Chapter 10

LabCaS for Ranking Potential Calpain Substrate Cleavage Sites from Amino Acid Sequence

Yong-Xian Fan, Xiaoyong Pan, Yang Zhang, and Hong-Bin Shen

Abstract

Calpains are a family of Ca^{2+}-dependent cysteine proteases involved in many important biological processes, where they selectively cleave relevant substrates at specific cleavage sites to regulate the function of the substrate proteins. Presently, our knowledge about the function of calpains and the mechanism of substrate cleavage is still limited due to the fact that the experimental determination and validation on calpain bindings are usually laborious and expensive. This chapter describes LabCaS, an algorithm that is designed for predicting the calpain substrate cleavage sites from amino acid sequences. LabCaS is built on a conditional random field (CRF) statistic model, which trains the cleavage site prediction on multiple features of amino acid residue preference, solvent accessibility information, pair-wise alignment similarity score, secondary structure propensity, and physical-chemistry properties. Large-scale benchmark tests have shown that LabCaS can achieve a reliable recognition of the cleavage sites for most calpain proteins with an average AUC score of 0.862. Due to the fast speed and convenience of use, the protocol should find its usefulness in large-scale calpain-based function annotations of the newly sequenced proteins. The online web server of LabCaS is freely available at http://www.csbio.sjtu.edu.cn/bioinf/LabCaS.

Key words Protease substrate recognition, Cleavage site prediction, Sequence labeling, Ensemble learning, Calpain, Conditional random fields

1 Introduction

1.1 Background

Calpains are an important family of Ca^{2+}-regulated cysteine proteases that serve to regulate gene expression, signal transduction, cell death, and apoptosis [1, 2] through interacting and cleaving their substrates [3]. Many previous studies have shown that calpains are involved in pathophysiological mechanisms, and as a result, their malfunctions may lead to the occurrence of various diseases, such as muscular dystrophies, diabetes, and tumorigenesis [4, 5]. Although several promising calpain-targeted therapeutic strategies have been reported, there are still remaining challenges for drug discovery [4] due to the fact that the mechanisms of calpain modulation and the cleavage specificity of calpain substrates are still unclear.

Jeannette S. Messer (ed.), *Calpain: Methods and Protocols*, Methods in Molecular Biology, vol. 1915,
https://doi.org/10.1007/978-1-4939-8988-1_10,

As one of the most important functions of calpain is recognizing and cleaving their substrates, knowing the exact positions of the corresponding substrates' cleavage sites is very important for understanding their working mechanisms because the locations of the cleavage sites are closely related to how calpains will modulate substrate functions [6]. Although the cleavage sites can be determined based on various wet-lab experimental approaches, the process is very laborious and time-consuming to test all the residues throughout the substrate sequence. Considering the fact that the number of known protein sequences has been exponentially increasing with the development of modern sequencing technologies, automated computational tools are highly desired to screen out the most probable cleavage sites from the whole sequence space; this would reduce the search space and provide a small subset of high confidence for further experimental verification. Motivated by this, the development of calpain-related bioinformatics tools has been an important area of research, and many databases [7–11] and computational methods [12–18] have been reported in the literature.

1.2 Public Databases Related with Calpain

With the efforts of past decades, many calpain-based data resources have been constructed that are publicly available and can be accessed through the Internet. Table 1 lists some of these resources. For instance, the MEROPS database is a manually curated resource for proteolytic enzymes, their inhibitors, and substrates [7]. The CaMPDB database consists of three resources: calpain, substrates, and calpastatin [8]. The ENZYME is another repository of the nomenclature of enzymes where a large amount of calpain-related data exists. This database primarily describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been attached [9]. The CutDB focuses on the annotations of individual proteolytic events, both actual and computationally predicted [10], while PMAP provides physiological substrate cleavage data and pathways and networks for proteolytic-related systems [11].

1.3 Publicly Available Bioinformatics Predictors Related with Calpain

Apart from the efforts of collecting calpain-related data from various large biological pools, another branch of bioinformatics has dealt with the development of intelligent predictors, which have been employed for the prediction of the most probable targets utilized in wet-lab experiments; this helps in the reduction of both time and expense. For instance, they can be applied to rank the most potential cleavage sites from the entire substrate amino acid sequence. Most of these substrate cleavage site predictors will need an annotated dataset, which is referred to as training set and consists of the substrate sequences with well-annotated true cleavage sites. Subsequently, four general categories of techniques can be used to construct the cleavage site ranking predictors: (1) statistical

Table 1
Some publicly available calpain-related databases

Database name	Description	Website	Reference
MEROPS	Database of proteolytic enzymes, their substrates and inhibitors	https://www.ebi.ac.uk/merops	[7]
CaMPDB	Resource for calpain and modulatory proteolysis	http://calpain.org	[8]
ENZYME	Enzyme nomenclature database	http://enzyme.expasy.org	[9]
CutDB	Proteolytic event database	http://cutdb.burnham.org	[10]
PMAP	Databases for analyzing proteolytic events and pathways	http://pmap.burnham.org/proteases	[11]

Table 2
The state-of-the-art predictors for calpain substrate cleavage sites

Predictors	Description	Address	Reference
Preference matrix	Sequential determinants of calpain cleavage	/	[12]
PoPS	Modeling and predicting protease specificity	http://pops.csse.monash.edu.au	[13]
Site prediction	Predicting the cleavage of proteinase substrates	http://www.dmbr.ugent.be/prx/bioit2-public/SitePrediction	[14]
GPS-CCD	Predicting calpain cleavage sites	http://ccd.biocuckoo.org	[15]
SVM (MKL)	Calpain cleavage site prediction	http://calpain.org	[16]
LabCaS	Labeling calpain substrate cleavage sites	http://www.csbio.sjtu.edu.cn/bioinf/LabCaS	[17]
Binary QSAR model	Prediction of cleavability of calpain proteolysis	/	[18]

propensity score-based method, (2) machine learning-based cleavage/non-cleavage two-class classifier, (3) the sequential learning-based labeling method, and (4) the quantitative structure activity-based method.

In the propensity score method, the amino acid propensities around the true cleavage sites are first calculated according to the training dataset. In the following step, a total score of a peptide segment in a pre-defined fixed-size sliding window is obtained that is judged with a derived optimal threshold to determine whether the central residue located in the segment is cleavable or not. The advantage of such a propensity score-based approach is that it is

very straightforward and fast. However, it is very sensitive to the size of the dataset. In general, a larger training set generates more reliable propensity statistics, while an optimal threshold often varies with the different dataset. In the machine learning-based two-class classification approach, the positive and negative (cleavable and non-cleavable, respectively) subsets are first structured by partitioning the whole dataset. The two subsets are then accessed by machine learning-based models, i.e., artificial neural networks (ANN) and support vector machine (SVM), to learn the classification rules. After this training procedure, a trained classifier is used for the prediction for a new query, which is not contained in the training set. The merit of the two-class classifiers is that they can partially reduce the small sample size effects. However, the classification performance can be significantly affected by the imbalance between positive and negative samples. The sequential learning-based method, which is suitable for sequential labeling and classification problems, is often insensitive to the ratio between positive and negative training subsets, and hence all the negative samples can be used to establish the labeling model that can avoid information loss. When the substrate proteins have solved structures, the method based on quantitative structure-based activity analysis can be applied. Different state-of-the-art computational methods that are available for calpains analysis are tabulated in Table 2.

Here, we have adopted the typical conditional random field (CRF) sequential learning model and developed a label calpain substrate cleavage site predictor (LabCaS). The LabCaS tool has integrated multiple sequence-derived features to enhance the prediction performance. LabCaS is freely accessible at http://www.csbio.sjtu.edu.cn/bioinf/LabCaS. Through web browsers such as Safari, Internet Explorer, and Firefox, users can use the web server by providing the sequence of query proteins in FASTA format as input. For each query sequence, LabCaS ranks the potential cleavage sites and gives the confidence scores belonging to the cleavage sites.

2 Methods for Constructing LabCaS

The LabCaS predictor can directly label the cleavable residues from the entire sequence using CRF algorithm, where five different sequential features are integrated. When a sequence is submitted to the LabCaS tool, it first extracts the features from the sequence that are fed into the CRF model for prediction.

2.1 Extracted Features for Prediction

Five different sequential features are extracted from the sequences:

1. Amino acid preference feature

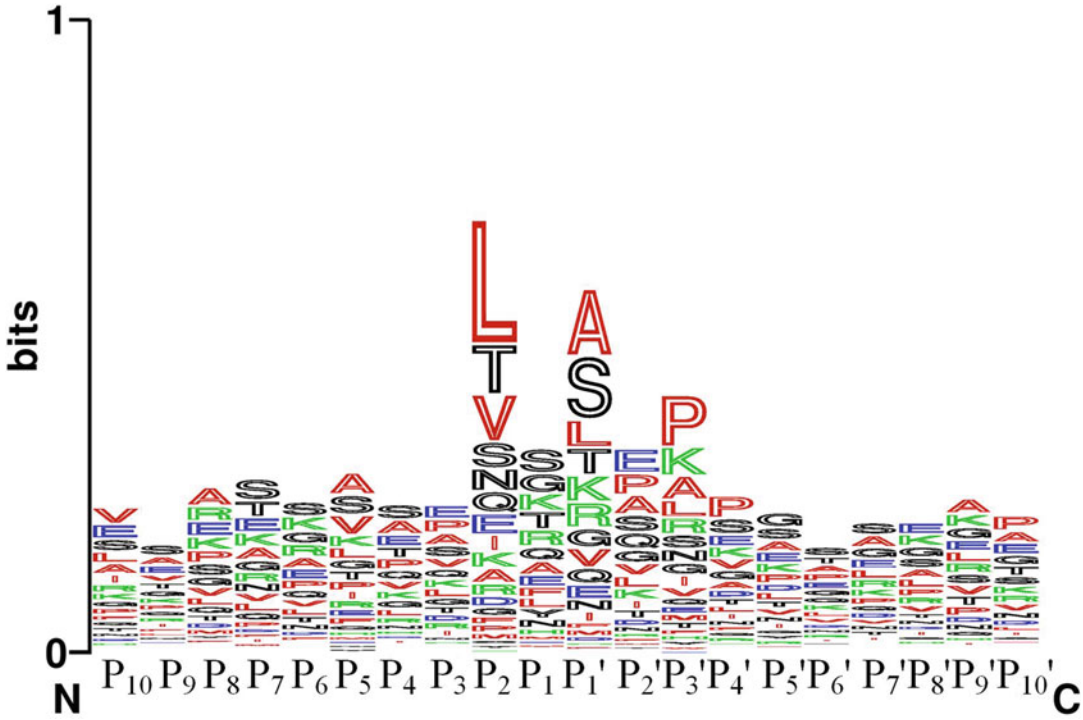

Fig. 1 A statistical view of the amino acid occurrences in a sliding window

Considering a sliding window of 20 residues (P_{10} to P_{10}' as shown in Fig. 1) in a peptide, we determine the probability of the occurrences of amino acids at each position. Such a statistics reflects the residue conservations around the cleavage sites [17]. We did find that such a preference feature is useful for the final predictions [17].

2. Sequence-based solvent accessibility information

From a structural point of view, the solvent accessibility of a residue is closely related to its interaction with other molecules. Hence, the solvent accessibility is also extracted from the sequence. In LabCaS, we apply I-TASSER package [19, 20] to extract solvent accessibility score for each residue that ranges from 0 (buried residue) to 9 (highly exposed residue).

3. Pair-wise alignment similarity score

We also construct a pool of cleavable peptides from the training dataset. Given an amino acid sequence, we split it into many peptides by using the sliding window approach. Each of these segments is aligned against the pre-collected cleavable segments from the pool to generate the pair-wise alignment similarity score, which is considered as a feature.

4. Sequence-based secondary structure information

The cleavable residues are often related with their secondary structures. Therefore, a feature with secondary structures predicted by PSIPRED [21] is consisted in LabCaS.

5. Physical-chemistry property information

In order to incorporate the amino acid physical-chemistry properties [22–24], we grouped the 20 amino acids into five groups: [(V, A, F, I, L, and M), (C, G, P, H, N, Q, S, and T), (W and Y), (D and E), and (R and K)]. The common physical-chemistry properties in each group are fed into the LabCaS program.

2.2 Prediction Models

LabCaS assigns a cleavable label to each residue of a calpain substrate sequence by using CRFs. According to the Hammersley-Clifford theorem of random field [25], the conditional distribution over a labeled sequence $\vec{y}$ given a calpain substrate corresponding sequence $\vec{x}$ can be calculated as:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{t=1}^{T} \exp\left(\sum_i \sum_j \lambda_{ij} f_{ij}(y_{t-1}, y_t) + \sum_j \sum_k \mu_{jk} g_{jk}(y_t, \boldsymbol{x})\right) \tag{1}$$

where $Z(\boldsymbol{x})$ is a normalization factor; $f_{ij}(y_{t-1}, y_t)$ is a transition feature function of the labels at position t and t–1 in the labeled sequence; $g_{jk}(y_t, \boldsymbol{x})$ is a state feature function of the label at position t and the observation sequence; λ_{ij} and μ_{jk} are model parameters corresponding to feature functions $f_{ij}(\cdot)$ and $g_{jk}(\cdot)$; i and j denote the ith and jth kind labels, respectively; and k represents the kth kind sequence pattern.

Five CRF models can thus be constructed from the five features, and the models are further combined using the product rule to obtain the final prediction by LabCaS:

$$\text{LabCaS}(\boldsymbol{x}) = \sqrt[5]{\prod_{j=1}^{5} p_j^*(\boldsymbol{y}|\boldsymbol{x})} \tag{2}$$

3 How to Use LabCaS

3.1 Query Input Page of LabCaS

Figure 2 shows the main page with the sequence entry section, where the sequences for calpain substrate are manually inputted in FASTA format by the users to predict cleavage sites. Upon the submission of each job, a link is provided to the user for tracking the progress of the job and for accessing the final prediction results. The following points are important when using LabCaS. First, LabCaS focuses on the prediction of calpain cleavage sites only. Second, the input sequence length should be at least 50 amino acid

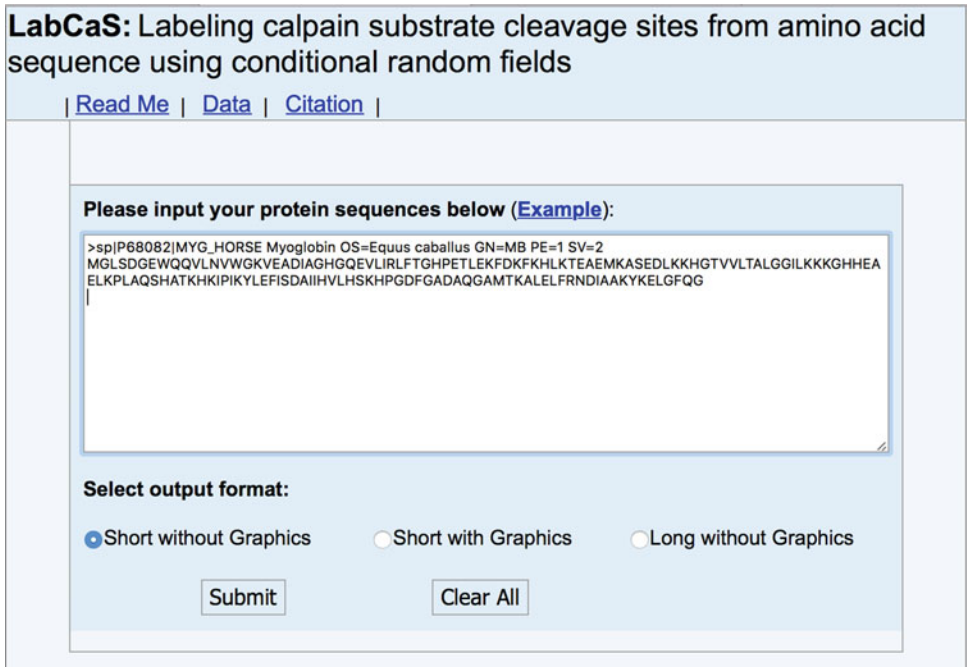

Fig. 2 Screenshot to show the query input page of LabCaS

residues and no more than 6000 residues. Additionally, the number of submitted sequences should not be more than three at a time. Third, there are three output formats of the predicted results, where "short without graphics" format outputs the predicted results at the high threshold; "long without graphics" outputs the results at the high, middle, and low thresholds without the corresponding graphics; and "short with graphics" outputs the results at the high threshold with the corresponding graphics.

3.2 Output Page of LabCaS

Once the submitted job is completed, the results are presented through both webpage and as a downloadable file. An illustration of a typical output is shown in Fig. 3, with a tab for display. The results show predicted calpain substrate cleavage sites, which are presented in a descending order of the confidence score. Additionally, a link is provided to view and download a TXT file comprising the results from LabCaS.

3.3 Case Study for Using LabCaS

Previously, a binary QSAR model was developed to predict the cleavability of calpain proteolysis [18]. Based on this binary QSAR model, the authors obtained 12 predicted cleavage sites of the sequence of horse myoglobin (MYO). Two of these sites, $Lys^{51}\downarrow Thr^{52}$ and $Gly^{151}\downarrow Phe^{152}$, were cleaved by calpains, as demonstrated by experiments utilizing LC-MS/MS. Table 3 shows the results when we submit the sequence to the LabCaS server. The first $Gly^{151}\downarrow Phe^{152}$ cleavage site is ranked as first, and

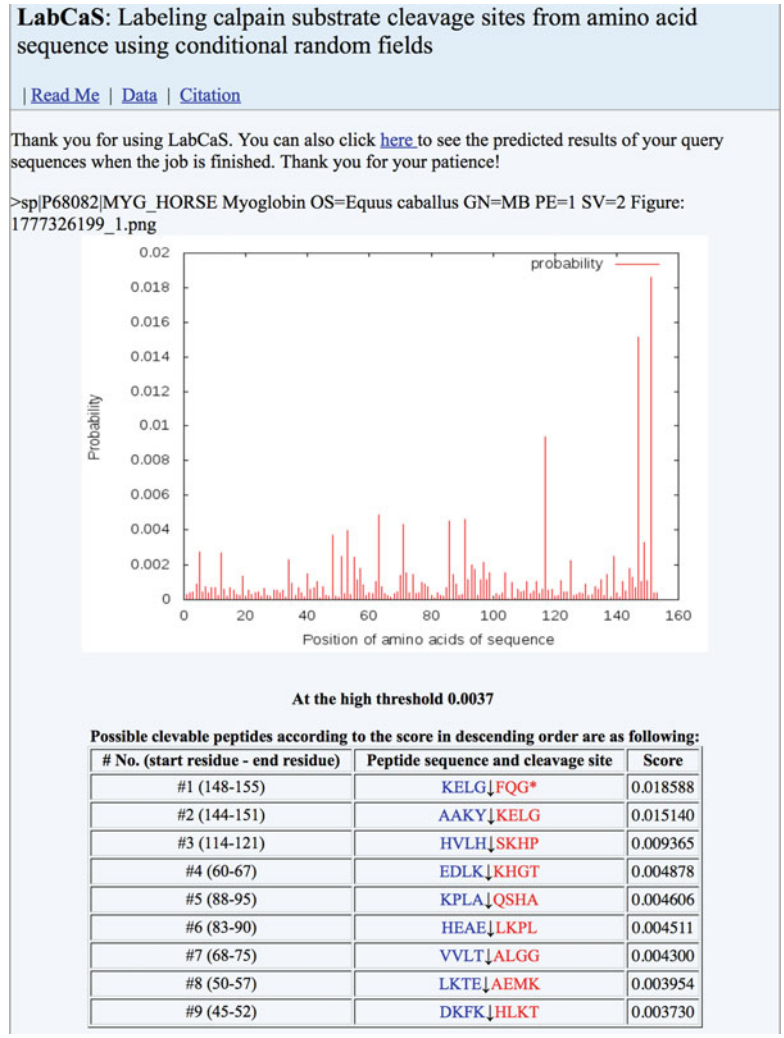

Fig. 3 Screenshot to show an example page of predicted cleavage sites by LabCaS

Table 3
Ranking results for horse myoglobin (accession number: P68082) by LabCaS

Rank	LabCaS
From first to fifth	$Gly^{151}\downarrow Phe^{152}$ (first)
From sixth to tenth	/
From 11th to 15th	$Lys^{51}\downarrow Thr^{52}$ (13th)

the second cleavage site of $Lys^{51}\downarrow Thr^{52}$ is ranked as 13th, as shown in Table 3. These results demonstrate that LabCaS can confidently rank the true cleavage sites and provide a hypothesis for wet-lab verifications.

4 Conclusions

In this chapter, we first briefly reviewed several bioinformatics tools, which include many databases and computational methods related to calpains. Then, the prediction of calpain substrate cleavage sites was formulated as a sequence labeling problem that was achieved by a novel ensemble method called LabCaS, which combines the models based on CRF algorithm. As a web server implementation of our approach, LabCaS is freely available for academic use at http://www.csbio.sjtu.edu.cn/bioinf/LabCaS, which is expected to become a powerful tool for in silico recognition of calpain substrate cleavage sites.

Acknowledgment

We are grateful to Mr. Wallace Chan and Dr. S M Golam Mortuza for proofreading the manuscript. This work was supported in part by the National Natural Science Foundation of China (No. 61462018, 61762026, 61671288, 91530321, 61725302, and 61603161), Guangxi Natural Science Foundation (No. 2017GXNSFAA198278), Guangxi Key Laboratory of Trusted Software (No. kx201403), Guangxi Colleges and Universities Key Laboratory of Intelligent Processing of Computer Images and Graphics (No. GIIP201502), Science and Technology Commission of Shanghai Municipality (No. 16JC1404300, 17JC1403500), and the National Science Foundation (ABI 1564756).

References

1. Campbell RL, Davies PL (2012) Structure-function relationships in calpains. Biochem J 447:335–351
2. Franco SJ, Huttenlocher A (2005) Regulating cell migration: calpains make the cut. J Cell Sci 118:3829–3838
3. Storr SJ, Carragher NO, Frame MC et al (2011) The calpain system and cancer. Nat Rev Cancer 11:364–374
4. Bertipaglia I, Carafoli E (2007) Calpains and human disease. Subcell Biochem 45:29–53
5. Croall DE, Ersfeld K (2007) The calpains: modular designs and functional diversity. Genome Biol 8:218
6. Friedrich P, Bozoky Z (2005) Digestive versus regulatory proteases: on calpain action in vivo. Biol Chem 386:609–612
7. Rawlings ND, Barrett AJ, Finn R (2016) Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 44:D343–D350
8. Duverle D, Takigawa I, Ono Y et al (2010) CaMPDB: a resource for calpain and modulatory proteolysis. Genome Inform 22:202–213
9. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28:304–305
10. Igarashi Y, Eroshkin A, Gramatikova S et al (2007) CutDB: a proteolytic event database. Nucleic Acids Res 35:D546–D549
11. Igarashi Y, Heureux E, Doctor KS et al (2009) PMAP: databases for analyzing proteolytic events and pathways. Nucleic Acids Res 37: D611–D618
12. Tompa P, Buzder-Lantos P, Tantos A et al (2004) On the sequential determinants of calpain cleavage. J Biol Chem 279:20775–20785
13. Boyd SE, Pike RN, Rudy GB et al (2005) PoPS: a computational tool for modeling and

predicting protease specificity. J Bioinforma Comput Biol 3:551–585

14. Verspurten J, Gevaert K, Declercq W et al (2009) SitePredicting the cleavage of proteinase substrates. Trends Biochem Sci 34:319–323
15. Liu Z, Cao J, Gao X et al (2011) GPS-CCD: a novel computational program for the prediction of calpain cleavage sites. PLoS One 6: e19001
16. Duverle DA, Ono Y, Sorimachi H et al (2011) Calpain cleavage prediction using multiple kernel learning. PLoS One 6(5):e19035
17. Fan YX, Zhang Y, Shen HB (2013) LabCaS: labeling calpain substrate cleavage sites from amino acid sequence using conditional random fields. Proteins 81:622–634
18. Shinkai-Ouchi F, Koyama S, Ono Y et al (2016) Predictions of cleavability of calpain proteolysis by quantitative structure-activity relationship analysis using newly determined cleavage sites and catalytic efficiencies of an oligopeptide array. Mol Cell Proteomics 15:1262–1280
19. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738
20. Xu D, Zhang J, Roy A et al (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 79 Suppl 10:147–160
21. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202
22. Granseth E, Von Heijne G, Elofsson A (2005) A study of the membrane-water interface region of membrane proteins. J Mol Biol 346:377–385
23. Mak MW, Wang W, Kung SY (2009) Fusion of conditional random field and signalp for protein cleavage site prediction. In: In acoustics, speech and signal processing. Taipei, pp 716–721
24. Fan YX, Song J, Shen HB et al (2011) PredCSF: an integrated feature-based approach for predicting conotoxin superfamily. Protein Pept Lett 18:261–267
25. Hammersley J, Clifford P (1971) Markov field on finite graphs and lattices. Unpublished manuscript