03, is shown. These proteins are the same as those discussed in Tables 1 and 2. A prediction result using each feature combination is output in a separate file, but here we show results of two combinations, Phylo+GE+GI+DOR+NET and PPI+Phylo+GE, in one figure. "mp" or "nonmp" for a query protein indicates that it is predicted to be moonlighting or non-moonlighting proteins, respectively

random forest outperformed the other two. Hence, we recommend using the two best feature combinations (Phylo+GE+GI+DOR +NET or PPI+Phylo+GE) with random forest as the final classifier in the MPFit package.

3.4 Case Studies

Here, we show example predictions of five MPs and five non-MPs (Table 1). In Table 1, the first five proteins are MPs, while the latter five are non-MPs. Due to its dual-functional nature, GO term annotations of MPs are classified into a larger number of clusters than non-MPs' as shown in the rightmost column in Table 1. For the prediction, the two recommended feature combinations were used, Phylo+GE+GI+DOR+NET and PPI+Phylo+GE, with random forest as the final classifier. For the five MPs, as shown in Table 2, both feature combinations correctly predicted four proteins as MPs, P30041, P06745, Q9CW03, and P10809. For one

Table 1
Functions of moonlighting and non-moonlighting proteins used in the case study

Protein name	UniProt ID	Function 1 (related GO terms)[a]	Function 2 (related GO terms)	# of GO terms[b]	# of MF-GO Clusters at (0.1, 0.5) SS cutoff[c]
Structural maintenance of chromosomes protein 3	Q9CW03	Form a cohesion complex that maintains proper sister chromatid cohesion (GO:003603, GO:0030893)	Involved in the control of cell growth and transformation. (GO:0051301, GO:0019827)	30	(3, 7)
Ferredoxin-dependent glutamate synthase	Q43155	Glutamate synthase (GO:0006537, GO:0015930)	Subunit of UDP-sulfoquinovose synthase (GO:0051536, GO:0051538)	17	(3, 3)
Peroxiredoxin-6	P30041	Acidic calcium-independent Phospholipase (GO:0006629, GO:0009395)	Can reduce H_2O_2 and short chain organic, fatty acid, and phospholipid hydroperoxides (GO:0016491, GO:0051920)	22	(3, 4)
Glucose-6-phosphate isomerase	P06745	Catalyzes interconversion of glucose-6-phosphate and fructose-6-phosphate (GO:0004347, GO:0016853)	Binds to target cells and causes pre-B cells to mature into antibody secreting cells (GO:0008083, GO:0005125)	17	(2, 7)
60 kDa heat shock protein, mitochondrial	P10809	Protein chaperone, prevents proteins from misfolding, promotes correct refolding (GO:0006986, GO:0006457)	Receptor for HDL affinity for apolipoprotein apoA-II (GO:0002039, GO:0001530)	53	(2, 5)

Sec-independent protein translocase protein TatC	P69423	Transports large folded proteins (GO:0005887,GO:0005622)	-	8	(1, 1)
Probable aminoglycoside efflux pump	P24177	Participates in the efflux of aminoglycosides. (GO:0016021, GO:0016020)	-	8	(1, 1)
C4-dicarboxylic acid transporter DauA	P0AFR2	Responsible for the aerobic transport of succinate from the periplasm (GO:0005829,GO:0005887)	-	9	(1, 1)
PTS system glucose-specific EIICB component	P69786	catalyzes the phosphorylation of incoming sugar substrates (GO:0016021,GO:0005886)	-	10	(1, 1)
Low-affinity inorganic phosphate transporter 1	P0AFJ7	Low-affinity inorganic phosphate transport (GO:0005737, GO:0005886)	-	10	(1, 1)

The first five are moonlighting proteins and the latter five are non-moonlighting proteins

[a]In the columns for function 1 and function 2, representative GO terms related to the functions are shown. Non-moonlighting proteins do not have the secondary function, and thus the function 2 column is left as empty (-)

[b]# of GO terms, the number of GO terms in UniProt

[c]The numbers of clusters of the GO terms constructed by using Semantic functional Similarity (SS) score cutoff of 0.1 and 0.5 are shown. Moonlighting proteins tend to have more clusters, which indicate that they have more diverse GO term annotations reflecting their dual functions

Table 2
Prediction results for the five moonlighting proteins

UniProt ID	Q9CW03	Q43155	P30041	P06745	P10809
Phylo+GE+GI+DOR+NET/Random Forest	No	Yes	Yes	Yes	Yes
PPI+Phylo+GE/Random Forest	Yes	No	Yes	Yes	Yes

Table 3
Prediction results for the five non-moonlighting proteins

UniProt ID	P69423	P24177	P0AFR2	P69786	P0AFJ7
Phylo+GE+GI+DOR+NET/Random Forest	No	No	No	Yes	Yes
PPI+Phylo+GE/Random Forest	No	No	No	Yes	No

protein, Q43155, the five feature combination (Phylo+GE+GI +DOR+NET) predicted correctly as MPs, but the other did not. Table 3 shows the results for the five non-MPs. Three proteins were correctly predicted by the two feature sets as non-MPs; P0AFJ7 was predicted as non-MPs when PPI+Phylo+GE was used, but not when the other feature set was used. For P69786, both feature sets incorrectly predicted it as MPs.

Acknowledgments

This work was partly supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM097528) and the National Science Foundation (IIS1319551, DBI1262189, IOS1127027, DMS1614777).

References

1. Campbell RM, Scanes CG (1995) Endocrine peptides 'moonlighting' as immune modulators: roles for somatostatin and GH-releasing factor. J Endocrinol 147(3):383–396
2. Jeffery CJ (1999) Moonlighting proteins. Trends Biochem Sci 24(1):8–11
3. Weaver DT (1998) Telomeres: moonlighting by DNA repair proteins. Curr Biol 8(14): R492–R494
4. Jeffery CJ (2011) Proteins with neomorphic moonlighting functions in disease. IUBMB Life 63(7):489–494
5. Jeffery CJ (2009) Moonlighting proteins—an update. Mol Biosyst 5(4):345–350
6. Jeffery CJ (2004) Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. Curr Opin Struct Biol 14(6):663–668
7. Jeffery CJ (2003) Moonlighting proteins: old proteins learning new tricks. Trends Genet 19 (8):415–417
8. Ozimek P, Kotter P, Veenhuis M, van der Klei IJ (2006) Hansenula polymorpha and Saccharomyces cerevisiae Pex5p's recognize different, independent peroxisomal targeting signals in alcohol oxidase. FEBS Lett 580(1):46–50
9. Chen XJ, Wang X, Kaufman BA, Butow RA (2005) Aconitase couples metabolic regulation to mitochondrial DNA maintenance. Science 307(5710):714–717
10. Banerjee S, Nandyala AK, Raviprasad P, Ahmed N, Hasnain SE (2007) Iron-dependent RNA-

binding activity of Mycobacterium tuberculosis aconitase. J Bacteriol 189(11):4046–4052

11. Gomez A, Domedel N, Cedano J, Pinol J, Querol E (2003) Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? Bioinformatics **19**(7):895–896
12. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15(6):1550–1556
13. Khan IK, Wei Q, Chitale M, Kihara D (2015) PFP/ESG: automated protein function prediction servers enhanced with Gene Ontology visualization tool. Bioinformatics 31(2):271–272
14. Hawkins T, Chitale M, Luban S, Kihara D (2009) PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins 74(3):566–582
15. Chitale M, Hawkins T, Park C, Kihara D (2009) ESG: extended similarity group method for automated protein function prediction. Bioinformatics 25(14):1739–1745
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
17. Khan I, Chitale M, Rayon C, Kihara D (2012) Evaluation of function predictions by PFP, ESG,and PSI-BLAST for moonlighting proteins. BMC Proc 6(Suppl 7):S5
18. Khan I, Chen Y, Dong T, Hong X, Takeuchi R, Mori H, Kihara D (2014) Genome-scale identification and characterization of moonlighting proteins. Biol Direct 9:30
19. Chapple CE, Robisson B, Spinelli L, Guien C, Becker E, Brun C (2015) Extreme multifunctional proteins identified from a human protein interaction network. Nat Commun 6:7412
20. Pritykin Y, Ghersi D, Singh M (2015) Genome-wide detection and analysis of multifunctional genes. PLoS Comput Biol 11(10): e1004467
21. Hernandez S, Franco L, Calvo A, Ferragut G, Hermoso A, Amela I, Gomez A, Querol E, Cedano J (2015) Bioinformatics and moonlighting proteins. Front Bioeng Biotechnol 3:90
22. Hernandez S, Amela I, Cedano J, Pinol J, Perez-Pons J, Mozo-Villarias A, Querol E (2012) Do moonlighting proteins belong to the intrinsically disordered protein class? J Proteomics Bioinform 5:262–264
23. Gomez A, Hernandez S, Amela I, Pinol J, Cedano J, Querol E (2011) Do protein–protein interaction databases identify moonlighting proteins? Mol Biosyst 7(8):2379–2382
24. Khan IK, Kihara D (2016) Genome-scale prediction of moonlighting proteins using diverse protein association information. Bioinformatics 32(15):2281–2288
25. Consortium GO (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43 (Database issue):D1049–D1056
26. Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, Zabad S, Patel B, Thakkar J, Jeffery CJ (2015) MoonProt: a database for proteins that are known to moonlight. Nucleic Acids Res 43(Database issue):D277–D282
27. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
28. UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43 (Database issue):D204–D212
29. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452
30. Okamura Y, Aoki Y, Obayashi T, Tadaka S, Ito S, Narise T, Kinoshita K (2015) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Res 43(Database issue):D82–D86
31. Oughtred R, Chatr-aryamontri A, Breitkreutz BJ, Chang CS, Rust JM, Theesfeld CL, Heinicke S, Breitkreutz A, Chen D, Hirschman J, Kolas N, Livstone MS, Nixon J, O'Donnell L, Ramage L, Winter A, Reguly T, Sellam A, Stark C, Boucher L, Dolinski K, Tyers M (2016) BioGRID: a resource for studying biological interactions in yeast. Cold Spring Harb Protoc (1):pdb top080754
32. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, Dunker AK, Gough J (2013) D(2)P(2): database of disordered protein predictions. Nucleic Acids Res 41(Database issue):D508–D516