POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles

Supplementary Material

SUPPLEMENTAL INFORMATION

Table S1 provides a comprehensive list of a wide range of research areas and application topics within the literature for which PSSM profile-based features have proved to be useful.

Table S1. Research topics and areas of PSSM profile-based features in the literature.

Research Area	Feature Descriptors by the Corresponding Research Work	References
Protein structural class prediction	**AAC-PSSM**, **DPC-PSSM**, and **AADP-PSSM**	(Liu, et al., 2010)
	AAC-PSSM, and **PSSM-AC**	(Liu, et al., 2012)
	AAC, and **PSSM**	(Chen, et al., 2008)
	AAC-PSSM, **PSSM-AC**, consensus sequence descriptors, and physicochemical property features	(Dehzangi, et al., 2013)
	RPSSM, and secondary structures	(Ding, et al., 2014)
	tri-gram-PSSM	(Tao, et al., 2015)
	PSSM, physicochemical property features, and GO feature descriptors	(Li, et al., 2014)
	EDP, **EEDP**, and **MEDP**	(Zhang, et al., 2014)
	AAC-PSSM, **TPC**, and **AATP**	(Zhang, et al., 2012)
	PSSM	(Xia, et al., 2012)
Post-translational modification site prediction	**PSSM**, disorder scores, secondary structures, solvent accessibilities, AAIndex, and AAC	(Jiang, et al., 2013)
	AAC, AGG, BLOSUM62, charge-hyd, CKSAAP, binary profiles, disorder scores, KNN, and **PSSM**	(Chen, et al., 2015)

	AAIndex, physicochemical descriptors, **PSSM**, evolutionary conservation scores, CKSAAP; predicted disordered regions, predicted secondary structures, predicted solvent accessibilities; BP, cellular component, molecular function, functional domain from InterPro, pathway information, functional domain from Pfam, protein-protein interaction annotations; functional domain annotations, nucleotide-binding site annotations, disulfide bond annotations, post-translational modified residue annotations, active site annotations, natural variant annotations, metal ion-binding site annotations, and other binding site annotations	(Li, et al., 2015)
	PSSM, AAC, DPC, solvent accessible surface areas, BLOSUM62, PWM, AAIndex	(Bui, et al., 2016)
	binary profiles, AAC, secondary structures, solvent accessible surface areas, and **PSSM**	(Chauhan, et al., 2012)
	PSSM, AAIndex, secondary structures, solvent accessible surface areas, and disorder scores	(Zhang, et al., 2014)
Protein fold recognition	**PSSM**, profile-profile alignments, secondary-structure specific gap-penalties, classic pair and solvation potentials	(Lobley, et al., 2009)
	Sequence and family information; sequence-sequence alignment; sequence-profile alignment; profile-profile alignment (including **PSSM**), and structural information	(Cheng and Baldi, 2006)
	***k*-separated-bigrams-PSSM**	(Sharma, et al., 2013)
	***k*-separated-bigrams-PSSM**	(Saini, et al.)
	PSSM-AC, and **PSSM-CC**	(Dong, et al., 2009)
	tri-gram-PSSM	(Paliwal, et al., 2014)
	PSSM	(Hong, et al., 2011)
Prediction of protein-protein interactions	**D-FPSSM**, and **S-FPSSM**	(Zahiri, et al., 2013)
	physicochemical descriptors, **PSSM-AC**, and **PSSM-CC**	(Guo, et al., 2008)
	physicochemical descriptors, evolutionary conservation scores, information entropy, **PSSM**, ASA, NC_a, and NC_r	(Deng, et al., 2009)
	PSSM, and predicted solvent accessibility	(Murakami and Mizuguchi, 2010)
	PSSM, and **PSSM-AC**	(Gao, et al., 2016)
	PSSM, and ***k*-separated-bigrams-PSSM**	(An, et al., 2016)
	PSSM, and solvent accessible surface areas	(Melo, et al., 2016)

Membrane protein topology prediction	**Pse-PSSM**	(Chou and Shen, 2007)
	PSSM, and IAMPC (Integrated Approach for Membrane Protein Classification)	(Pu, et al., 2007)
	physicochemical descriptors, and **PSSM**	(Hayat and Khan, 2012)
	PSSM, and secondary structures	(Yan, et al., 2015)
	PSSM, AAC, DPC, physicochemical descriptors, and biochemical feature descriptors	(Mishra, et al., 2014)
	PSSM, and biochemical feature descriptors	(Chen, et al., 2011)
Prediction of protein subcellular localization	**PSSM**	(Xie, et al., 2005)
	DP-PSSM	(Juan, et al., 2009)
	Pse-PSSM	(Juan, et al., 2008)
	PSSM, and PSFM	(Guo, et al., 2006)
	PseAAC, and **PSSM-AC**	(Wang and Li, 2013)
Bacterial protein prediction	AAC, secondary structures, solvent accessibilities, physicochemical descriptors, and **PSSM**	(Yang, et al., 2013)
	AAC, DPC, **PSSM-composition**, and **PSSM-AC**	(Zou, et al., 2013)
	AAC, DPC, and **PSSM**	(Garg and Gupta, 2008)
	AAC, DPC, MM, and **PSSM**	(Selvaraj, et al., 2016)
	AAC, DPC, physicochemical property features, and **PSSM**	(Restrepo-Montoya, et al., 2011)
HIV 1 protease cleavage prediction	**PSSM**	(Jensen, et al., 2003)
	PSSM	(Jensen, et al., 2006)
	geno2pheno, and **PSSM**	(Seclen, et al., 2011)
	geno2pheno, and **PSSM**	(Bunnik, et al., 2011)
Protein disorder prediction	**PSSM**, and BLOSUM62	(Jones and Cozzetto, 2015)
	PSSM	(Jones and Ward, 2003)
	PSSM, and physicochemical property features	(Shimizu, et al., 2007)
	PSSM, secondary structures, and solvent accessibilities	(Becker, et al., 2013)

	PSSM, and physicochemical descriptors	(Su, et al., 2006)
Protein secondary structure prediction	**PSSM**	(Bouziane, et al., 2011)
	PSSM, and SPSSM	(Li, et al., 2012)
	PSSM	(Tang, et al., 2011)
	conformation parameters, **PSSM**, net charges, hydrophobic and side chain mass	(Huang and Chen, 2013)
Prediction of DNA-binding sites	**PSSM**	(Ahmad and Sarai, 2005)
	biochemical descriptors and **PSSM**	(Wang, et al., 2010)
	AAC, DPC and **PSSM**	(Kumar, et al., 2007)
	physicochemical descriptors, biochemical descriptors and **PSSM**	(Huang, et al., 2011)
	binary profile, BLOSUM62 and **PSSM**	(Hwang, et al., 2007)
Prediction of RNA-binding sites	**PSSM**, **smoothed-PSSM**	(Cheng, et al., 2008)
	physicochemical descriptors, hydrophobicity, relative accessible surface areas, secondary structures, **PSSM**, and side-chain environment	(Liu, et al., 2010)
	PSSM	(Kumar, et al., 2008)
	PSSM, residue interface propensity, predicted residue accessibility values, and residue hydrophobicity scores	(Murakami, et al., 2010)
	biochemical property features, and **PSSM**	(Wang, et al., 2010)
	PSSM, **smoothed-PSSM**, and sequence-derived descriptors	(Walia, et al., 2012)
Protein function prediction	**AB-PSSM**, **RPM-PSSM**, and physicochemical property features	(Jeong, et al., 2011)
	PSSM, UniProtKB/Swiss-Prot text mining, amino acid trigram mining, FFPRED, orthologous groups, profile-profile comparison, and functional space	(Cozzetto, et al., 2013)
	GO annotations, and **PSSM**	(Wass and Sternberg, 2008)

*PSSM denotes that the original PSSM profile was directly used in the corresponding paper.

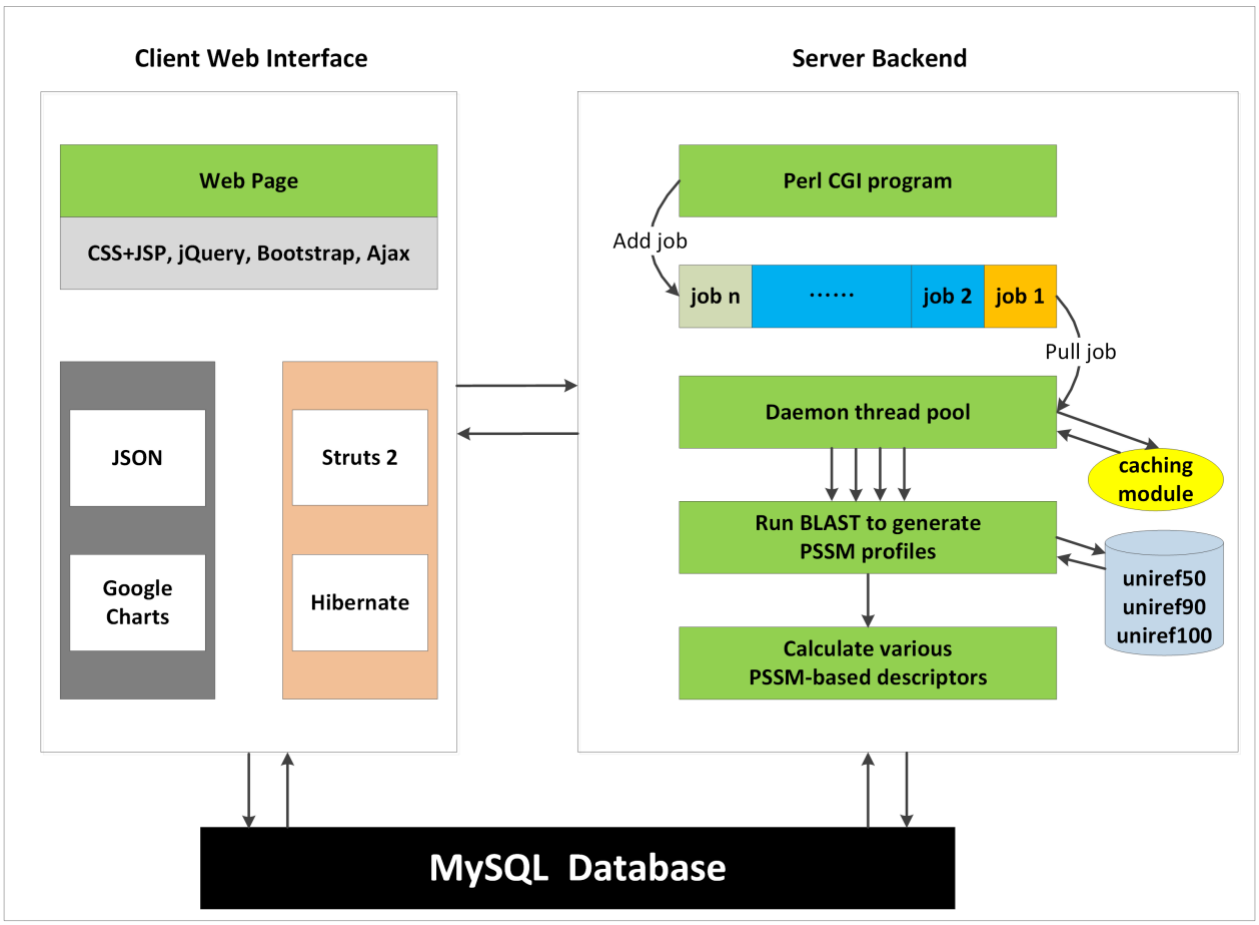

Fig. S1. The architecture of the POSSUM web server.

The architecture of the POSSUM server is illustrated in Fig. S1. There are two main components to this architecture: Client Web Interface and Server Backend. These two components can interactively exchange the data of submitted jobs, and inform each other. Please refer to the main text of the manuscript for a detailed description and discussion.

The POSSUM server is currently configured and hosted on an extensible cloud computing facility provided by the e-Research Centre at Monash University, equipped with 4 cores, 16GB memory and a 1TB hard disk. Importantly, this configuration can be readily expanded and upgraded in accordance with the increasing user demand of the webserver.

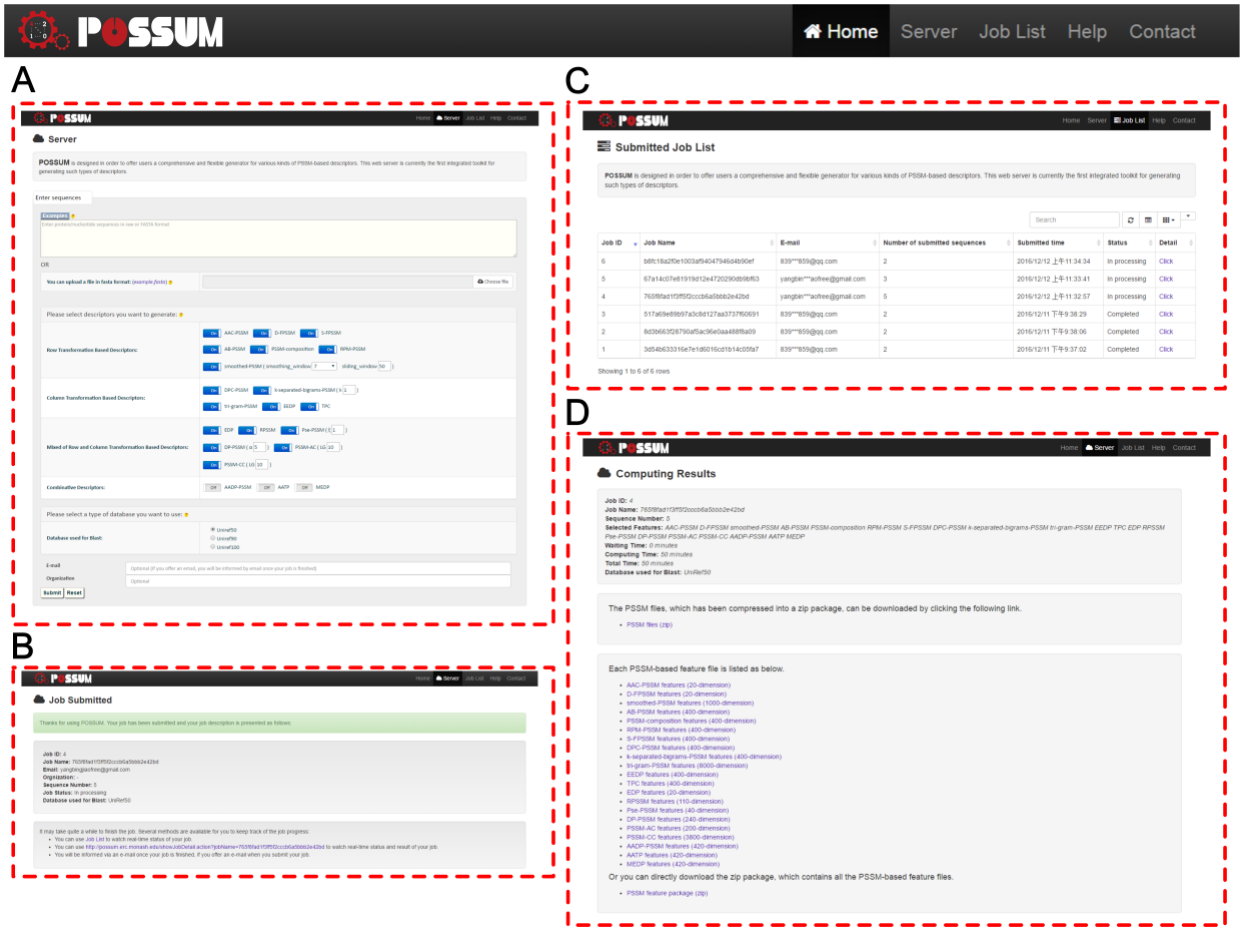

Fig. S2. An example of the user interface of the POSSUM server: (A) Webpage displaying users' submission options; (B) Webpage summarizing the submitted information; (C) Webpage listing status of all submitted jobs, and (D) The result page containing the original PSSM files and calculated descriptors by POSSUM, as well as the links for downloading the corresponding PSSM-based feature files.

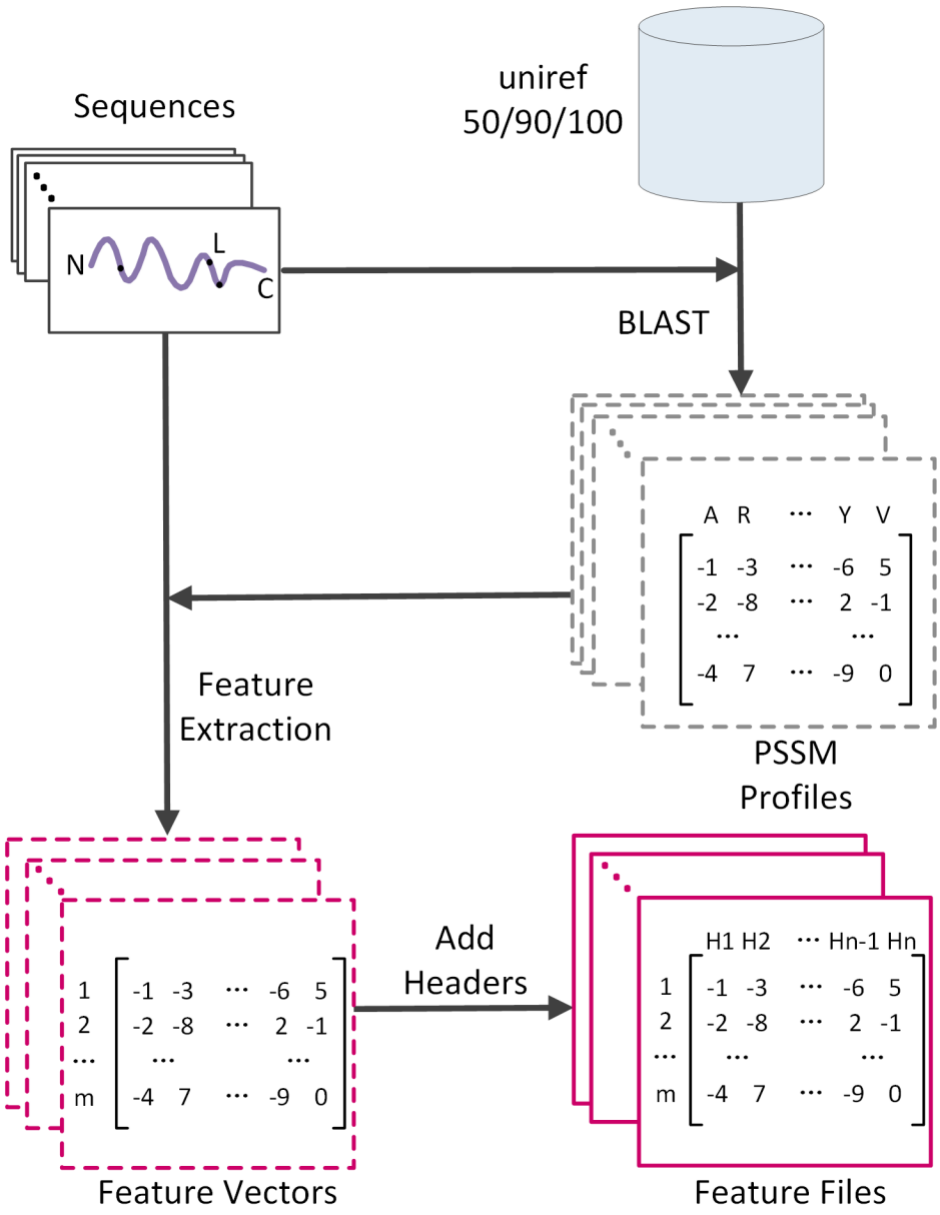

Fig. S3. Workflow of the POSSUM server.

The workflow of the POSSUM server is displayed in Fig. S3.

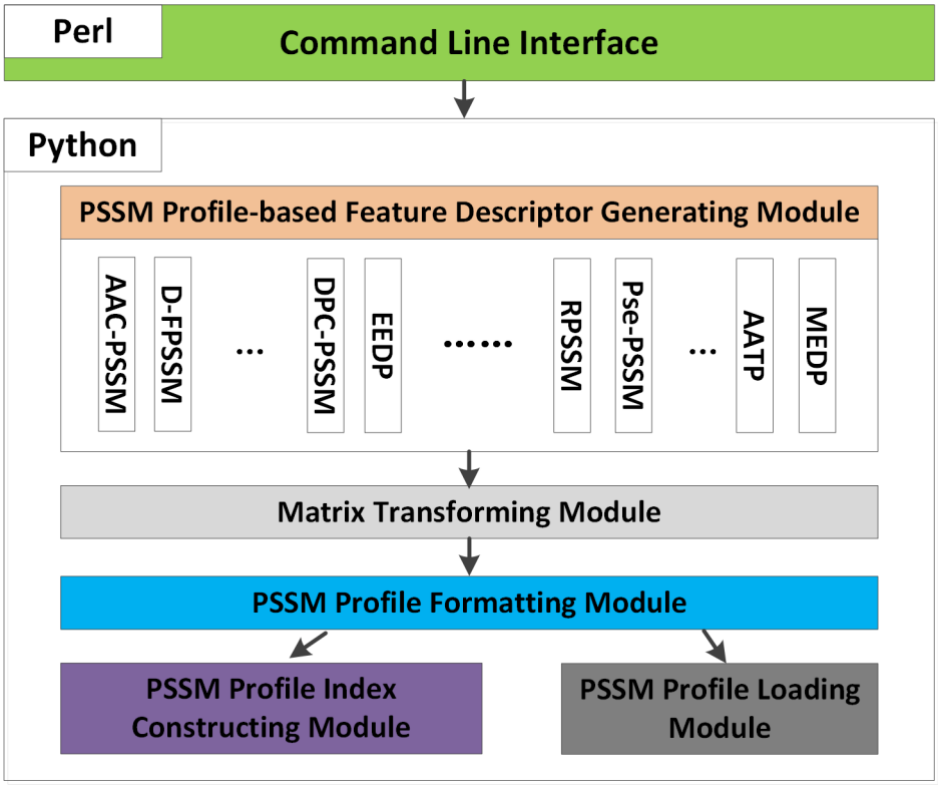

Fig. S4. Architecture of the POSSUM standalone toolkit.

The architecture of the POSSUM standalone toolkit is displayed in Fig. S4. The toolkit was implemented in Python (for core function implementation) and Perl (for universal command line interface). The major components of the toolkit are briefly described as follows:

- Command Line Interface: This module is made available to provide a universal and user-friendly command line interface, via which users can effectively interact with the toolkit. This module allows users to specify and apply different parameters and it invokes the descriptor generating process.

- PSSM Profile-based Feature Descriptor Generating Module: This module can be used to wrap up and output the descriptor files based on the raw descriptor vectors (generated by the Matrix Transforming Module) in accordance with the user-specified parameters.

- Matrix Transforming Module: This module can be used to transform the PSSM matrix (which is abstracted from the original PSSM profile) to generate user-specified raw descriptor vectors. Various applicable matrix transformation functions in groups of row transformations, column transformations, and mixture of row and column transformations are available within this module.

- PSSM Profile Formatting Module: This module can be used to abstract the PSSM matrix from the PSSM profile.

- PSSM Profile Index Constructing Module: This module is a fundamental part of the program that scans the FASTA sequences and the PSSM profile folder to build a hash map for each query sequence and its corresponding PSSM profile.

- PSSM Profile Loading Module: This module looks up the hash table (built by the PSSM Profile Index Constructing Module) to check the availability of the PSSM profile for a sequence and loads the corresponding PSSM profile into the memory.

Comparison of the computational time of PSSM profile-based feature descriptor generation by POSSUM on different uniref databases

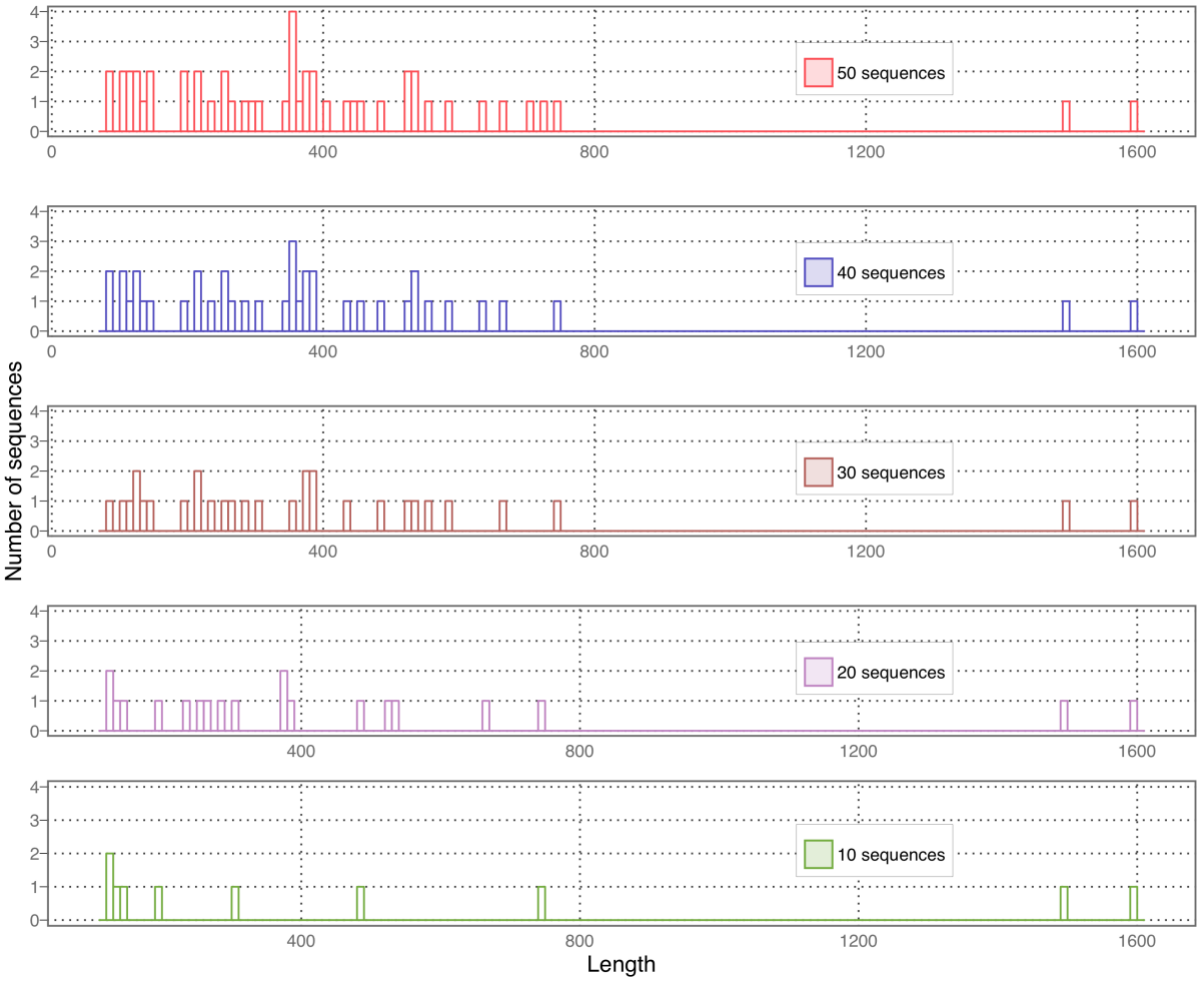

Fig. S5. The distribution of submitted sequence lengths.

Next, in order to illustrate the computational power of POSSUM, we randomly selected 50 sequences from the UniProt database (http://www.uniprot.org/). We subsequently evaluated POSSUM server's CPU computing time for generating PSSM profile-based feature descriptors on the three different uniref databases (i.e. uniref50, uniref90 and uniref100). Specifically, we submitted 10, 20, 30, 40 and 50 sequences to the POSSUM server to generate all 21 types of PSSM profile-based feature descriptors. The distributions of sequence lengths for these tasks, their computational time against different uniref databases, and the distributions of the computational time over a certain task (generating PSSM profile-based feature descriptors for 50 sequences on uniref50) are shown in Fig. S5, Fig. S6 and Fig. S7, respectively.

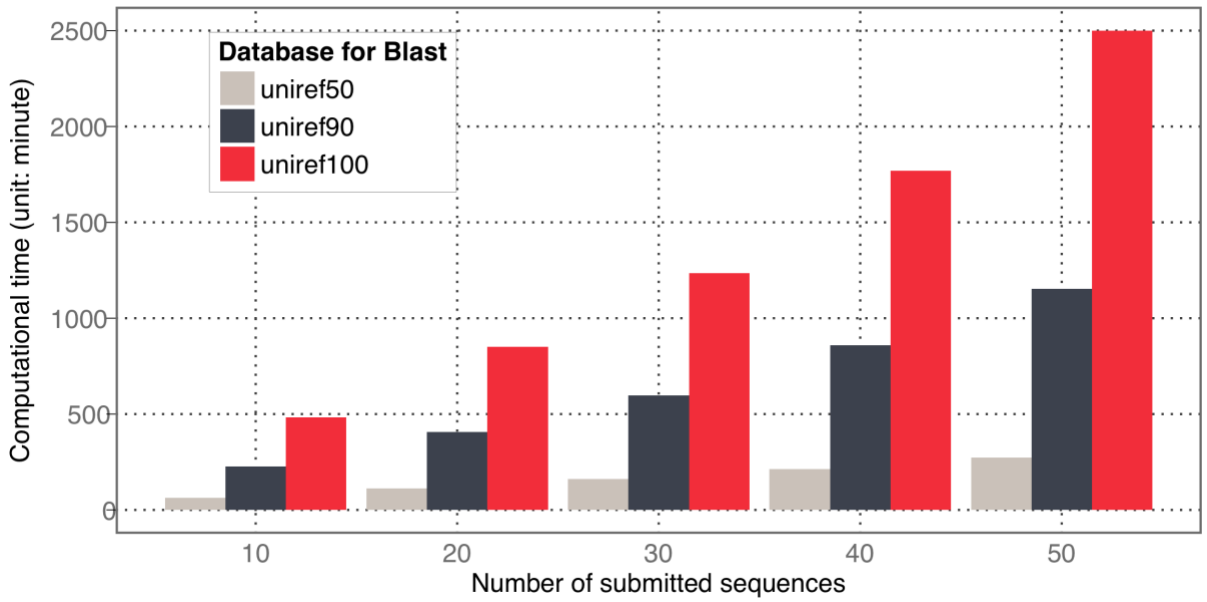

Fig. S6. Comparisons of the computational time for the POSSUM server to process and generate the PSSM profile-based feature descriptors of varying numbers of sequences using three different uniref databases (i.e. uniref50, uniref90 and uniref100). The three databases were generated based on different sequence identity thresholds. The computational time on the y-axis indicates the total computational time for submitted sequences (unit: minute).

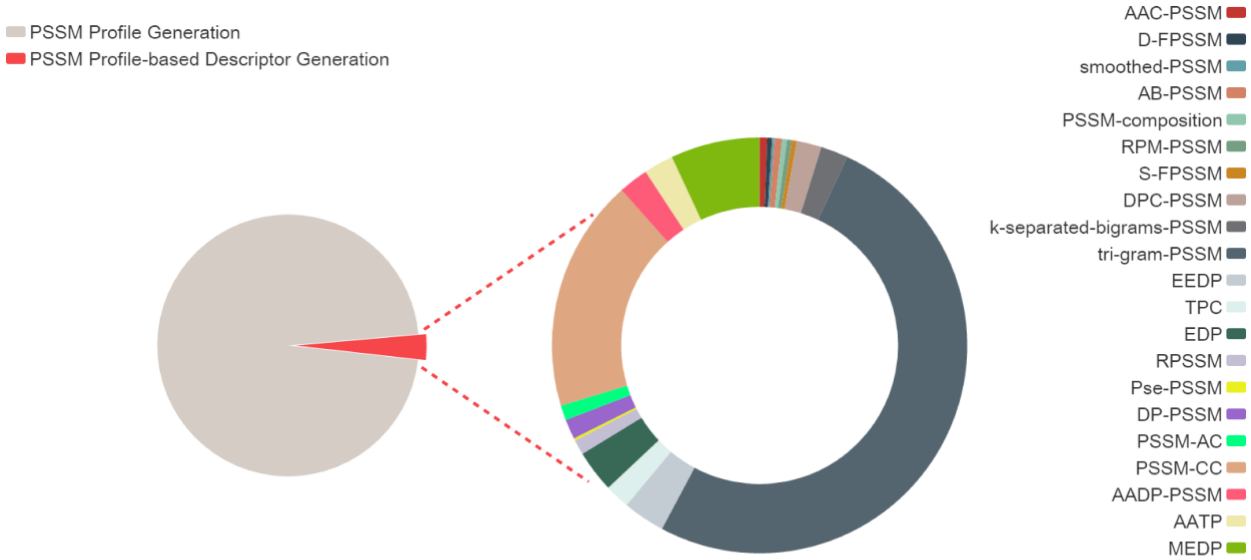

Fig. S7. Distribution of the computational time involved in the task of generating all types of PSSM profile-based feature descriptors as a whole. The results were obtained over the 50 randomly selected sequences based on the uniref50 database.

Fig. S6 suggests a near linear relationship between the CPU computational time and the number of submitted sequences, provided the same uniref database was used. Nevertheless, the computational time considerably varied depending on which uniref database was used for the same task. Users should keep in mind there is a trade-off between the quality of the PSSM profiles generated and computational efficiency, and select which options would best suit their practical needs.

Furthermore, generating a PSSM profile is the most time-consuming step during the entire feature descriptor generation process (Fig. S7, left panel), accounting for 96.8% of the computing time. In this regard, parallelization of the PSSM profile generation is expected to significantly boost the throughput of the POSSUM server. In addition, we also notice that during the calculation of PSSM profile-based feature descriptors (Fig. S7, right panel), the tri-gram-PSSM is the most time-consuming step due to a very large number of features (described as a vector in a 8000-dimensional space) required to be generated.

Application of POSSUM-calculated features to the prediction of type IV secretion effectors and performance evaluation based on the 10 times of 5-fold cross-validation tests

To demonstrate the usefulness of PSSM-based features generated by POSSUM, we further applied POSSUM features to the prediction of type IV secretion effector proteins and examined the performance of machine learning models trained using these features. We employed the dataset prepared in (Zou, et al., 2013) as the benchmark dataset for the performance comparison, which included 340 type IV effectors and 1132 non-effectors. After removing the sequence redundancy, 338 positive and 338 negative samples were finally selected. Based on this dataset, all 21 types of feature descriptors were generated using POSSUM. In addition, some well-known sequence-based descriptors were used as a reference, such as composition of k-spaced amino acid pairs (CKSAAP) (Chen, et al., 2011), amphiphilic pseudo-amino acid composition (APAAC), pseudo-amino acid composition (PAAC), and quasi-sequence-order (QSO), which are originally proposed in (Chou, 2000; Chou, 2001) and implemented using the protr package (Xiao, et al., 2015).

Table S2. The list of performances of various descriptors.

Descriptors groups	Descriptor	SN	SP	ACC	F-value	MCC
Row transformation	AAC-PSSM	0.883±0.007	0.919±0.009	0.901±0.005	0.899±0.005	0.803±0.011
	D-FPSSM	0.829±0.010	0.895±0.008	0.862±0.007	0.856±0.008	0.725±0.014
	smoothed-PSSM	0.835±0.005	0.919±0.005	0.877±0.003	0.871±0.003	0.757±0.007

	AB-PSSM	0.868±0.004	0.925±0.007	0.896±0.005	0.893±0.004	0.795±0.009
	PSSM-composition	0.879±0.008	0.908±0.003	0.894±0.004	0.891±0.004	0.789±0.007
	RPM-PSSM	0.866±0.007	**0.935±0.008**	0.900±0.003	0.896±0.003	0.803±0.007
	S-FPSSM	0.843±0.008	0.923±0.006	0.883±0.005	0.877±0.005	0.769±0.010
Column transformation	DPC-PSSM	0.873±0.006	0.915±0.006	0.894±0.004	0.891±0.005	0.789±0.009
	k-separated-bigrams-PSSM	0.859±0.007	0.916±0.011	0.888±0.006	0.884±0.006	0.777±0.013
	tri-gram-PSSM	0.869±0.007	0.890±0.009	0.880±0.007	0.878±0.007	0.760±0.014
	EEDP	0.878±0.005	0.931±0.007	**0.904±0.005**	0.901±0.005	**0.810±0.010**
	TPC	0.904±0.005	0.897±0.007	0.901±0.004	0.901±0.004	0.802±0.007
Mixed of row and column transformation	EDP	0.854±0.005	0.915±0.004	0.884±0.003	0.880±0.004	0.771±0.006
	RPSSM	0.871±0.006	0.922±0.004	0.897±0.003	0.893±0.003	0.794±0.006
	Pse-PSSM	0.874±0.007	0.926±0.006	0.900±0.005	0.897±0.006	0.801±0.011
	DP-PSSM	0.873±0.007	0.933±0.005	0.903±0.004	0.900±0.005	0.808±0.007
	PSSM-AC	0.770±0.008	0.914±0.010	0.842±0.006	0.829±0.006	0.691±0.013
	PSSM-CC	0.815±0.007	0.912±0.007	0.863±0.006	0.855±0.005	0.730±0.011
Combination of above descriptors	AADP-PSSM	0.876±0.005	0.912±0.004	0.894±0.004	0.891±0.004	0.789±0.007
	AATP	**0.905±0.007**	0.902±0.005	0.903±0.005	**0.903±0.005**	0.807±0.010
	MEDP	0.875±0.006	0.929±0.002	0.902±0.003	0.899±0.004	0.806±0.005
Sequence-based descriptors	AAC	0.778±0.008	0.826±0.005	0.802±0.006	0.797±0.006	0.605±0.012
	DPC	0.788±0.010	0.824±0.013	0.806±0.009	0.801±0.009	0.613±0.020
	CKSAAP	0.797±0.011	0.830±0.007	0.814±0.007	0.810±0.008	0.629±0.014
	APAAC	0.766±0.011	0.806±0.017	0.786±0.011	0.781±0.010	0.573±0.022
	PAAC	0.769±0.013	0.805±0.015	0.787±0.008	0.782±0.008	0.575±0.017

	QSO	0.762±0.006	0.842±0.009	0.802±0.005	0.794±0.005	0.606±0.010

The rows highlighted by grey are the descriptors achieving MCC values of 0.800 or larger.

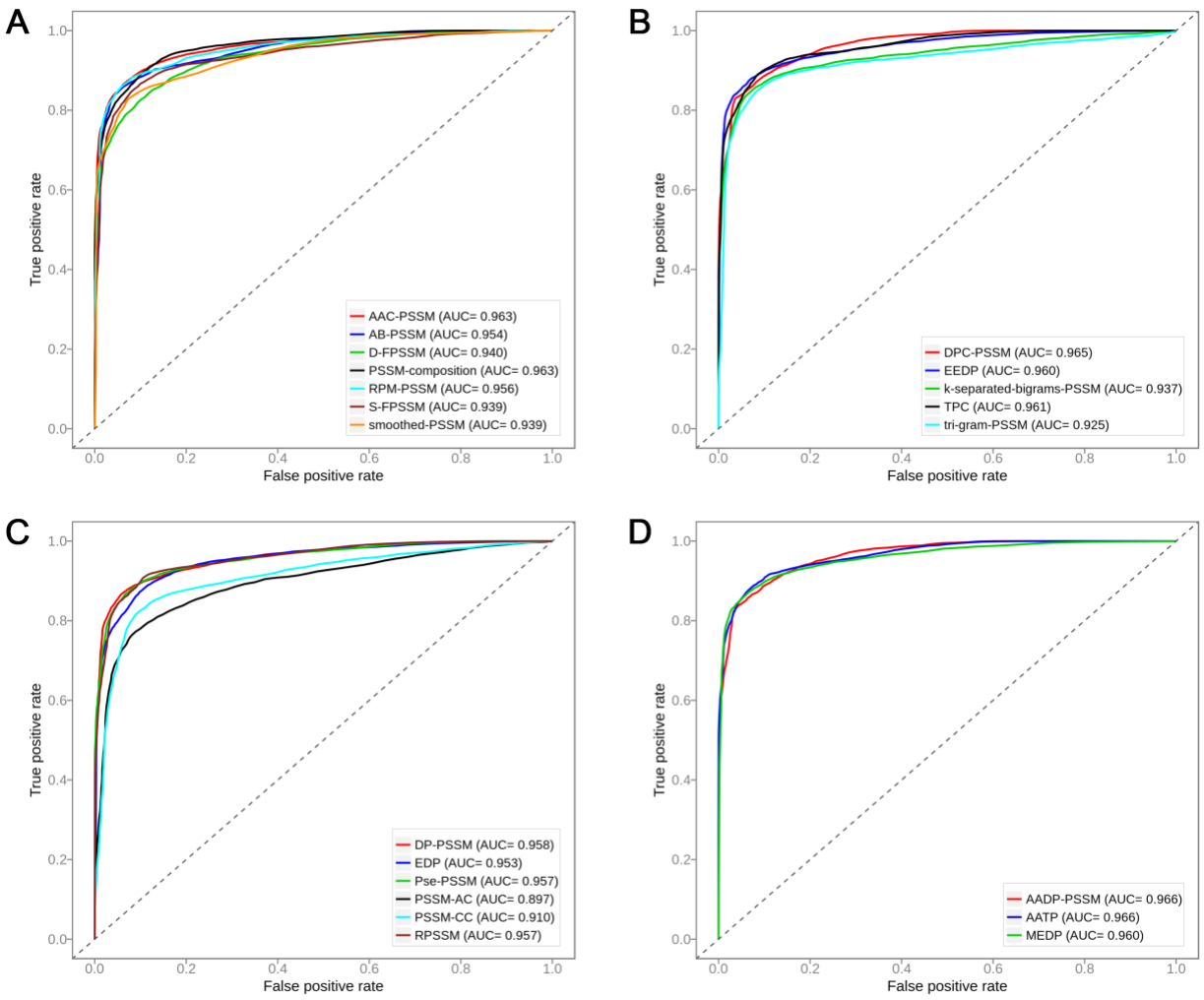

Fig. S8. Prediction performance of type IV secretion effectors using random forest classifiers, trained using multiple different feature descriptors generated by POSSUM as input features. The performance results were evaluated based on the 10 times randomization tests of 5-fold cross-validation. (A) ROC curves of random forest classifiers trained with feature descriptors within the row-transformation group; (B) ROC curves of random forest classifiers trained with feature descriptors within the column-transformation group; (C) ROC curves of random forest classifiers trained with feature descriptors within the mixture of row-transformation and column-transformation group, and (D) ROC curves of random forest classifiers trained with feature descriptors by combinations of rest groups.

For each type of PSSM-based features, the random forest classifier was trained and validated based on the 10-time randomization tests of 5-fold cross-validation. Respective results are shown in Table S2 and Fig. S8.

As can be observed from Table S2, PSSM-based descriptors performed much better when compared with sequence-based descriptors in terms of ACC, F-value and MCC scores. These results indicate that PSSM descriptors are much more informative, significantly contributing to the model performance. On the other hand, the RF classifiers trained using different types of PSSM-derived features achieved a varying performance, in terms of ACC (ranging from 0.842 to 0.904), F-value (ranging from 0.829 to 0.903) and MCC (ranging from 0.691 to 0.810), depending on the particular PSSM feature type used for training the RF models. The performance discrepancy implies that selection of optimal PSSM features that best suit the specific classification task should be exercised with caution. POSSUM is a tool that offers the opportunity to do the latter, by allowing interested users to address this technically challenging yet important question and meet their specific needs and facilitate their efforts to optimize the model performance within a homogenous framework. Statistically quantifying the contribution of various PSSM-based features to the prediction performance of the machine learning models is a relevant question of interest, as well as combining different feature selection techniques to identify a condensed subset of the most important PSSM features that collectively determine the model performance.

Furthermore, and rather surprisingly, certain uncommon (not well known) descriptors such as DP-PSSM and EEDP achieved reasonable performances. In contrast, some popular descriptors such as PSSM-AC and PSSM-CC performed poorly in this assessment (Fig. S8C). Taken together, we recommend that PSSM matrix transformations be a requisite for the application of POSSUM-calculated PSSM features to protein class classification and prediction tasks. In addition, various PSSM-based descriptors should be comprehensively assessed based on a well-prepared benchmark dataset for the purpose of identifying the best-performing descriptors. As can be seen from Fig. S8D, feature groups based on the combinations of other individual types of descriptors achieved a high and stable prediction performance, suggesting that the combinations of descriptors are likely to further improve the performance. This can be further validated and examined by assessing the performance of different approaches in a real application, e.g. protein classification (Nanni, et al., 2014). Nanni *et al*. reported that models trained based on the fusion of PSSM-based features and sequence-derived features could outperform those trained using only PSSM features. In summary, the application of PSSM-based features to the prediction of bacterial secreted effectors serves as a demonstration of the usefulness of POSSUM, and validates the need to develop and make available such tool to the wider research community.

Finally, it is worth mentioning that bioinformatics applications of the variety of PSSM-based feature descriptors that can be calculated by POSSUM need not be restricted to prediction of bacterial secretion effector proteins; in fact, these versatile and informative PSSM features can be applied to address a wide range of sequence-based classification tasks related to e.g. protein sequence analysis, remote homology detection, protein family prediction, protein structure and function prediction, in combination with other complementary features. We hope the new bioinformatics tool presented in this work, POSSUM, can be adopted as a useful starting point to develop more accurate predictors for bioinformatics' open questions.

References:

Ahmad, S. and Sarai, A. (2005) PSSM-based prediction of DNA binding sites in proteins, *BMC bioinformatics*, **6**, 33.

An, J.Y., *et al.* (2016) Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model, *Protein science : a publication of the Protein Society*.

Becker, J., Maes, F. and Wehenkel, L. (2013) On the encoding of proteins for disordered regions prediction, *PloS one*, **8**, e82252.

Bouziane, H., Messabih, B. and Chouarfia, A. (2011) Profiles and majority voting-based ensemble method for protein secondary structure prediction, *Evolutionary bioinformatics online*, **7**, 171-189.

Bui, V.M., *et al.* (2016) SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites, *BMC genomics*, **17 Suppl 1**, 9.

Bunnik, E.M., *et al.* (2011) Detection of inferred CCR5- and CXCR4-using HIV-1 variants and evolutionary intermediates using ultra-deep pyrosequencing, *PLoS pathogens*, **7**, e1002106.

Chauhan, J.S., *et al.* (2012) GlycoPP: a webserver for prediction of N- and O-glycosites in prokaryotic protein sequences, *PloS one*, **7**, e40155.

Chen, K., Kurgan, L.A. and Ruan, J. (2008) Prediction of protein structural class using novel evolutionary collocation‐based sequence representation, *Journal of computational chemistry*, **29**, 1596-1604.

Chen, S.A., *et al.* (2011) Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties, *Bioinformatics*, **27**, 2062-2067.

Chen, Z., *et al.* (2011) Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs, *PloS one*, **6**, e22930.

Chen, Z., *et al.* (2015) Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features, *Briefings in bioinformatics*, **16**, 640-657.

Cheng, C.W., *et al.* (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information, *BMC bioinformatics*, **9 Suppl 12**, S6.

Cheng, J. and Baldi, P. (2006) A machine learning information retrieval approach to protein fold recognition, *Bioinformatics*, **22**, 1456-1463.

Chou, K.C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, *Biochemical and biophysical research communications*, **278**, 477-483.

Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo‐amino acid composition, *Proteins: Structure, Function, and Bioinformatics*, **43**, 246-255.

Chou, K.C. and Shen, H.B. (2007) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, *Biochemical and biophysical research communications*, **360**, 339-345.

Cozzetto, D., *et al.* (2013) Protein function prediction by massive integration of evolutionary analyses and multiple data sources, *BMC bioinformatics*, **14**, S1.

Dehzangi, A., *et al.* (2013) A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem, *IEEE/ACM transactions on computational biology and bioinformatics*, **10**, 564-575.

Deng, L., *et al.* (2009) Prediction of protein-protein interaction sites using an ensemble method, *BMC bioinformatics*, **10**, 426.

Ding, S., *et al.* (2014) A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile, *Biochimie*, **97**, 60-65.

Dong, Q., Zhou, S. and Guan, J. (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, *Bioinformatics*, **25**, 2655-2662.

Gao, Z.G., *et al.* (2016) Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM, *BioMed research international*, **2016**, 4563524.

Garg, A. and Gupta, D. (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens, *BMC bioinformatics*, **9**, 62.

Guo, J., Lin, Y. and Liu, X. (2006) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins, *Proteomics*, **6**, 5099-5105.

Guo, Y., *et al.* (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, *Nucleic acids research*, **36**, 3025-3030.

Hayat, M. and Khan, A. (2012) MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM, *Journal of theoretical biology*, **292**, 93-102.

Hong, Y., *et al.* (2011) Predicting protein folds with fold-specific PSSM libraries, *PloS one*, **6**, e20557.

Huang, H.L., *et al.* (2011) Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties, *BMC bioinformatics*, **12 Suppl 1**, S47.

Huang, Y.F. and Chen, S.Y. (2013) Extracting physicochemical features to predict protein secondary structure, *TheScientificWorldJournal*, **2013**, 347106.

Hwang, S., Gou, Z. and Kuznetsov, I.B. (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins, *Bioinformatics*, **23**, 634-636.

Jensen, M.A., *et al.* (2006) A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences, *Journal of virology*, **80**, 4698-4704.

Jensen, M.A., *et al.* (2003) Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences, *Journal of virology*, **77**, 13376-13388.

Jeong, J.C., Lin, X. and Chen, X.W. (2011) On position-specific scoring matrix for protein function prediction, *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, **8**, 308-315.

Jiang, Y., *et al.* (2013) Prediction and Analysis of Post-Translational Pyruvoyl Residue Modification Sites from Internal Serines in Proteins, *PloS one*, **8**, e66678.

Jones, D.T. and Cozzetto, D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity, *Bioinformatics*, **31**, 857-863.

Jones, D.T. and Ward, J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices, *Proteins*, **53 Suppl 6**, 573-578.

Juan, E.Y., Jhang, J. and Li, W. (2008) Predicting protein subcellular localization using PsePSSM and suport vector machines. *Proceedings of the 11th Join Conference on Information Sciences*. pp. 1-6.

Juan, E.Y., *et al.* (2009) Predicting Protein Subcellular Localizations for Gram-Negative Bacteria using DP-PSSM and Support Vector Machines. *Complex, Intelligent and Software Intensive Systems, 2009. CISIS'09. International Conference on*. IEEE, pp. 836-841.

Kumar, M., Gromiha, M.M. and Raghava, G.P. (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles, *BMC bioinformatics*, **8**, 463.

Kumar, M., Gromiha, M.M. and Raghava, G.P. (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile, *Proteins*, **71**, 189-194.

Li, D., *et al.* (2012) A novel structural position-specific scoring matrix for the prediction of protein secondary structures, *Bioinformatics*, **28**, 32-39.

Li, F., *et al.* (2015) GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome, *Bioinformatics*, **31**, 1411-1419.

Li, L., *et al.* (2014) PSSP-RFE: accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations, *PloS one*, **9**, e92863.

Liu, T., *et al.* (2012) Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles, *Amino acids*, **42**, 2243-2249.

Liu, T., Zheng, X. and Wang, J. (2010) Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, *Biochimie*, **92**, 1330-1334.

Liu, Z.P., *et al.* (2010) Prediction of protein-RNA binding sites by a random forest method with combined features, *Bioinformatics*, **26**, 1616-1622.

Lobley, A., Sadowski, M.I. and Jones, D.T. (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination, *Bioinformatics*, **25**, 1761-1767.

Melo, R., *et al.* (2016) A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces, *International journal of molecular sciences*, **17**.

Mishra, N.K., Chang, J. and Zhao, P.X. (2014) Prediction of membrane transport proteins and their substrate specificities using primary sequence information, *PloS one*, **9**, e100278.

Murakami, Y. and Mizuguchi, K. (2010) Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites, *Bioinformatics*, **26**, 1841-1848.

Murakami, Y., *et al.* (2010) PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences, *Nucleic acids research*, **38**, W412-W416.

Nanni, L., Lumini, A. and Brahnam, S. (2014) An empirical study of different approaches for protein classification, *TheScientificWorldJournal*, **2014**, 236717.

Paliwal, K.K., *et al.* (2014) A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition, *IEEE transactions on nanobioscience*, **13**, 44-50.

Pu, X., *et al.* (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices, *Journal of theoretical biology*, **247**, 259-265.

Restrepo-Montoya, D., *et al.* (2011) NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins, *BMC bioinformatics*, **12**, 21.

Saini, H., *et al.* Protein Fold Recognition Using Genetic Algorithm Optimized Voting Scheme and Profile Bigram.

Seclen, E., *et al.* (2011) High concordance between the position-specific scoring matrix and geno2pheno algorithms for genotypic interpretation of HIV-1 tropism: V3 length as the major cause of disagreement, *Journal of clinical microbiology*, **49**, 3380-3382.

Selvaraj, M., *et al.* (2016) BacHbpred: Support Vector Machine Methods for the Prediction of Bacterial Hemoglobin-Like Proteins, *Advances in bioinformatics*, **2016**, 8150784.

Sharma, A., *et al.* (2013) A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition, *Journal of theoretical biology*, **320**, 41-46.

Shimizu, K., Hirose, S. and Noguchi, T. (2007) POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix, *Bioinformatics*, **23**, 2337-2338.

Su, C.T., Chen, C.Y. and Ou, Y.Y. (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder, *BMC bioinformatics*, **7**, 319.

Tang, Z., *et al.* (2011) Improving the performance of beta-turn prediction using predicted shape strings and a two-layer support vector machine model, *BMC bioinformatics*, **12**, 283.

Tao, P., *et al.* (2015) Prediction of protein structural class using tri-gram probabilities of position-specific scoring matrix and recursive feature elimination, *Amino acids*, **47**, 461-468.

Walia, R.R., *et al.* (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art, *BMC bioinformatics*, **13**, 1.

Wang, L., *et al.* (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, *BMC systems biology*, **4 Suppl 1**, S3.

Wang, X. and Li, G.-Z. (2013) Multilabel learning via random label selection for protein subcellular multilocations prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **10**, 436-446.

Wass, M.N. and Sternberg, M.J. (2008) ConFunc--functional annotation in the twilight zone, *Bioinformatics*, **24**, 798-806.

Xia, X.Y., *et al.* (2012) Accurate prediction of protein structural class, *PloS one*, **7**, e37653.

Xiao, N., *et al.* (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences, *Bioinformatics*, btv042.

Xie, D., *et al.* (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST, *Nucleic acids research*, **33**, W105-W110.

Yan, R., *et al.* (2015) Prediction of structural features and application to outer membrane protein identification, *Scientific reports*, **5**, 11586.

Yang, X., *et al.* (2013) Effective identification of Gram-negative bacterial type III secreted effectors using position-specific residue conservation profiles, *PloS one*, **8**, e84439.

Zahiri, J., *et al.* (2013) PPIevo: protein-protein interaction prediction from PSSM based evolutionary information, *Genomics*, **102**, 237-242.

Zhang, L., Zhao, X. and Kong, L. (2014) Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition, *Journal of theoretical biology*, **355**, 105-110.

Zhang, N., *et al.* (2014) Discriminating between lysine sumoylation and lysine acetylation using mRMR feature selection and analysis, *PloS one*, **9**, e107464.

Zhang, S., Ye, F. and Yuan, X. (2012) Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM, *Journal of biomolecular structure & dynamics*, **29**, 634-642.

Zou, L., Nan, C. and Hu, F. (2013) Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles, *Bioinformatics*, **29**, 3135-3142.