*Bioinformatics*, 33(17), 2017, 2756–2758
doi: 10.1093/bioinformatics/btx302
Advance Access Publication Date: 10 May 2017
Applications Note

OXFORD

Sequence analysis

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

**Jiawei Wang[1,†], Bingjiao Yang[2,†], Jerico Revote[1], André Leier[3], Tatiana T. Marquez-Lago[3], Geoffrey Webb[4], Jiangning Song[1,4,5,*], Kuo-Chen Chou[6,7,8] and Trevor Lithgow[1,*]**

[1]Biomedicine Discovery Institute, Monash University, VIC 3800, Australia, [2]College of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China, [3]Informatics Institute and Department of Genetics, School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA, [4]Monash Centre for Data Science, Faculty of Information Technology, [5]ARC Centre of Excellence for Advanced Molecular Imaging, Monash University, VIC 3800, Australia, [6]Gordon Life Science Institute, Boston, MA 02478, USA, [7]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China and [8]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

*To whom correspondence should be addressed.
[†]The authors wish it to be known that these authors contributed equally.
Associate Editor: John Hancock

Received on March 12, 2017; revised on April 14, 2017; editorial decision on April 30, 2017; accepted on May 9, 2017

## Abstract

**Summary:** Evolutionary information in the form of a Position-Specific Scoring Matrix (PSSM) is a widely used and highly informative representation of protein sequences. Accordingly, PSSM-based feature descriptors have been successfully applied to improve the performance of various predictors of protein attributes. Even though a number of algorithms have been proposed in previous studies, there is currently no universal web server or toolkit available for generating this wide variety of descriptors. Here, we present POSSUM (Position-Specific Scoring matrix-based feature generator for machine learning), a versatile toolkit with an online web server that can generate 21 types of PSSM-based feature descriptors, thereby addressing a crucial need for bioinformaticians and computational biologists. We envisage that this comprehensive toolkit will be widely used as a powerful tool to facilitate feature extraction, selection, and benchmarking of machine learning-based models, thereby contributing to a more effective analysis and modeling pipeline for bioinformatics research.
**Availability and implementation:** http://possum.erc.monash.edu/.
**Contact:** trevor.lithgow@monash.edu or jiangning.song@monash.edu
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Feature extraction or feature encoding is a fundamental step in the construction of high-quality machine learning-based models. Specifically, this step is key to determining the effectiveness of trained models in bioinformatics applications (Chou, 2011). In the last two decades, a variety of feature encoding schemes have been proposed in order to exploit useful patterns from protein sequences. Such schemes are often based on sequence information or representation of physicochemical properties. Although direct features derived from sequences themselves (such as amino acid compositions, dipeptide compositions and counting of *k*-mers) are regarded as essential for training models, an increasing number of studies

have shown that evolutionary information in the form of PSSM profiles is much more informative than sequence information alone (An *et al.*, 2016). Accordingly, PSSM-based feature descriptors have been commonly used as indispensable primary features to construct models, filling a major gap in the current bioinformatics research. For example, PSSM-based feature descriptors have successfully improved the prediction performance of structural and functional properties of proteins across a wide spectrum of bioinformatics applications (See Supplementary Table S1 in the Supplementary Material for a comprehensive lists of applications). These include for example protein fold recognition (Lobley *et al.*, 2009) and the prediction of protein structural classes (Liu *et al.*, 2010), protein-protein interactions (Zahiri *et al.*, 2013), protein subcellular localization (Xie *et al.*, 2005), RNA-binding sites (Cheng *et al.*, 2008) and protein functions (Radivojac *et al.*, 2013), to name a few.

A number of servers and standalone software packages have been developed to derive a variety of feature descriptors from protein, DNA and RNA sequences, including PROFEAT (Rao *et al.*, 2011), PseAAC (Shen and Chou, 2008), propy (Cao *et al.*, 2013), repDNA (Liu *et al.*, 2015), protr/ProtrWeb (Xiao *et al.*, 2015), Pse-in-One (Liu *et al.*, 2015; Liu *et al.*, 2017a), repRNA (Liu *et al.*, 2016) and Pse-Analysis (Liu *et al.*, 2017b). Despite their usefulness and popularity, these tools primarily focus on the generation of features related to sequence-based and/or physicochemical descriptors, instead of PSSM profile-based features. Indeed, there are over 20 different PSSM-based algorithms that calculate and model PSSM-based feature descriptors. However, to the best of our knowledge, there is currently no consolidated web server or toolkit available for generating these PSSM-based feature descriptors. Here, we present a bioinformatics toolkit, POSSUM, an effective tool that enables users to generate a broad spectrum of PSSM-based numerical representation schemes for protein sequences. It implements a wide range of algorithms available in the literature, provides an easy-to-use interface, and offers much needed functionality and flexibility for users to derive and customize these descriptors. We demonstrate the usage of POSSUM-calculated PSSM features for the prediction of bacterial secretion effector proteins (cf. Supplementary Material results).

## 2 Implementation

The POSSUM server consists of two major components: the client web interface and the server backend (See Supplementary Fig. S1). The former was implemented using jQuery, Bootstrap, Struts and Hibernate. Users can interact with the client web interface to input their protein sequences and choose the specific feature descriptors to be generated. Submitted jobs are then forwarded to the server backend. For the latter, a Perl CGI program lines up submitted jobs in a queue and invokes a Perl daemon thread for each job to execute the descriptor generation process. This architecture guarantees that multiple jobs can be executed simultaneously, within the maximum number of allowed threads predefined in the server, while any remaining jobs are queued until processing slots become available.

With the client web interface, users can upload a protein sequence file in the FASTA format, or directly input protein sequences (Supplementary Figs S2 and S3). Next, users need to customize parameters to generate PSSM profiles, which is followed by selection of the feature descriptors needed to be calculated. POSSUM generates PSSM profiles of the submitted sequences by running PSI-BLAST. Depending on the length of the input protein sequence, the PSSM profile generation process can be computationally time-consuming. To address this issue, we implemented a caching module in POSSUM, allowing re-use of generated PSSM profiles instead of computing them again. Based on the PSSM profiles, POSSUM can calculate the corresponding feature descriptors in the background inside the server backend. Users do not need to wait for job progress: they can track the progress of submitted jobs through a unique link, or be informed by email (if they opted for this in the client interface) once their jobs are finished. Both the raw PSSM files and resulting descriptors can then be downloaded from their unique link.

**Table 1.** A full list of PSSM-based feature descriptors that can be generated by POSSUM

| Descriptors groups | Descriptor | Number | Original |
|---|---|---|---|
| Row transformations | AAC-PSSM | 20 | (Liu *et al.*, 2010) |
| | D-FPSSM | 20 | (Zahiri *et al.*, 2013) |
| | smoothed-PSSM | –[a] | (Cheng *et al.*, 2008) |
| | AB-PSSM | 400 | (Jeong *et al.*, 2011) |
| | PSSM-composition | 400 | (Zou *et al.*, 2013) |
| | RPM-PSSM | 400 | (Jeong *et al.*, 2011) |
| | S-FPSSM | 400 | (Zahiri *et al.*, 2013) |
| Column transformations | DPC-PSSM | 400 | (Liu *et al.*, 2010) |
| | k-separated-bigrams-PSSM | 400 | (Saini *et al.*, 2016) |
| | tri-gram-PSSM | 8000 | (Paliwal *et al.*, 2014) |
| | EEDP | 400 | (Zhang *et al.*, 2014) |
| | TPC | 400 | (Zhang *et al.*, 2012) |
| Mixture of row and column transformations | EDP | 20 | (Zhang *et al.*, 2014) |
| | RPSSM | 110 | (Ding *et al.*, 2014) |
| | Pse-PSSM | 40 | (Chou and Shen, 2007) |
| | DP-PSSM | –[a] | (Juan *et al.*, 2009) |
| | PSSM-AC | –[a] | (Dong *et al.*, 2009) |
| | PSSM-CC | –[a] | (Dong *et al.*, 2009) |
| Combination of above descriptors | AADP-PSSM | 420 | (Liu *et al.*, 2010) |
| | AATP | 420 | (Zhang *et al.*, 2012) |
| | MEDP | 420 | (Zhang *et al.*, 2014) |

[a]The number of feature descriptor values depends on the choice of the parameter.

For users who prefer to apply their own parameter settings for specific research purposes and users who have the capacity to perform high throughput generation of PSSM files for a very large dataset using their local computers, an open source standalone software toolkit is also available. The standalone version of POSSUM (See Supplementary Fig. S4) was developed using Python and Perl, and can be executed on Unix/Linux, Windows and Mac OS. As an open source software, users can access, modify and redistribute the source codes, allowing users to tailor POSSUM according to their specific requirements.

PSSM-based algorithms are based on matrix transformations from original PSSM profiles, which can be categorized into three types: row transformations, column transformations, or a mixture of row and column transformations. For POSSUM, these descriptors are divided into four groups (Table 1). The first group consists of AAC-PSSM, D-FPSSM, smoothed-PSSM, AB-PSSM, PSSM-composition, RPM-PSSM and S-FPSSM, which are generated by row transformations of the original PSSM. The second group contains the descriptors generated by column transformations, including DPC-PSSM, k-separated-bigrams-PSSM, tri-gram-PSSM, EEDP and TPC. The third group includes EDP, RPSSM, Pse-PSSM, DP-PSSM, PSSM-AC and PSSM-CC, which are generated by a mixture of row and column transformations. The fourth group comprises of AADP-PSSM, AATP and MEDP, which simply combine descriptors in the former three groups.

## 3 Results

In this work, we present POSSUM, a comprehensive, flexible, user-friendly and publicly accessible toolkit (with both local standalone software and online webserver) that we developed to allow users to easily generate more than 20 types of PSSM profile-based feature descriptors. It greatly facilitates feature generation, analysis, training and benchmarking of machine-learning models and predictions. POSSUM has been extensively benchmarked to guarantee correctness of computations, and was deliberately designed to ensure workflow efficiency. To the best of our knowledge, this is the first toolkit for generating such a great variety of evolutionary feature descriptors. Future work will include parallelization of PSSM profile generation to improve the throughput of POSSUM server. POSSUM is freely accessible at http://possum.erc.monash.edu/.

## Acknowledgements

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC) (1092262) and the Australian Research Council (ARC). G.I.W. is a recipient of Discovery Outstanding Research Award (DORA) of the ARC. T.L. is an ARC Australian Laureate Fellow.

*Conflict of Interest*: none declared.

## References

An,Y. *et al.* (2016) Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. *Brief Bioinform.*, bbw100.

Cao,D.S. *et al.* (2013) Propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **29**, 960–962.

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

Chou,K.-C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.*, **273**, 236–247.

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. *Biochem. Biophys. Res. Commun.*, **360**, 339–345.

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. *et al.* (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics*, **25**, 2655–2662.

Jeong,J.C. *et al.* (2011) On position-specific scoring matrix for protein function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform./IEEE, ACM*, **8**, 308–315.

Juan,E.Y. *et al.* (2009) Predicting protein subcellular localizations for gram-negative bacteria using dp-pssm and support vector machines. In: *International Conference on Complex, Intelligent and Software Intensive Systems, 2009. CISIS'09*, pp. 836–841. IEEE Press, Fukuoka, Japan.

Liu,B. *et al.* (2015) repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, **31**, 1307–1309.

Liu,B. *et al.* (2016) repRNA: a web server for generating various feature vectors of RNA sequences. *Mol. Genet. Genom. MGG*, **291**, 473–481.

Liu,B. *et al.* (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–W71.

Liu,B. *et al.* (2017a) Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Natural Science*, **9**, 67.

Liu,B. *et al.* (2017b) Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget*, **8**, 13338–13343.

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

Lobley,A. *et al.* (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, **25**, 1761–1767.

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 Trans. Nanobiosci.*, **13**, 44–50.

Radivojac,P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.

Rao,H.B. *et al.* (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **39**, W385–W390.

Saini,H. *et al.* (2016) Protein fold recognition using genetic algorithm optimized voting scheme and profile bigram. *J Softw.*, **11**, 756–767.

Shen,H.B. and Chou,K.C. (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **373**, 386–388.

Xiao,N. *et al.* (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, **31**, 3555–3557.

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

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

Zhang,L. *et al.* (2014) Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.*, **355**, 105–110.

Zhang,S. *et al.* (2012) Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM. *J. Biomol. Struct. Dyn.*, **29**, 634–642.

Zou,L. *et al.* (2013) Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics*, **29**, 3135–3142.