RESEARCH ARTICLE

SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity

Ying Hong Li1☯, Jing Yu Xu1,4☯, Lin Tao1,2☯, Xiao Feng Li1, Shuang Li1, Xian Zeng2, Shang Ying Chen2, Peng Zhang2, Chu Qin2, Cheng Zhang2, Zhe Chen3, Feng Zhu1*, Yu Zong Chen2

1 Innovative Drug Research and Bioinformatics Group, Innovative Drug Research Centre and School of Pharmaceutical Sciences, Chongqing University, Chongqing, 401331, China, 2 Bioinformatics and Drug Discovery group, Department of Pharmacy, National University of Singapore, Singapore, 117543, Singapore, 3 Zhejiang Key Laboratory of Gastro-intestinal Pathophysiology, Zhejiang Hospital of Traditional Chinese Medicine, Zhejiang Chinese Medical University, Hangzhou, P. R. China, 4 School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China

☯ These authors contributed equally to this work.
* zhufeng@cqu.edu.cn

Abstract

Knowledge of protein function is important for biological, medical and therapeutic studies, but many proteins are still unknown in function. There is a need for more improved functional prediction methods. Our SVM-Prot web-server employed a machine learning method for predicting protein functional families from protein sequences irrespective of similarity, which complemented those similarity-based and other methods in predicting diverse classes of proteins including the distantly-related proteins and homologous proteins of different functions. Since its publication in 2003, we made major improvements to SVM-Prot with (1) expanded coverage from 54 to 192 functional families, (2) more diverse protein descriptors protein representation, (3) improved predictive performances due to the use of more enriched training datasets and more variety of protein descriptors, (4) newly integrated BLAST analysis option for assessing proteins in the SVM-Prot predicted functional families that were similar in sequence to a query protein, and (5) newly added batch submission option for supporting the classification of multiple proteins. Moreover, 2 more machine learning approaches, K nearest neighbor and probabilistic neural networks, were added for facilitating collective assessment of protein functions by multiple methods. SVM-Prot can be accessed at http://bidd2.nus.edu.sg/cgi-bin/svmprot/svmprot.cgi.

Introduction

The knowledge of protein function is essential for studying biological processes [1], understanding disease mechanisms [2], and exploring novel therapeutic targets [3,4]. Apart from experimental methods, a number of in-silico approaches have been developed and extensively
used for protein function prediction. These methods include sequence similarity [5], sequence clustering [6], evolutionary analysis [7], gene fusion [8], protein interaction [9], protein remote homology detection [10,11], protein functional family classification based on sequence-derived [12,13] or domain [1] features, and the integrated approaches that combine multiple methods, algorithms and/or data sources for enhanced functional predictions [5,14–16]. A protein functional family is a group of proteins with specific type of molecular functions (e.g. proteases [17]), binding activities (e.g. RNA-binding [18]), or involved in specific biological processes defined by the Gene Ontology [19] (e.g. DNA repair [20]). Moreover, models of protein function prediction have been constructed for more broadly-defined functional families such as transmembrane [21], virulent [22] and secretory [23] proteins, and a large-scale community-based critical assessment of protein function annotation (CAFA) revealed that the improvements of current protein function prediction tools were in urgent need [24]. Despite the development and extensive exploration of these methods, there is still a huge gap between proteins with and without functional characterizations. Continuous efforts are therefore needed for developing new methods and improving existing methods. These efforts have been made possible by the rapidly expanding knowledge of protein sequence [25], structural [26], functional [19] and other [27–30] data.

The uncharacterized proteins comprise a substantial percentage of the predicted proteins in many genomes, and some of these proteins are of no clear sequence or structural similarity to a protein of known function [31,32]. A particular challenge is to predict the function of these proteins from their sequence without the knowledge of similarity, clustering or interaction relationship with a known protein. As part of the collective efforts in developing such prediction methods, we have developed a web-based software SVM-Prot that employs a machine learning method, support vector machines (SVM), for predicting protein functional families from protein sequences irrespective of sequence or structural similarity [12], which have shown good predictive performances [33–40] to complement other methods or as part of the integrated approaches in predicting the function of diverse classes of proteins including the distantly-related proteins and homologous proteins of different functions.

The previous version of SVM-Prot covered 54 functional families. Its predictive accuracies of these families were ranging from 53.03% to 99.26% in sensitivity and from 82.06% to 99.92% in specificity [12]. Since the early 2000s, the number of proteins with sequence information had dramatically expanded from 2 million to more than 48.7 million entries in the UniProt database, and the number of annotated functional families with more than 100 sequence entries had significantly increased from 54 to 192 [25]. Our analysis on all “reviewed” protein entries in the UniProt database revealed that the overwhelming majority (80.23%) of these entries were from those 192 families. The enriched protein sequence data could be employed to expand the coverage and improve the predictive performance of SVM-Prot. Moreover, our earlier study suggested that the prediction performance of SVM could be substantially enhanced by the use of a more diverse set of protein descriptors for representing more comprehensive classes of proteins [41]. Thus, SVM-Prot was upgraded by using the enriched protein data and more diverse protein descriptors to train models for all 192 functional families and to improve the predictive performance of SVM-Prot. The prediction models for an additional set of Gene Ontology [19] functional families will be developed and added into SVM-Prot in the near future.

To facilitate the analysis of specific proteins of the SVM-Prot predicted functional families that might be relevant to a query protein, a new option conducting BLAST sequence alignment was provided to [42] search proteins of the SVM-Prot predicted functional families that were similar to the query protein. Moreover, a batch submission option for loading multiple protein sequences was also included. Given that the functional prediction capacity could be enhanced
by the integration of multiple methods [5,14,15], two machine learning prediction tools, K nearest neighbor (kNN) and probabilistic neural networks (PNN), were integrated into this version of SVM-Prot to facilitate the collective assessment of protein functional families. These two tools had been explored for functional prediction of proteins [43–46] and other biomolecules [47]. Since these two tools had been extensively used for developing over 39 protein functional family prediction models (S1 Table), and because of their potential utility in complementing SVM from the nearest neighbor and neural network perspectives, SVM-Prot could serve the community by providing the alternative protein functional family prediction tools based on these and other machine learning methods.

Results and Discussion

To evaluate the predictive performance of models in the SVM-Prot, the sensitivity (SE), precision (PR), and specificity (SP) of the independent evaluation datasets were calculated and demonstrated in Table 1 and S2 Table. SE, PR and SP of the SVM model were in the range of 50.00~99.99%, 5.31~99.99% and 82.06~99.99%, respectively. In the kNN model, the performances were 51.06~99.99% for SE, 17.86~94.49% for PR and 90.19~99.99% for SP. Moreover, SE, PR and SP of the PNN model were in the range of 60.49~99.99%, 25.00~99.75% and 97.34~99.99%, respectively. The SEs and PRs of the SVM classifier were generally lower and with larger variations than the SPs. This was partly due to the imbalanced training sets with the numbers of non-members greatly surpassing those of the members. Imbalanced training sets were known to adversely affect the machine learning prediction performance, particularly the minority class [48,49]. Moreover, not all functional families were sufficiently covered by the known proteins, particularly those with < 100 known protein members, the inadequate coverage of the respective training sets likely affect SEs to varying degrees.

To further evaluate the capability of SVM-Prot in predicting the functional families of novel proteins, a comprehensive literature search for recently reported novel proteins was conducted using the keyword “novel” in combination with “protein”, “enzyme”, “transporter”, “DNA binding”, “RNA-binding”, “viral”, or “bacterial”. As a result, 42 novel proteins published in 2015 or 2014 that had been explicitly described as novel in the literature were identified. These proteins were not in the SVM-Prot training datasets but with available sequence in the literature or public databases.

S3 Table summarized the prediction results of those 42 novel proteins by SVM-Prot, FFPred 3 [50] and NCBI BLAST [51], and the detailed prediction results were further provided in S4 Table. The function of a novel protein was considered as matched to a computer identified functional family when these two exactly matched at a specifically defined class level. Take the formate-nitrite transporter as an example, it belongs to the formate transporter family, the major intrinsic protein superfamily and the transporter TC1.A class. This families or classes are considered as specifically defined class levels, but the transporter family is too broadly defined. Overall, the number of functional families predicted or outputted by SVM-Prot for each novel protein was in the range of 3~18, and that by FFPred was in the range of 16~55 (if predictions of low reliability were included, the number should change to 45~101). Moreover, the function of 13 out of those 42 novel proteins was correctly assigned to one functional family predicted by SVM-Prot, and 7 (if prediction results of low reliability were included, the number should change to 12) were correctly matched by FFPred (S3 Table). In particular, amongst those 13 proteins predicted by SVM-Prot, 7 were ranked as top-1 in the list of predicted functional families, 2 were ranked as top-2, and 4 were ranked as top-5. However, for FFPred, only one protein was ranked as top-1, another one was ranked as top-2, and 2 more proteins were ranked as top-10. The majority (8 proteins) of the predicted proteins by FFPred were ranked
Table 1. Partial list of the protein functional families covered by SVM-Prot and the prediction performance of the SVM, kNN and PNN models on the independent testing sets. The complete list is provided in S2 Table. The predicted results are given in Sensitivity SE = TP/(TP+FN), Specificity SP = TN/(TN+FP), Precision PR = TP/(TP + FP), where TP = true positive, FN = false negative, TN = true negative, and FP = false positive respectively.

| Family Name                  | GO Id       | Training Dataset | Testing Dataset | Independent Dataset | SVM   | KNN   | PNN   |
|------------------------------|-------------|------------------|-----------------|---------------------|-------|-------|-------|
|                              |             | Positive | Negative | Positive | Negative | SE (%) | SP (%) | PR (%) | SE (%) | SP (%) | PR (%) | SE (%) | SP (%) | PR (%) |
| Actin capping                | GO:0051693  | 652      | 41797    | 128      | 39584      | 95.1   | 99.99  | 93.3   | 73.3    | 99.9   | 55.0   | 91.2    | 99.9    | 71.0    |
| Calmodulin-binding           | GO:0005516  | 465      | 41405    | 223      | 39198      | 87.2   | 99.99  | 90.5   | 70.0    | 99.4   | 41.6   | 82.9    | 99.9    | 84.0    |
| DNA recombination            | GO:0006310  | 1678     | 10614    | 3382     | 18224      | 85.7   | 97.4   | 92.1   | 67.5    | 99.3   | 80.3   | 77.6    | 98.9    | 77.0    |
| DNA repair                   | GO:0008281  | 2142     | 10643    | 1179     | 17646      | 88.7   | 96.8   | 85.9   | 67.6    | 96.8   | 68.0   | 64.3    | 99.3    | 90.4    |
| DNA-directed DNA polymerase  | GO:0003887  | 825      | 9588     | 963      | 19524      | 81.9   | 98.7   | 88.5   | 51.1    | 99.4   | 41.2   | 80.2    | 99.7    | 66.4    |
| EC 1.5 Oxidoreductases (CH-NH donors) | GO:0016645 | 276      | 8755     | 59       | 15283      | 58.6   | 99.6   | 66.1   | 84.5    | 95.8   | 76.0   | 64.2    | 99.2    | 92.6    |
| EC 2.9 Transferases (selenium-containing) | GO:0016785 | 693      | 41834    | 620      | 39620      | 96.0   | 99.99  | 99.3   | 83.7    | 99.7   | 81.4   | 92.4    | 99.9    | 92.5    |
| EC 3.7 Acting on carbon-carbon bonds | GO:0016822 | 1429     | 41786    | 76       | 39543      | 96.5   | 99.9   | 94.6   | 84.4    | 99.4   | 78.0   | 91.2    | 99.9    | 95.2    |
| EC 4.4 Carbon-sulfur lyases  | GO:0016846  | 182      | 8999     | 76       | 15086      | 60.3   | 99.9   | 83.3   | 77.0    | 99.0   | 82.7   | 83.8    | 99.2    | 86.8    |
| EC 5.1 Racemases and Epimerases | GO:0016854 | 379      | 8796     | 95       | 15268      | 53.0   | 99.4   | 53.9   | 80.7    | 93.8   | 80.0   | 69.3    | 98.7    | 94.5    |
| EC 6.6 Forming nitrogen-metal bonds | GO:001002  | 1590     | 41758    | 348      | 39629      | 89.3   | 99.9   | 91.7   | 88.4    | 98.7   | 55.7   | 79.5    | 99.99   | 94.0    |
| Elongation factor activity   | GO:003746   | 1069     | 41788    | 938      | 39570      | 97.5   | 99.99  | 98.8   | 95.8    | 99.6   | 83.7   | 84.1    | 99.9    | 94.0    |
| G protein coupled receptors  | GO:004930   | 927      | 8320     | 4998     | 2016       | 95.6   | 98.1   | 94.5   | 96.6    | 98.9   | 64.1   | 94.1    | 99.9    | 93.4    |
| Growth factor activity       | GO:0008083  | 423      | 41680    | 301      | 39458      | 88.9   | 99.9   | 88.5   | 76.7    | 99.9   | 81.9   | 86.0    | 99.9    | 86.7    |
| GTPase activation            | GO:005096   | 429      | 41584    | 207      | 39359      | 92.9   | 99.9   | 83.3   | 61.8    | 99.6   | 42.2   | 86.7    | 99.9    | 78.4    |
| Heparin-binding              | GO:008201   | 182      | 41591    | 123      | 39344      | 89.1   | 99.9   | 79.9   | 70.7    | 99.9   | 75.0   | 90.2    | 99.9    | 61.0    |
| Lipid degradation            | GO:0018042  | 403      | 8775     | 233      | 20635      | 78.9   | 99.9   | 97.4   | 64.8    | 99.8   | 72.0   | 75.1    | 99.9    | 89.6    |
| Lipid-binding                | GO:0008289  | 274      | 8530     | 166      | 20926      | 84.4   | 99.9   | 93.4   | 72.8    | 99.6   | 71.2   | 66.9    | 99.7    | 72.1    |
| rRNA-binding protein         | GO:0019843  | 708      | 7972     | 1245     | 16044      | 94.1   | 98.7   | 59.0   | 96.5    | 98.3   | 91.4   | 95.8    | 98.7    | 93.6    |
| Sigma factor activity        | GO:0016887  | 101      | 41835    | 60       | 39616      | 87.0   | 99.99  | 85.5   | 68.3    | 99.9   | 50.6   | 83.3    | 99.99   | 81.8    |

doi:10.1371/journal.pone.0155290.t001
within the range of top-27 to top-70. Thus, SVM-Prot is capable of predicting the functional families of novel proteins at comparable yield and reduced false hit rates with respect to FFPred. It should be strongly cautioned that these two servers for protein function prediction were upgraded at different times with varying coverage of training datasets, so the difference in the prediction results may not reflect the true prediction capability of these servers.

As a further comparison, the performance of BLAST on those 42 novel proteins was also evaluated. The number of similarity proteins with E-value < 0.05 for each novel protein was in the range of 0–112, and the function of 30 out of those 42 novel proteins were correctly matched to one of the BLAST identified similarity proteins (20, 2, 6, 2 are ranked as top-1, top-2, top-4 and top-10, respectively) (S3 Table). However, caution needs to be raised about the straightforward comparison of the BLAST results with those of the SVM-Prot and FFPred. BLAST searched proteins may cover the previously or recently deposited similarity proteins that are of the same or similar functions with respect to our tested novel proteins, while some of these similarity proteins may not be in the training set of both SVM-Prot and FFPred. Nonetheless, the better prediction performance of BLAST on these novel proteins suggests a need for more frequent upgrade of the SVM-Prot and FFPred by enriched up-to-date training datasets.

One useful strategy for overcoming the imbalanced datasets problem is to re-construct the training sets into more balanced ones by either over sampling the minority class [48] or under sampling the majority one [49], which might compromise the training datasets by introducing noises to the minority class or reducing the diversity of the majority one. In SVM-Prot, the training sets of the non-members were constructed from the minimal set of representative proteins from the Pfam domain families. Our study showed that further reduction of the training sets by one protein per Pfam family significantly reduced the SPs without much improvement of the SEs. Therefore, no further reduction of the training data was made. Another effective strategy for reducing the negative influence of imbalanced data is to separately optimize the pair of cost parameters of SVM models at the same time [52], particularly the cost for the errors on the positive samples compared to negative ones. In the development of SVM models, due to the very high diversity of each training dataset (containing 7613–46,223 proteins), both the separate and uniform cost parameter optimization scheme led to very high cost parameters for both positive and negative samples that achieve similar levels of prediction performance.

The capability of protein function prediction can be affected by multiple factors, including insufficient diversity of proteins in some functional families, inadequate coverage or representation of certain important structural and/or physicochemical features by the current datasets and protein descriptors, deficiency of the computational algorithms and parameter optimization procedures. The capability of the machine learning functional prediction tools has been enhanced by the expanded protein data, improvement of computational algorithms and exploitation of integrated prediction strategies using multiple methods [53]. In addition to the employment of the continuously expanding protein data, SVM-Prot may be improved by exploring the newly developed computational methods. In particular, there have been new progresses in the development and the use of a new machine learning method, deep learning, for predicting protein secondary structure and other local structural properties [54–56], which may be potentially extended for protein function prediction. SVM-Prot can also be improved by integrating multiple methods and algorithms for enhanced functional predictions [5,14,15].

As an effective ensemble classifier, LibD3C [57] was widely cited by the recent publications aiming at identifying the DNA-binding proteins [58,59], predicting the cytokine-receptor interactions [60] and discovering immunoglobulins [61]. S5 Table summarized the prediction performances of the SVM, LibD3C, kNN and PNN on the independent testing sets of 10 randomly selected representative families covered by the SVM-Prot. These 10 protein families
included 4 enzyme families (EC1.5, EC2.9, EC4.4 and EC5.1), actin capping family, DNA recombination family, DNA repair family, elongation factor activity family, GPCR family and lipid-binding protein family. PR, SE and SP of the SVM model were in the range of 53.9–99.3%, 53.00–97.5% and 96.8–99.99%, respectively. In the LibD3C models the corresponding performances were 52.39–90.51% for PR, 79.23–99.03% for SE and 96.86–99.89% for SP. The kNN method resulted in the performances of 55.0–83.7% for PR, 67.5–96.6% for SE and 93.8–99.9% for SP. Moreover, PR, SE and SP of the PNN model were in the range of 71.0–94.5%, 64.2–94.1% and 98.7–99.9%, respectively. As demonstrated in S5 Table, prediction performances (PR, SE and SP) were comparable among SVM, LibD3C, kNN and PNN, indicating that each method was an effective complement to other methods. It should be strongly cautioned that those 10 randomly tested families may not be enough in representing the prediction performances of all protein families covered by the current SVM-Prot. Therefore, a comprehensive analysis on all SVM-Prot families using above classifiers is needed for the next update of the SVM-Prot.

Methods

Instead of direct alignment or clustering of sequences, the SVM-Prot classification models classifies a protein into functional families based on the analysis of sequence-derived structural and physicochemical properties [33,34]. Proteins known to be in a functional family (e.g., proteases) and those outside the family (e.g., representatives of all non-protease proteins) are used to train a classification model, which recognizes specific sequence-derived features for classifying proteins either into or outside the functional family. Proteins of specific functional family share common structural and physicochemical features [62,63], which may be recognized by a machine learning classification model given the availability of sufficiently diverse training datasets [64].

Data collection

Table 1 and S2 Table provided a partial and complete list of the protein functional families covered by the upgraded SVM-Prot and the predictive performances of the SVM, kNN and PNN models. These families included G-protein coupled receptor family from GPCRDB [63], nuclear receptor family from NucleaRDB [63], 50 enzyme families from BRENDA [62], 20 transporter families from TCDB [65], 1 channel family from LGICdb [66], 24 molecular binding families (e.g. DNA-binding, RNA-binding, iron-binding), 67 Gene Ontology (30 molecular function and 37 biological process) families, and 28 broadly defined functional families from the UniProt database [25]. The 19 broadly defined functional families were selected on the following basis: either the prediction models for these families have been developed (e.g. allergen proteins [47]), or the relevant functions have some common features exploitable for developing prediction models (e.g., cAMP binding). As illustrated above, the reason why protein functional families were derived from multiple sources was partly because of their complementary coverage and different functional perspectives. For instance, 122 functional families predictable in SVM-Prot were not covered by FFPred [50], while 391 functional families provided in FFPred were not covered by SVM-Prot. Thus, SVM-Prot may serve to complement other prediction servers by providing different coverage of protein functional families.

Datasets construction

To prepare datasets for constructing the model of each functional family, the training, testing and independent datasets were carefully prepared by following a strict procedure. Firstly, protein names of members in each family were collected from the UniProt [25], and protein members of the same name but different species origin were grouped together. Secondly, protein
members in each group were iteratively selected and put into the training, testing, and independent datasets as positive samples. Thirdly, to generate negative samples, protein members in each functional family were mapped into the pfam [67] protein families. The pfam families with at least one member of the functional family were named as “positive family”, while the rest of the pfam families were named as “negative family”. Fourthly, 3 representative proteins from each “negative family” were randomly selected and iteratively put into the training, testing, and independent datasets as negative samples.

During the model construction, the parameter optimization for each training set was tested by testing set. When the optimized parameter was found, the training and testing sets were combined together to form a new training set, and the optimized parameter was further applied to train a new model. Then, independent dataset was used to evaluate the performance of the newly constructed model and to detect the overfitting problem. Once the optimized parameter passed the evaluation, it was used to train a final model by integrating training, testing, and independent datasets. All duplicated proteins in each training, testing, independent evaluation dataset or among them were removed before the model construction.

Protein representation

Extensive efforts were applied to the exploration of web-based or stand-alone tools for extracting the features from protein sequences [68,69]. For example, the Pse-in-One is a server for generating various modes of pseudo components of DNA, RNA, and protein sequences [68]. In this work, each sequence is represented by various physicochemical properties including 9 properties of the early version SVM-Prot (amino acid composition, polarity, hydrophobicity, surface tension, charge, normalized Van der Waals volume, polarizability, secondary structure and solvent accessibility) and 4 additional properties in this version SVM-Prot (molecular weight, solubility, number of hydrogen bond donor in side chain, and number of hydrogen bond acceptor in side chain) [69]. All properties are encoded in 3 descriptors, named as composition (C), transition (T), and distribution (D) [70]. C is the fraction of amino acids with a particular property. T characterizes the percent frequency of amino acids of a particular property neighbored by amino acids of another specific property. D measures the fractional chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property is located.

Take a hypothetical protein (AEAAEAEEAEAAAEEAEAAEEAEAEAAE) with 16 alanines ($n_1 = 16$) and 14 glutamic acids ($n_2 = 14$) as an example. The composition (C) for these two amino acids are $n_1/(n_1 + n_2) = 0.53$ and $n_2/(n_1 + n_2) = 0.47$, respectively. Moreover, this protein contains 15 A-to-E and E-to-A transitions (T) with percent frequency of $15/29 = 0.52$. Furthermore, the first, 25%, 50%, 75% and 100% of amino acid A are located within the first, 5, 12, 20, and 29 residues, respectively. Therefore, the distribution (D) for amino acid A can be calculated as $(1/30 = 0.03, 5/30 = 0.17, 12/30 = 0.40, 20/30 = 0.67, 29/30 = 0.97$, and that for amino acid E can also be calculated in the same way. Overall, the amino acid descriptors for this sequence are $C = (0.53, 0.47), T = (0.51)$, and $D = (0.03, 0.17, 0.40, 0.67, 0.97, 0.07, 0.27, 0.60, 0.77, 1.00)$, respectively. In most studies, amino acids are divided into three classes for each property. The combined descriptors for each property consist of 21 elements (3 for C, 3 for T and 15 for D). Moreover, the Moreau-Broto autocorrelation [71] of amino acid index and Pseudo-amino acid composition [72] are added for presenting correlation of the structural and physicochemical properties within each protein sequence.

Protein functional family prediction models

Three types of classification models were developed for predicting protein functional families in SVM-Prot. The first model is SVM, which is based on the structural risk minimization
EC2.7 Transferases transferring phosphorus-containing group family as the top family for this

Fig 1 illustrates an example of SVM-Prot prediction of an EGFR protein sequence, which predicted EC2.7 Transferases transferring phosphorus-containing group family as the top family for this
protein, a click of the BLAST search further indicated that this protein is a receptor protein-tyrosine kinase.

Performance measurement

The performance of SVM, kNN and PNN models were assessed by three different measurements. The first one is using sensitivity (SE), specificity (SP, also known as recall) and precision (PR) to evaluate the predictive performance of the independent validation datasets, which are defined as below:

\[
SE = \frac{TP}{TP + FN} \quad (6)
\]

\[
SP = \frac{TN}{TN + FP} \quad (7)
\]

\[
PR = \frac{TP}{TP + FN} \quad (8)
\]

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively. In the real world, the number of proteins outside a specific
functional family should significantly surpass that within the family. Thus, a slight decline of specificity (SP) would induce tremendous false positive prediction results, which reminds us to primarily focus on the SP when evaluating the model’s prediction performance.

The second measurement is the use of platt’s posterior class probability [50,76] for scoring the predicted functional families of a query protein. This probability has been used for scoring the machine learning classification of protein functional families [50], fold classes [77], transmembrane topology [78], secondary structures [79], and the effect of missense mutations on protein function [80]. It has also been built into such popular machine learning software as LibSVM [73], in which the posterior probability takes the form of a sigmoid function:

$$Pr(y = 1|f) \approx P_{\text{post}}(f) \equiv \frac{1}{1 + \exp(Af + B)}$$  (9)

where $f = f(x)$ is the output of the SVM and the parameters $A$ and $B$ are optimized via cross validation of the training sets.

The last measurement is the test of these models by a set of newly published novel proteins (reported in 2014 and 2015) with their functions reported in the respective publications, and a comparative analysis between SVM-Prot and two popular protein function prediction tools were provided.

**Supporting Information**

S1 Table. List of literature-reported protein functional family prediction models developed by using kNN and PNN methods. (DOCX)

S2 Table. Complete list of the protein functional families covered by SVMProt and the prediction performance of the SVM, kNN and PNN models on the independent testing sets. (DOCX)

S3 Table. List of the novel proteins published in 2015 and 2014 that are not in the SVMProt training sets and have available sequence in the literature or public databases. (DOCX)

S4 Table. The detailed results of the prediction of the functional families of the 42 novel proteins by SVMProt, FFPred and NCBI BLAST. (DOCX)

S5 Table. 10 representative protein functional families covered by SVM-Prot and the prediction performance of the LibD3C, SVM, kNN and PNN models on the independent testing sets. (DOCX)

**Acknowledgments**

This work was funded by the Fundamental Research Funds for the Central Universities (CDJZR14468801, CDJKXB14011, 2015CDJXY); Ministry of Science and Technology, 863 Hi-Tech Program (2007AA02Z160); Key Special Project Grant 2009ZX09501-004 China; and Singapore Academic Research Fund grant R-148-000-208-112.

**Author Contributions**

Conceived and designed the experiments: FZ YZC.
Performed the experiments: YHL JYX LT XFL.
Analyzed the data: YHL JYX LT XFL.
Contributed reagents/materials/analysis tools: YHL JYX LT XFL SL XZ SYC PZ CQ CZ ZC.
Wrote the paper: FZ YZC.
Design the web interface: YHL LT XFL.

References

1. Das S, Stilitee I, Lee D, Lees JG, Dawson NL, Ward J, et al. CATH FunFHMMer web server: protein functional annotations using functional family assignments. Nucleic Acids Res. 2015; 43: W148–153. doi: 10.1093/nar/gkv488 PMID: 25964299
2. Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature. 2009; 461: 1071–1078. doi: 10.1038/nature08467 PMID: 19847258
3. Weinberg SE, Chandel NS. Targeting mitochondria metabolism for cancer therapy. Nat Chem Biol. 2015; 11: 9–15. doi: 10.1038/nchembio.1712 PMID: 25517383
4. Yang H, Qin C, Li YH, Tao L, Zhou J, Yu CY, et al. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. Nucleic Acids Res. 2016; 44: D1069–1074. doi: 10.1093/nar/gkv1230 PMID: 26578601
5. Piovesan D, Giollo M, Leonardi E, Ferrari C, Tosatto SC. INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic Acids Res. 2015; 43: W134–140. doi: 10.1093/nar/gkv523 PMID: 26019177
6. Rentzsch R, Orengo CA. Protein function prediction using domain families. BMC Bioinformatics. 2013; 14 Suppl 3: S5. doi: 10.1186/1471-2105-14-S3-S5 PMID: 23514456
7. Sahranei SM, Luo KR, Brenner SE. SIFTER search: a web server for accurate phylogeny-based protein function prediction. Nucleic Acids Res. 2015; 43: W141–147. doi: 10.1093/nar/gkv461 PMID: 25979264
8. Date SV, Marcotte EM. Protein function prediction using the Protein Link EXPlorer (PLEX). Bioinformatics. 2005; 21: 2558–2559. PMID: 15701682
9. Kotlyar M, Pastrello C, Pivetta F, Lo Sardo A, Cumbaa C, Li H, et al. In silico prediction of physical protein interactions and characterization of interactome orphans. Nat Methods. 2015; 12: 79–84. doi: 10.1038/nmeth.3178 PMID: 25402006
10. Liu B, Chen J, Wang X. Application of learning to rank to protein remote homology detection. Bioinformatics. 2015; 31: 3492–3498. doi: 10.1093/bioinformatics/btv413 PMID: 26136933
11. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014; 30: 472–479. doi: 10.1093/bioinformatics/btt709 PMID: 24318998
12. Cai CZ, Han LY, Ji ZL, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 2003; 31: 3692–3697. PMID: 12824396
13. Lobley AE, Nugent T, Orengo CA, Jones DT. FFPred: an integrated feature-based function prediction server for vertebrate proteomes. Nucleic Acids Res. 2008; 36: W297–302. doi: 10.1093/nar/gkn193 PMID: 18463141
14. Wass MN, Barton G, Sternberg MJ. CombFunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res. 2012; 40: W466–470. doi: 10.1093/nar/gks489 PMID: 22641853
15. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30: 1236–1240. doi: 10.1093/bioinformatics/btu031 PMID: 24451626
16. Xue W, Wang P, Li B, Li Y, Xu X, Yang F, et al. Identification of the inhibitory mechanism of FDA approved selective serotonin reuptake inhibitors: an insight from molecular dynamics simulation study. Phys Chem Chem Phys. 2016; 18: 3260–3271. doi: 10.1039/c5cp05771j PMID: 26745505
17. Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. J Mol Biol. 2005; 345: 187–199. PMID: 15567421
18. Han LY, Cai CZ, Lo SL, Chung MC, Chen YZ. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA. 2004; 10: 355–368. PMID: 14970381
19. Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015; 43: D1049–1056. doi: 10.1093/nar/gku1178 PMID: 25428369
20. Guan Y, Myers CL, Hess DC, Barutcuoglu Z, Caudy AA, Troyanskaya OG. Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biol. 2008; 9 Suppl 1: S3. doi:10.1186/gb-2008-9-s1-s3 PMID: 18613947

21. Wang M, Yang J, Liu GP, Xu JZ, Chou KC. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Eng Des Sel. 2004; 17: 509–516. PMID: 15314209

22. Garg A, Gupta D. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics. 2008; 9: 62. doi:10.1186/1471-2105-9-62 PMID: 18226234

23. Garg A, Raghava GP. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. In Silico Biol. 2008; 8: 129–140. PMID: 18928201

24. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013; 10: 221–227. doi:10.1038/nmeth.2340 PMID: 23353650

25. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43: D204–212. doi:10.1093/nar/gku989 PMID: 25348405

26. Rose PW, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, et al. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. 2015; 43: D345–356. doi:10.1093/nar/gku1214 PMID: 25428375

27. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2015; 43: D213–221. doi:10.1093/nar/gku1243 PMID: 25428371

28. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42: D222–230. doi:10.1093/nar/gkt1223 PMID: 24288371

29. Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res. 2012; 40: D1128–1136. doi:10.1093/nar/gkr979 PMID: 21948793

30. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, et al. Update of TTD: Therapeutic Target Database. Nucleic Acids Res. 2010; 38: D787–791. doi:10.1093/nar/gkp1014 PMID: 19933260

31. Bork P. Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res. 2000; 10: 398–400. PMID: 10779480

32. Hu P, Janga SC, Babu M, Diaz-Mejia JJ, Butland G, Yang W, et al. Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. 2009; 7: e96. doi:10.1371/journal.pbio.1000275

33. Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. Proteins. 2004; 55: 66–76. PMID: 14997540

34. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids Res. 2004; 32: 6437–6444. PMID: 15585667

35. Song L, Li D, Zeng X, Wu Y, Guo L, Zou Q. nDNA-Prot: identification of DNA-binding proteins based on unbalanced classification. BMC Bioinformatics. 2014; 15: 298. doi:10.1186/1471-2105-15-298 PMID: 25196432

36. Lin C, Zou Y, Qin J, Liu X, Jiang Y, Ke C, et al. Hierarchical classification of protein folds using a novel ensemble classifier. PLoS One. 2013; 8: e56499. doi:10.1371/journal.pone.0056499 PMID: 24347146

37. Cheng XY, Huang WJ, Hu SC, Zhang HL, Wang H, Zhang JX, et al. A global characterization and identification of multifunctional enzymes. PLoS One. 2012; 7: e38979. doi:10.1371/journal.pone.0038979 PMID: 22723914

38. Zou Q, Wang Z, Guan X, Liu B, Wu Y, Lin Z. An approach for identifying cytokines based on a novel ensemble classifier. Biomed Res Int. 2013; 2013: 696090. doi:10.1155/2013/696090 PMID: 24027761

39. Wei L, Liao M, Gao X, Zou Q. An Improved Protein Structural Prediction Method by Incorporating Both Sequence and Structure Information. IEEE Trans Nanobioscience. 2014; 14: 339–349.

40. Wei L, Liao M, Gao X, Zou Q. Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique. IEEE Trans Nanobioscience. 2015; 14: 649–659. doi:10.1109/TNB.2015.2450239 PMID: 26355556

41. Ong SA, Lin HH, Chen YZ, Li ZR, Cao Z. Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinformatics. 2007; 8: 300. PMID: 17705863
42. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421. doi:10.1186/1471-2105-10-421 PMID: 20003500

43. Nath N, Mitchell JB. Is EC class predictable from reaction mechanism? BMC Bioinformatics. 2012; 13: 60. doi:10.1186/1471-2105-13-60 PMID: 22530800

44. Naveed M, Khan A. GPCR-MPredictor: multi-level prediction of G protein-coupled receptors using genetic ensemble. Amino Acids. 2012; 42: 1809–1823. doi:10.1007/s00726-011-0902-6 PMID: 21505826

45. Khan ZU, Hayat M, Khan MA. Discrimination of acidic and alkaline enzyme using Chou’s pseudo amino acid composition in conjunction with probabilistic neural network model. J Theor Biol. 2015; 365: 197–203. doi:10.1016/j.jtbi.2014.10.014 PMID: 25452135

46. Wang P, Yang F, Yang H, Xu X, Liu D, Xue W, et al. Identification of dual active agents targeting 5-HT1A and SERT by combinatorial virtual screening methods. Biomed Mater Eng. 2015; 26 Suppl 1: S2233–2239. doi:10.3233/BME-151529 PMID: 26406003

47. Cui J, Han LY, Li H, Ung CY, Tang ZQ, Zheng CJ, et al. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. J Theor Biol. 2015; 365: 197–203. doi:10.1016/j.jtbi.2014.10.014 PMID: 25452135

48. Majid A, Ali S, Iqbal M, Kausar N. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. Comput Methods Programs Biomed. 2014; 113: 792–808. doi:10.1016/j.cmpb.2014.01.001 PMID: 24472367

49. Dai HL. Imbalanced Protein Data Classification Using Ensemble FTM-SVM. IEEE Trans Nano-bioscience. 2015; 14: 350–359.

50. Minneci F, Piovesan D, Cozzetto D, Jones DT. FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. PLoS One. 2013; 8: e63754. doi:10.1371/journal.pone.0063754 PMID: 23717476

51. Bernardes JS, Pedreira CE. A review of protein function prediction under machine learning perspective. Recent Pat Biotechnol. 2013; 7: 122–141. PMID: 23848274

52. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep. 2015; 5: 11476. doi:10.1038/srep11476 PMID: 26098304

53. Lin C, Chen WQ, Qiu C, Wu YF, Krishnan S, Zou Q. LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing. 2014; 123: 424–435.

54. Xu RF, Zhou JY, Wang HP, He YL, Wang XL, Liu B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. BMC Syst Biol. 2015; 9: S10. doi:10.1038/srep15479 PMID: 26552440

55. Wei LY, Zou Q, Liao MH, Lu HJ, Zhao YM. A Novel Machine Learning Method for Cytokine-Receptor Interaction Prediction. Comb Chem High Throughput Screen. 2016; 19: 144–152. PMID: 26652440

56. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 2002; 30: 47–49. PMID: 11752250

57. Horn F, Friend G, Cohen FE. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. Nucleic Acids Res. 2001; 29: 346–349. PMID: 11251313
64. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics. 2002; 18: 147–159. PMID: 11836223

65. Saier MH, Jr. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev. 2000; 64: 354–411. PMID: 10839820

66. Le Novère N, Changeux JP. LGICdb: the ligand-gated ion channel database. Nucleic Acids Res. 2001; 29: 294–295. PMID: 11125117

67. Bateman A, Birney E, Cerruti L, Durbin R, Ettwiller L, Eddy SR, et al. The Pfam Protein Families Database. Nucleic Acids Res. 2002; 30: 276–280. PMID: 11752314

68. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015; 43: W65–71. doi: 10.1093/nar/gkv458 PMID: 25958395

69. Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 2011; 39: W385–390. doi: 10.1093/nar/gkr284 PMID: 21609959

70. Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of Protein-Folding Class Using Global Description of Amino-Acid-Sequence. Proc Natl Acad Sci U S A. 1995; 92: 8700–8704. PMID: 7568000

71. Broto P, Moreau G, Vandycke C. Molecular-Structures—Perception, Auto-Correlation Descriptor and Sar Studies—Use of the Auto-Correlation Descriptor in the Qsar Study of 2 Non-Narcotic Analgesic Series. Eur J Med Chem. 1984; 19: 79–84.

72. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 2006; 34: W32–37. PMID: 16845018

73. Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011; 2: 1–27.

74. Fix E, Hodges JL. Discriminatory analysis: Non-parametric discrimination: Consistency properties. Texas: USAF School of Aviation Medicine; 1951. pp. 261–279.

75. Specht DF. Probabilistic neural networks. Neural Networks. 1990; 3: 109–118.

76. Lin HT, Lin CJ, Weng RC. A note on Platt's probabilistic outputs for support vector machines. Mach Learn. 2007; 68: 267–276.

77. Grassmann J, Reczko M, Suhai S, Edler L. Protein fold class prediction: new methods of statistical classification. Proc Int Conf Intell Syst Mol Biol. 1999; 1999: 106–112.

78. Reynolds SM, Kall L, Riffle ME, Blimes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol. 2008; 4: e1000213. doi: 10.1371/journal.pcbi.1000213 PMID: 18989393

79. Guermeur Y, Geourjon C, Gallinari P, Deleage G. Improved performance in protein secondary structure prediction by inhomogeneous score combination. Bioinformatics. 1999; 15: 413–421. PMID: 10366661

80. Needham CJ, Bradford JR, Bulpitt AJ, Care MA, Westhead DR. Predicting the effect of missense mutations on protein function: analysis with Bayesian networks. BMC Bioinformatics. 2006; 7: 405. PMID: 16956412