EvoStruct-Sub: An Accurate Gram-positive Protein Subcellular Localization Predictor Using Evolutionary and Structural Features

Md. Raihan Uddin, Alok Sharma, Dewan Md Farid,
Md. Mahmudur Rahman, Abdollah Dehzangi, Swakkhar Shatabda

PII: S0022-5193(18)30055-9
DOI: 10.1016/j.jtbi.2018.02.002
Reference: YJTBI 9345

To appear in: *Journal of Theoretical Biology*

Received date: 30 December 2017
Revised date: 18 January 2018
Accepted date: 3 February 2018

Please cite this article as: Md. Raihan Uddin, Alok Sharma, Dewan Md Farid, Md. Mahmudur Rahman, Abdollah Dehzangi, Swakkhar Shatabda, EvoStruct-Sub: An Accurate Gram-positive Protein Subcellular Localization Predictor Using Evolutionary and Structural Features, *Journal of Theoretical Biology* (2018), doi: 10.1016/j.jtbi.2018.02.002

Highlights

- Structural Features for Classification
- Evolutionary Features for Classification
- E□ective Prediction
- Feature Selection

EvoStruct-Sub: An Accurate Gram-positive Protein Subcellular Localization Predictor Using Evolutionary and Structural Features

Md. Raihan Uddin[1], Alok Sharma[2,3,5], Dewan Md Farid[1], Md. Mahmudur Rahman[4], Abdollah Dehzangi[4,*] and Swakkhar Shatabda[1,*]

[1] *Department of Computer Science and Engineering, United International University, Bangladesh*
[2] *School of Engineering and Physics, University of the South Pacific, Fiji*
[3] *Institute for Integrated and Intelligent Systems, Griffith University, Australia*
[4] *Department of Computer Science, Morgan State University, United States*
[5] *RIKEN Center for Integrative Medical Sciences, Yokohama, Japan*
[*] *Corresponding authors*

Abstract

Determining subcellular localization of proteins is considered as an important step towards understanding their functions. Previous studies have mainly focused solely on Gene Ontology (GO) as the main feature to tackle this problem. However, it was shown that features extracted based on GO is hard to be used for new proteins with unknown GO. At the same time, evolutionary information extracted from Position Specific Scoring Matrix (PSSM) have been shown as another effective features to tackle this problem. Despite tremendous advancement using these sources for feature extraction, this problem still remains unsolved. In this study we propose EvoStruct-Sub which employs predicted structural information in conjunction with evolutionary information extracted directly from the protein sequence to tackle this problem. To do this we use several different feature extraction method that have been shown promising in subcellular localization as well as similar studies to extract effective local and global discriminatory information. We then use Support Vector Machine (SVM) as our classification technique to build EvoStruct-Sub. As a result, we are able to enhance Gram-positive subcellular localization prediction accuracies by up to 5.6% better than previous studies including the studies that used GO for

feature extraction.

Keywords: Proteins subcellular localization, Evolutionary-based Features, Structural-based Features, Classification, Support Vector Machine, Feature Selection

1. INTRODUCTION

The functioning of a protein depends on its location in the cell. In fact, it just functions properly in one or a few locations in the cell. Knowing those locations can provide important information about functioning of the proteins and how they interact with other micro-molecules. Therefore, determining protein subcellular localization is considered as an important step towards understanding its functioning [1, 2].

Of all proteins, bacterial proteins are the most important proteins to determine their functions because of their biological aspects which are both harmful and useful [3]. Some bacteria can cause a wide range of diseases while some others play the role of catalyst in biological interactions. Some bacteria are also widely used to produce antibiotics. Bacteria are categorized as a kind of prokaryotic microorganism that can be divided in two groups, Gram-positive and Gram-negative [4]. Gram-positive bacteria are those that are stained dark blue or violet by Gram-staining while Gram-negative bacteria cannot retain the stain, instead taking up the counter-stain and appearing red or pink [3].

During the past two decades and since the introduction of bacterial proteins subcellular localization, a wide range of machine learning methods with many different combination and types of features have been proposed to tackle this problem [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 24, 25, 26]. For example, PSORT (predictor) used sequence features based on sorting signal [27], SubLoc (predictor) uses SVM with AAC to obtain higher accuracy [28], and TargetP (predictor) uses ANN and N-terminal sequence to predict subcellular locations [29]. In addition Pierleoni et al. used N-terminal, AAC and alignment profile to predict the subcellular localization [30]. Similarly, Tamura

and Akutsu used alignment of block sequence [31] and Chang et al. developed and used gapped-dipeptide and probabilistic latent semantic analysis method for prediction of Gram-negative bacterial protein [32]. Lee et al. Despite all the e□orts have been made so far, the protein subcellular localization prediction problem for bacterial proteins have remained unsolved.

As pointed in a recent review [33], in the last decade or so, a number of web-servers were also developed for predicting the subcellular localization of proteins with both single site and multiple sites based on their sequences information alone. They can be roughly classified into two series [34]. One is the PLoc series and the other is iLoc series. The PLoc series contains the six web-servers [4], [35], [36], [37], [38], [39] to deal with eukaryotic, human, plant, Gram-positive, Gram-negative, and virus proteins, while the iLoc series contains the seven web-servers [40], [41], [42] , [43], [3], [40], [44] to deal with eukaryotic, human, plant, animal, Gram-positive, Gram-negative, and virus proteins, respectively.

In addition, Huang and Yuan analyzed series of classifiers for subcellular localization, but these were limited to single location site. For multi label prediction, Gpos-mplock and Gneg-mplock (predictor) are proposed [36], [38] to predict protein localization in Gram-positive and Gram-negative bacteria; and Plant-mploc (predictor) is developed [37] which uses top down strategy to predict single or multiple protein localization in plant protein. Virus-mploc (predictor) [39] was developed with fusion of classifiers and features of functional domain and gene ontology to predict virus proteins. To increase the quality of prediction, three revised version of the prediction systems were developed: iloc-Gpos (predictor) [3], iloc-plant (predictor) [42], iloc-virus (predictor) [44]. Huang and Yuan used AAC, evolution information and PseACC with backward propagation (BP) and radial basis function (RBF) neural network to predict both single and multi-site subcellular proteins.

Many of those studies that have mentioned earlier relied on Gene Ontology as their feature to tackle this problem [45, 46, 47, 48, 49]. Despite promising results achieved using GO, it is hard or even for some cases impossible to use these features with new proteins with unknown GO. Therefore, there is an

emphasis on proposing methods that rely on features directly extracted from protein sequence without using any other extra information or meta data that are available for just known proteins.

Early studies have focused on using features that are extracted from the occurrence of amino acids from the protein sequence [50, 45, 51]. Later studies try to incorporate physicochemical based features to enhance the prediction performance [52]. However, the protein subcellular localization prediction problem remained limited using these sources of features.

More recent studies have started using evolutionary information to tackle this problem [2, 53, 54, 5, 7, 55, 6]. Using these features they demonstrated significant enhancement and even achieved comparable performance compared to use of GO as input feature. In fact, application of evolutionary based features have demonstrated its superiority over using occurrence or physicochemical based features in many similar studies found in the literature [56, 57, 58, 59, 60, 61]. However, further enhancement have remained out of reach relying on these features. In addition, many of those studies tried to address Gram-positive and Gram-negative subcellular localization at the same time. Despite lots of similarities, still they have their own di□erences in nature based on their biological properties. Therefore, similar to all the other subcellular-localization prediction problems, a well designed method that is tailored for that specific task (either Gram-positive or Gram-negative) has a better chance to achieve more promising results.

To develop a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications [20, 21, 22, 23, 24, 33, 12, 62, 63, 64, 65, 66], one should observe the Chou's 5-step rule [67]; i.e., making the following five steps very clear: (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the biological sequence samples with an e□ective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) how to properly perform cross-validation tests to objectively

evaluate the anticipated accuracy of the predictor; (v) how to establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by- one.

To address these problems, here we propose EvoStruct-Sub which in addition to evolutionary information uses predicted structural information to specifically predict Gram-positive subcellular localization. To do this, we first extract evolutionary information from Position Specific Scoring Matrix (PSSM) [68] and predicted secondary structure using SPDER 2.0 [69, 70]. We then extract global and local discriminatory information using segmentation technique for our classification task [71, 72]. We finally use Support Vector Machine (SVM) to our extracted features to build EvoStruct-Sub. By applying EvoStruct-Sub to Gram-positive subcellular localization, we achieve up to 95.4% prediction accuracy for this task. In addition, we achieve to over 90.0% prediction accuracy for this problem when we use our method for multi-label samples. These results are over 5.0% better than previously reported results found in the literature [73, 5, 6].

2. MATERIALS AND METHODS

In this section, we describe the materials and methods required to develop EvoStruct-Sub.

2.1. Benchmark Dataset

In this research we use a dataset which have been used widely in literature for Gram-positive subcellular localization [74], [75], [76], [77], [78]. The benchmark that we use in this study was introduced in [74], [75], [76], [77]. This benchmark contains total 523 protein samples which belongs to four Gram-positive subcellular localizations. Among this 523 samples there are total 519 different protein sample. Among 519 proteins there are total 515 protein samples which belongs to only one or single location while the rest 4 protein samples belongs to two locations. Thus Gram-positive bacterial protein benchmark contains total 523 (515 + 4 * 2) protein samples. The name and the number of

proteins in these four locations are shown in Table 1. This dataset is available at: http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi.

Table 1: Details of Gram-positive bacterial proteins' dataset

No.	subcellular location	Total protein samples
1	Cell membrane	174
2	Cell wall	18
3	Cytoplasm	208
4	Extracellular	123
Total number of locative proteins		523
Total number of different proteins		519

2.2. Feature Extraction

We extract evolutionary information from Position Specific Scoring Matrix (PSSM) [68] and structural information from the SPIDER 2.0 [70, 69].

PSSM which is produced as the output of PSIBLAST is an scoring matrix that provide substitution probabilities of a given amino acid with other amino acids based on its specific position in a protein[68]. PSSM is a $L \times 20$ matrix, where L is the length of the input protein. Here we use PSIBLAST with three iteration and its cut off value (E) set to 0.001 to produce PSSM .

SPIDER 2.0 (Scoring Protein Interaction Decoys using Exposed Residues) is an accurate method that predict different aspects of local structure such as secondary structure, torsion angle, and Accessible Surface Area (ASA), simultaneously [70, 69]. As an output it produces a $L \times 8$ matrix that include three columns of the probability of contribution of amino acids to each of the secondary structure elements (α-helix, β-strand, coil), one column for ASA, and four columns for the torsion angles (φ, ψ, θ, τ) [79, 80]. For the rest of this paper, we will refer to this matrix as SPD3 for simplicity. SPD3 has been recently used in many different fields and demonstrated promising results [81, 82, 83, 84].

Here we have extracted a wide range of features based on different concepts that have been investigate in the literature both for PSSM and SPD3. As a

result the combination of 6 feature groups attained the best result for our task. This might be due to the consistency of these sets of features with each other. However, further investigations in future can potentially provide better understanding of available discriminatory information in these feature groups and consequently provide further prediction enhancement. These 6 feature groups are explained in detail in the following sections.

2.2.1. PSSM-AAO Feature

PSSM-AAO is amino acid occurrence based on PSSM. This feature group is directly extracted from PSSM matrix. It aims at capturing global discriminatory information regarding the substitution probabilities of the amino acids with respect to their positions in the protein sequence [85, 86, 87]. This feature is extracted by summation of the substitution score of a given amino acid with all the amino acids along the protein sequence. The equation for this feature is given below:

$$PSSM - AAO_j = \sum_{i=1}^{L} N_{i,j} \quad (j = 1, ..., 20) \tag{1}$$

Here N is the corresponding matrix, L is the protein length and j is the respective column. The dimensionality of this feature vector is 20. Algorithm for extracting composition feature is shown in Algorithm 1.

2.2.2. PSSM-SD Feature

This method is specifically proposed to add more local discriminatory information about how the amino acids, based on their substitution probabilities (extracted from PSSM), are distributed along the protein sequence [88]. This method is explained in detail in [71] [5].

Algorithm for extracting PSSM-SD feature is shown in Algorithm 2.

As shown in [5] using $F_p = 25$ gives the best result for this method. As a result, in our final experiment we have adopted $F_p = 25$ which produces 80 ((100 ÷ 25) × 20 = 80) features.

Algorithm 1: PSSM-AAO Feature Extraction

```
N ← PSSM Matrix;
L ← Length of the Protein;
C ← Number of matrix column;
V ← Empty array of size C;
for j = 0; j < C; j = j + 1 do
    sum ← 0;
    for i = 0; i < L; i = i + 1 do
        sum = sum + N_{i,j};
    end
    V_j = sum;
end
```

2.2.3. PSSM-SAC Feature

This feature was introduced in [5, 89, 86]. It was shown that information about the interaction of neighboring amino acids along the protein sequence can play an important role in providing significant local discriminatory information and enhancing protein subcellular localization prediction accuracy [6, 87]. To extract this information, the concept of auto covariance has been used for different segments of proteins. This is done to enforce local discriminatory information extracted from PSSM. We use the similar approach that was adopted and explained in [5]. We also use the distance factor of 10 as it was also shown in this study as the most effective parameter to extract features for protein subcellular localization.

2.2.4. Auto Covariance of Predicted Secondary Structure

A correlation factor coupling adjacent residues along the protein sequence is known as Auto covariance (AC) [59, 90, 72]. It is also known as a kind of variant

[90] A. Dehzangi, A. Sattar, Ensemble of diversely trained support vector machines for protein fold recognition., in: ACIIDS (1), 2013, pp. 335–344.

[91] S. Wold, J. Jonsson, M. Sjörström, M. Sandberg, S. Rännar, Dna and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures, Analytica Chimica Acta 277 (2) (1993) 239–253.

[92] Y. Guo, M. Li, M. Lu, Z. Wen, Z. Huang, Predicting g-protein coupled receptors–g-protein coupling specificity based on autocross-covariance transform, Proteins: structure, function, and bioinformatics 65 (1) (2006) 55–60.

[93] Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences, Nucleic acids research 36 (9) (2008) 3025–3030.

[94] Q. Dong, S. Zhou, J. Guan, A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, Bioinformatics 25 (20) (2009) 2655–2662.

[95] J. Wu, M.-L. Li, L.-Z. Yu, C. Wang, An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition, The protein journal 29 (1) (2010) 62–67.

[96] Y.-h. Zeng, Y.-z. Guo, R.-q. Xiao, L. Yang, L.-z. Yu, M.-l. Li, Using the augmented chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, Journal of theoretical biology 259 (2) (2009) 366–372.

[97] T. Liu, X. Zheng, C. Wang, J. Wang, Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation, Protein and peptide letters 17 (10) (2010) 1263–1269.

[98] G. Taherzadeh, Y. Zhou, A. W.-C. Liew, Y. Yang, Structure-based prediction of protein–peptide binding regions using random forest, Bioinformatics.

[99] K.-C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Current topics in medicinal chemistry 17 (21) (2017) 2337–2358.

[100] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, K.-C. Chou, Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition, Analytical biochemistry 456 (2014) 53–60.

[101] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.-C. Chou, Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences, Nucleic acids research 43 (W1) (2015) W65–W71.

[102] W. Chen, H. Lin, K.-C. Chou, Pseudo nucleotide composition or pseknc: an e□ective formulation for analyzing genomic sequences, Molecular BioSystems 11 (10) (2015) 2620–2634.

[103] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 2013.

[104] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST) 2 (3) (2011) 27.

[105] X. Cheng, X. Xiao, K.-C. Chou, ploc-meuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key go information into general pseaac, Genomicsdoi:doi:10.1016/j.ygeno.2017.08.

[106] X. Cheng, S.-G. Zhao, X. Xiao, K.-C. Chou, iatc-misf: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (3) (2016) 341–346.

[107] X. Cheng, S.-G. Zhao, X. Xiao, K.-C. Chou, iatc-mhyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (35) (2017) 58494.

[108] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, K.-C. Chou, iptm-mlys: identifying multiple lysine ptm sites and their di□erent types, Bioinformatics 32 (20) (2016) 3116–3123.

[109] K.-C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Molecular Biosystems 9 (6) (2013) 1092–1100.

[110] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, isno-aapair: incorporating amino acid pairwise coupling into pseaac for predicting cysteine s-nitrosylation sites in proteins, PeerJ 1 (2013) e171.

[111] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, irspot-psednc: identify recombination spots with pseudo dinucleotide composition, Nucleic acids research 41 (6) (2013) e68–e68.

[112] W. Chen, H. Tang, J. Ye, H. Lin, K.-C. Chou, irna-pseu: Identifying rna pseudouridine sites, Molecular TherapyNucleic Acids 5 (7) (2016) e332.

[113] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.-C. Chou, irna-psecoll: Identifying the occurrence sites of di□erent rna modifications by incorporating collective e□ects of nucleotides into pseknc, Molecular Therapy-Nucleic Acids 7 (2017) 155–163.

[114] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.-C. Chou, irna-ai: identifying the adenosine to inosine editing sites in rna sequences, Oncotarget 8 (3) (2017) 4208.

[115] B. Liu, S. Wang, R. Long, K.-C. Chou, irspot-el: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (1) (2016) 35–41.

[116] Y. Xu, Z. Wang, C. Li, K.-C. Chou, ipreny-pseaac: identify c-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into pseaac, Medicinal Chemistry 13 (6) (2017) 544–551.

[117] L.-M. Liu, Y. Xu, K.-C. Chou, ipgk-pseaac: identify lysine phosphoglycerylation sites in proteins by incorporating four di□erent tiers of amino acid pairwise coupling information into the general pseaac, Medicinal Chemistry 13 (6) (2017) 552–559.

[118] A. Dehzangi, S. Sohrabi, R. He□ernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features, BMC bioinformatics 16 (4) (2015) S1.

[119] K.-C. Chou, H.-B. Shen, Recent advances in developing web-servers for predicting protein attributes, Natural Science 1 (02) (2009) 63.