PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

LabCaS: Labeling calpain substrate cleavage sites from amino acid sequence using conditional random fields

Yong-Xian Fan,[1] Yang Zhang,[2,3]* and Hong-Bin Shen[1,2]*

[1] Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109

[3] Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT

The calpain family of Ca^{2+}-dependent cysteine proteases plays a vital role in many important biological processes which is closely related with a variety of pathological states. Activated calpains selectively cleave relevant substrates at specific cleavage sites, yielding multiple fragments that can have different functions from the intact substrate protein. Until now, our knowledge about the calpain functions and their substrate cleavage mechanisms are limited because the experimental determination and validation on calpain binding are usually laborious and expensive. In this work, we aim to develop a new computational approach (LabCaS) for accurate prediction of the calpain substrate cleavage sites from amino acid sequences. To overcome the imbalance of negative and positive samples in the machine-learning training which have been suffered by most of the former approaches when splitting sequences into short peptides, we designed a conditional random field algorithm that can label the potential cleavage sites directly from the entire sequences. By integrating the multiple amino acid features and those derived from sequences, LabCaS achieves an accurate recognition of the cleave sites for most calpain proteins. In a jackknife test on a set of 129 benchmark proteins, LabCaS generates an AUC score 0.862. The LabCaS program is freely available at: http://www.csbio.sjtu.edu.cn/bioinf/LabCaS.

Proteins 2013; 81:622–634.

Key words: protease substrate specificity; cleavage site prediction; sequence labeling; ensemble learning.

INTRODUCTION

Calpains are a vital conserved family of Ca^{2+}-dependent cysteine proteases which catalyze the limited proteolysis of many specific substrates.[1,2] At present, there are at least 16 known calpain isoform genes in humans, among which 14 genes encode proteins that have cysteine protease domains and the other two encode smaller regulatory proteins that are associated with some catalytic subunits forming heterodimeric proteases.[3] Calpains play a crucial role through cleaving calpain substrates in numerous biological processes, including the regulation of gene expression, signal transduction, cell death and apoptosis, remodeling cytoskeletal attachments during cell fusion or motility, and cell cycle progression.[3,4] Many previous studies have demonstrated that calpain malfunction leads to a variety of diseases,[2,5] including muscular dystrophies,[6] diabetes,[7] and tumorigenesis.[1] Knowing the exact positions of the substrate cleavage sites is very important to revealing the working mechanisms of calpain because the locations of the cleavage sites are closely related to how calpains precisely modulate substrate functions.[8] Although cleavage sites can be determined with various conventional experimental approaches, it is both very laborious and time-consuming

Grant sponsor: National Natural Science Foundation of China; Grant numbers: 61222306, 91130033, and 61175024; Grant sponsor: Shanghai Science and Technology Commission; Grant number: 11JC1404800; Grant sponsor: Foundation for the Author of National Excellent Doctoral Dissertation of PR China; Grant number: 201048; Grant sponsor: Program for New Century Excellent Talents in University; Grant number: NCET-11-0330; Grant sponsor: Shanghai Jiao Tong University Innovation Fund for Postgraduates, the National Science Foundation Career Award; Grant number: DBI 0746198; Grant sponsor: National Institute of General Medical Sciences; Grant numbers: GM083107 and GM084222

*Correspondence to: Y. Zhang, Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109. E-mail: zhng@umich.edu or H. B. Shen, Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China. E-mail: hbshen@sjtu.edu.cn

Received 23 July 2012; Revised 8 November 2012; Accepted 12 November 2012
Published online 23 November 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI:10.1002/prot.24217

to test all of the residues throughout the substrate sequence. To bridge the gap left over by experiments, many computational methods have been proposed to attempt to identify potential calpain substrate cleavage sites using sequences information.[9–11]

Tompa *et al.*[10] computed the amino acid residue propensities around the cleavage sites and established a position-specific scoring matrix (PSSM)-based method by using all 106 calpain cleavage sites in 49 substrates. They found that in μ-calpain and m-calpain, leucine (L), threonine (T), and valine (V) residues often appears in the P_2 position and lysine (K), tyrosine (Y), and arginine (R) in the P_1 position. Boyd *et al.* established a web server called PoPS, which can help researchers to build their own computational models and predict protease specificity based on the specific training datasets submitted by the users. This method mainly scored each subsequence in the substrate by combining the PSSM and weight of the subsite with the sliding window technique.[12] Verspurten *et al.*[13] developed SitePrediction to predict substrate cleavage sites by using the frequency and substitution matrix scoring strategy. Recently, duVerle *et al.* constructed a web service for the prediction of calpain cleavage sites and then further updated their predictor using the multiple kernel learning approach.[14,15] Liu *et al.*[16] constructed a software package named GPS-CCD for the prediction of calpain cleavage sites based on the no interval alignment scoring method.

The above computational methods can be generally grouped into two categories: (1) propensity score method; and (2) machine learning-based two-class classification approaches. A common strategy in the first group is to first calculate the amino acid propensities around the true cleavage sites in the training dataset, and then calculate a total score in a predefined fixed size slide window, which is used to compare with a derived optimal threshold to judge whether the residue located in the center of the window is cleavable or not. The merit of this type of approach is that it is usually highly efficient, while one of the most important shortcomings is that it is very sensitive to the dataset size, where the generated threshold is often heavily biased, especially in small sample size problems. In the second group, by partitioning the whole dataset into positive (represented by cleavable peptides) and negative (represented by noncleavable peptides) subsets, a machine learning based classifier is used for prediction, where typical algorithms include artificial neural networks and support vector machine (SVM). The merit of the approaches in the second category is that they can partially reduce the small sample size effects, while the shortcoming is that the performance can be significantly affected by the extreme imbalance between positive and negative samples (ratio between the sizes of positive and negative subsets can be as small as 1:250). In most cases, to reduce the imbalance effects, one can apply the random downsampling in the negative subsets to generate a balanced training dataset, which however will ultimately greatly reduce the amount of useful information.

In this article, we present a novel method LabCaS to predict the substrate cleavage sites from the flanking sequences of substrates. We develop LabCaS based on the conditional random fields (CRFs) algorithm,[17] which is a sequential supervised machine learning technique. We find that the CRF model is particularly suitable for this study. As a solid machine learning algorithm, CRF is robust to the small sample size problem when learning predicting rules from limited experimentally verified calpain substrate cleavage sites. Another outstanding advantage of CRFs compared with traditional two-class classifiers applied in predicting calpain substrate cleavage sites is that it is a typical sequential learning machine and is insensitive to the ratio between positive and negative training subsets, so all the negative samples can be used to establish the prediction model that can avoid information loss in downsampling process. Considering that the single-view feature only represents part of the protein's information, multiple sequence derived features are integrated to be fed into LabCaS by two different ensemble fusion strategies, that is, feature level fusion and decision level fusion. Our results show that the decision level fusion is a better choice. Experimental results demonstrate the success of LabCaS.

MATERIALS AND METHODS

Materials

The most recent 130 calpain substrate sequences with 368 cleavage sites constructed by Liu *et al.*[16] are used for the training purpose in this study because it is the largest dataset up to now. These experimentally verified calpain substrates with their cleavage sites were obtained by searching the scientific literature from PubMed and then combining with the data collected by Tompa *et al.* and duVerle *et al.*[10,14] The pair-wise sequence identity in the 130 sequences is less than 40%. We removed one of the samples (ID: A2ASS6) in this study because its sequence is too long (35,213 residues) to be dealt with in the current CRF model. We obtained a total of 129 calpain substrate sequences consisting of 367 cleavable sites and 91,743 noncleavable sites.

Methods

Instead of the fragment-based two-class classification approach applied in the traditional way for prediction of the potential cleavage sites by splitting the whole sequence into many short peptides, the developed LabCaS works following a new protocol by labeling the cleavable residues directly from the entire amino acids sequence with CRF algorithm. Motivated by the fact the

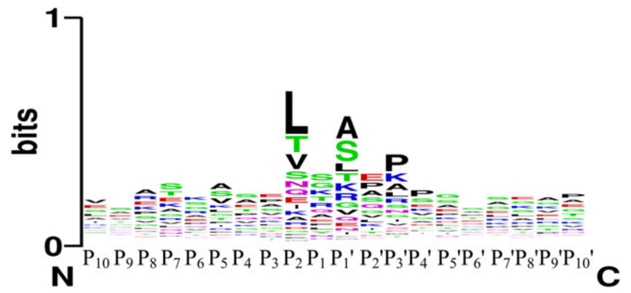

Figure 1

Sequence logo diagram representation of the occurrences of AA residues in the calpain substrate cleavage site from P_{10} to P_{10}'. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

calpain substrate recognition and proteolysis are not controlled by a single determinant but by multiple ones including secondary structure (SS), sequential motif score, and others,[10] we implemented ensemble learning in LabCaS by fusing predicted outputs from five base CRF models trained on different representation features. The improved cleavage site recognition accuracy from the ensemble strategy demonstrates the importance for taking the consideration of multiple determinants.

Prediction features

Amino acid residue preference feature. Sequence logos are a diagram representation of amino acids or nucleic acids in multiple sequence alignment in order to visualize and analyze sequence conservation patterns.[18] Each logo consists of stacks of symbols (amino acids or nucleic acids), one stack for each corresponding position in the sequence. The total height of each stack indicates the degree of the sequence conservation at the corresponding position, while the height of each symbol within the stack represents the relative occurrence of the amino or nucleic acid at that position. As an illustration, Figure 1 displays a sequence logo of 367 positive peptides raging from P_{10} to P_{10}' generated by the WebLogo 3 server,[19] where the cleavage site is between P_1 and P_1'.

Based on the statistics, we can compute the sequence conservation $R_{\mathrm{seq}}(\mathrm{P}_i)$ at a specific position P_i that is defined as the difference between the maximum possible entropy $E_{\max}(\mathrm{P}_i)$ and the entropy of the observed symbol distribution $E_{\mathrm{obs}}(\mathrm{P}_i)$:

$$R_{\mathrm{seq}}(\mathrm{P}_i) = E_{\max} - E_{\mathrm{obs}} = \log_2 N - \left(-\sum_{n=1}^{N} (p_n(\mathrm{P}_i)) \log_2 (p_n(\mathrm{P}_i)) \right), \quad (1)$$

where $p_n(\mathrm{P}_i)$ is the observed frequency of symbol n at the position P_i, and N is the number of distinct symbols, equal to 20 for a protein sequence in this study. Although Figure 1 does not contain strong evidence of sequence conservation throughout the 20-mer input, we can still find that pentapeptide P_2–P_3' shows some residue preferences, especially at the position P_2 with approximately 0.7204 bits according to Eq. (1) and Table I. So, we further analyze the amino acids propensities of amino acids at positions P_2–P_3'. Our results reveal that the most significant preference is that 28.5% of cleavable sites have a P_2 leucine (L). Aside from the P_2 site specificity, we note a modest preference for alanine (A) at P_1' and proline (P) at P_3'. Alanine (A) is present in 18.5% of all P_1' positions and proline (P) occurs in 20.1% of all P_3' positions.

Solvent accessibility sequences information. The solvent accessibility (AC) of a residue is related to its cleavability, and hence can be used to enhance the prediction performance of calpain substrate cleavage sites.[15] Several other methods, such as Cascleave[20] and SitePrediction,[13] have exploited the predicted AC to predict the substrate cleavage sites. Two-state (exposed or buried) AC can be predicted by using the SSpro grogram[21] as an additional feature description. The substrate cleavage sites are generally considered to be relatively exposed, but there are indeed examples where the proteolytic cleavages happen at solvent inaccessible regions.[22] Therefore, a naïve two-state AC cutoff may reject some true positive cleavage site predictions. In our study, we use the real-value AC predictions that was generated by the I-TASSER protein structure prediction package,[23,24] where the AC value was trained by the neural network machine with the combination of sequence profiles and protein three-dimensional (3D) structural models. The AC value ranges from 0 (buried residue) to 9 (highly exposed residue)

Table I

Propensities Analysis of Amino Acids Around Calpain Cleavage Site at Position P_2–P_3'

Amino acids	P_2	P_1	P_1'	P_2'	P_3'
A	0.035	0.068	0.185	0.095	0.092
C	0.011	0.003	0.003	0.005	0
D	0.024	0.019	0.016	0.035	0.014
E	0.049	0.057	0.041	0.12	0.03
F	0.019	0.052	0.019	0.022	0.022
G	0.022	0.098	0.054	0.065	0.049
H	0.008	0.035	0.011	0.014	0.022
I	0.038	0.008	0.03	0.046	0.046
K	0.038	0.087	0.073	0.057	0.111
L	0.285	0.052	0.076	0.06	0.068
M	0.016	0.024	0.019	0.019	0.03
N	0.052	0.046	0.033	0.033	0.052
P	0.019	0.011	0.011	0.106	0.201
Q	0.052	0.071	0.046	0.084	0.033
R	0.03	0.084	0.071	0.024	0.065
S	0.063	0.122	0.174	0.087	0.057
T	0.117	0.087	0.076	0.046	0.03
V	0.109	0.019	0.054	0.063	0.041
W	0.003	0.005	0.003	0	0.016
Y	0.011	0.052	0.005	0.019	0.022

which quantifies the degree of the surface area of a given residue that is accessible to the solvent. In the large-scale benchmark test,[25] the I-TASSER AC prediction was shown to have a correlation coefficient 0.83 with the real AC of experimental structures assigned by the DSSP.[26]

Pair-wise alignment similarity score (BL). Under the assumption that similar short peptides probably possess similar biological functions, we try to infer the cleavability of a query peptide based on the pair-wise similarity between the query and those in the training dataset. The similarity between the query and the cleavable sites in the dataset can be evaluated by the pair-wise sequence alignment with a substitution matrix, such as BLOSUM62. The similarity, $S(A_1, A_2)$, between two short peptides A_1 and A_2 of $(m + n)$ residues can be defined as:

$$S(A_1, A_2) = \sum_{-m \le i \le n} S(A_1^i, A_2^i), \qquad (2)$$

where A_1^i and A_2^i are the amino acids at the ith residue in peptides A_1 and A_2, respectively. Because some elements of the BLOSUM62 matrix are negative, $S(A_1, A_2)$ could be negative. We set $S(A_1, A_2) = 0$ if $S(A_1, A_2) < 0$, as has been done previously.[16] The final score of a query peptide is the average of all the similarity scores obtained by a pair-wise comparison between the query and each of the training samples. Here, m and n are set to 10 and 4 because a previous study[16] has proved that this was an optimal choice.

SS sequence information. As calpains hydrolyze, its substrate proteins in a limited manner, resulting in fragments keeping intact domains, it has been suggested that calpains prefer to cleave the substrates in flexible regions between structured domains.[6] Hence, the SS context is an important factor for determining whether the presence of a particular substrate motif can be accessed and cleaved by calpain. The dataset of calpain substrates used in this study allows us to perform a comprehensive analysis of the structural determinants that characterize the calpain substrate specificity. We predict the SS type (helix, strand, or coil) for each residue using PSIPRED.[27]

Physical-chemistry property sequence information. Grouping amino acids according to their physical-chemistry (PC) properties is helpful for reducing the noise caused by mutations, and thus improving the accuracy of protein structure and function predictions.[28–30] Considering this point, we divided 20 amino acids into five groups and each group stands for a PC property of the amino acids, as shown in Table II. Amino acid residues V, A, F, I, L, and M with strong hydrophobicity form the hydrophobic group, the residues C, G, P, H, N, Q, S, and T present obvious polarity so they form the polar group, W and Y are form the aromatic group, D and E are form the acidic group, and R and K are in the basic group.

Table II
Grouping of Amino Acids According to their Physical-Chemistry Properties

Property	Group
Hydrophobic	P1 = [V,A,F,I,L,M]
Polar	P2 = [C,G,P,H,N,Q,S,T]
Aromatic	P3 = [W,Y]
Acidic	P4 = [D,E]
Basic	P5 = [R,K]

Prediction models

Labeling calpain substrate cleavage sites using CRFs algorithm. The task of calpain substrate cleavage sites prediction is to assign a label from a finite set of labels to each residues of a calpain substrate sequence. CRFs, a sequential labeling algorithm, were first used for labeling natural language sequence data by Lafferty.[17] Given a random vector over sequences $\boldsymbol{x} = [x_1, x_2, \ldots, x_T]$, CRFs try to obtain the most probable random vector over the corresponding labeled sequences $\boldsymbol{y} = [y_1, y_2, \ldots, y_T]$, that is, $\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x})$. CRFs are undirected graphical models, and the conditional probability $P(\boldsymbol{y}|\boldsymbol{x})$ can be computed directly. Recently, CRFs have attracted much attention and been successfully applied in bioinformatics literature for dealing with biological sequences.[31]

We formulate the prediction of calpain substrate cleavage sites based on the CRF labeling approach, a calpain substrate's corresponding sequence can be denoted as $\boldsymbol{x} = [x_1, x_2, \ldots, x_T]$ $(x_i \in \Theta)$ where Θ varies in different representation modes. For example, Θ is the set of twenty amino acid letters when the corresponding sequence is amino acid sequence, 10 single digits if the corresponding sequence is the predicted solvent accessibilities or the BLOSOM62-based pair-wise alignment similarity scores, the set of H, C, and E when the corresponding sequence is the predicted SSs, the set of 1, 2, 3, 4, and 5 standing for the five different amino acid groups if the corresponding sequence is the PC properties. In the case of identifying the potential cleavage sites, the corresponding label sequence is denoted as $\boldsymbol{y} = [y_1, y_2, \ldots, y_T]$ $(y_i \in L)$, where L is the set of C and N which stand for the cleavage sites and noncleavage sites, respectively. According to the fundamental Hammersley–Clifford theorem of random field,[32] the conditional distribution over a labeled sequence $\boldsymbol{y}$ given a calpain substrate corresponding sequence $\boldsymbol{x}$ is as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{t=1}^{T} \exp\left(\sum_i \sum_j \lambda_{ij} f_{ij}(y_{t-1}, y_t) + \sum_j \sum_k \mu_{jk} g_{jk}(y_t, \boldsymbol{x}) \right), \qquad (3)$$

where

$$Z(\boldsymbol{x}) = \sum_{y'} \prod_{t=1}^{T} \exp\left(\sum_i \sum_j \lambda_{ij} f_{ij}(y'_{t-1}, y'_t) + \sum_j \sum_k \mu_{jk} g_{jk}(y'_t, \boldsymbol{x})\right), \quad (4)$$

where $Z(\boldsymbol{x})$ is a normalization factor; $f_{ij}(y_{t\text{-}1},y_t)$ 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 $g_{jk}(\cdot)$ which are typically Boolean functions; i and j denote the ith and jth kind labels, respectively; k represents the kth kind sequence pattern.

One of the most important things for applying CRFs in identifying the substrate cleavage sites is to identify the model parameters of Eq. (3), which can be typically learned on the training dataset using a maximum likelihood approach. That is, maximizing the conditional log likelihood of the training examples over the parameter space.[33] Given N (N = 129) substrate sequences with known labels of each residue of $D = [\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}]_{i=1}^{N}$, where $\boldsymbol{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}]$ is an observation sequence from a substrate, and $\boldsymbol{y}^{(i)} = [y_1^{(i)}, y_2^{(i)}, \ldots, y_T^{(i)}]$ is a desired label sequence. The conditional log likelihood can be defined as follows:

$$\ell(\theta) = \sum_{i=1}^{N} \log P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}), \quad (5)$$

where θ is the parameter vector. By substituting Eq. (3) into Eq. (5) and maximizing $\ell(\theta)$, we can finally achieve the appropriate parameter vector $\hat{\theta}$ and construct the CRFs model based on D. For detailed process of solving $\hat{\theta}$, please refer to ref [17]. With the constructed CRFs, the most probable label sequence probability $p^{\star}$ for an input sequence $\boldsymbol{x}$ can then be inferred according to the dynamic programming algorithms or some approximate inference algorithms as follows:

$$p^{\star} = \arg\max_{y} P_{\hat{\theta}}(\boldsymbol{y}|\boldsymbol{x}) \quad (6)$$

The pocket CRF (http://sourceforge.net/projects/pocket-crf-1/files/pocket_crf/), an open source implementation of CRFs for labeling sequential data, is adopted to perform our experiments.

Ensemble prediction. As we can represent a substrate with five different encoding methods as discussed above, that is, amino acid (AA), AC, BL, SS, and PC, hence, five CRF models can be obtained on each of the sequential representation accordingly denoted as CRF-AA, CRF-AC, CRF-BL, CRF-SS, and CRF-PC respectively. In this study, the five outputs are integrated together using the product rule as the final prediction:

$$E(\boldsymbol{x}) = \sqrt[5]{\prod_{j=1}^{5} p_j^{*}(\boldsymbol{y}|\boldsymbol{x})}. \quad (7)$$

Cross-validation and performance assessment

In this study, we tested our proposed method using leave-one-protein-out jackknife cross-validation which takes one protein sequence out for testing while keeping the remaining protein sequences for training. This procedure will be terminated when all the proteins have been tested individually. The predictive ability of LabCaS is assessed using several measures, namely, sensitivity (SN), specificity (SP), the Mathews correlation coefficient, and the overall accuracy (ACC). It should be pointed out that the aforementioned four measurements rely on the selected prediction thresholds. Hence, another threshold-independent criteria, the Area Under the ROC curve (AUC) is also applied for evaluating the performances. When all the residues of training sequences in the dataset have been labeled by the CRF algorithm based on the validation tests, we will get a continuous numeric value to represent the confidence of a residue belonging to its predicted class (cleavable or not). Then, gradually adjusting the classification threshold will produce a series of confusion matrices. From each confusion matrix, a ROC point, the coordinate of which is (TP/TP+FN, FP/FP+TN), can then be computed. A series of ROC points constitute the ROC curve, where the AUC value can be finally calculated.

RESULTS AND DISCUSSIONS

Statistical results affected by dataset scale

According to Figure 1, the AA prevalence at P_2 was observed to be leucine (L), threonine (T), and valine (V), consistent with Tompa's study.[10] Slight differences were observed at other positions in the current dataset compared to the previous study. For example, at the position P_1, serine (S), glycine (G), and lysine (K) are found to be the top three most popular AAs in the current study, while in the ref.[10] lysine (K), threonine (T), and arginine (R) were reported as the top three. According to Figure 1 alanine (A), serine (S), and leucine (L) are the most common AAs at position P_1'. This is contrary to Tompa *et al.* who found that threonine (T), lysine (K), and arginine (R) are the most common AAs at position P_1'. These differences could be caused by the different sizes of datasets used for statistics. The current study is based on 129 substrate sequences consisting of

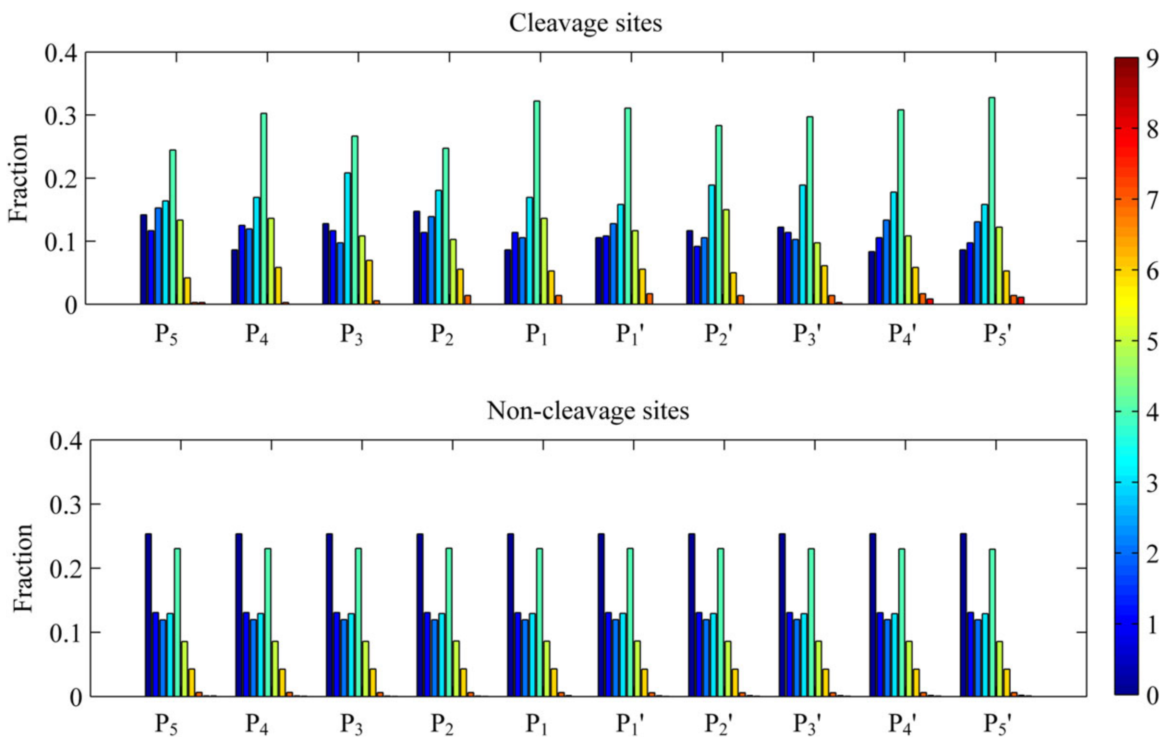

Figure 2

Distributions of predicted AC from P_5 to P_5' positions in the vicinity of cleavage sites and noncleavage sites; the values range from 0 (buried residue) to 9 (highly exposed residue) [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.].

367 cleavage sites, which is much larger than the dataset used in previous statistics that contains 49 substrate sequences of 106 cleavage sites.

Analysis of determinants that characterize calpain substrate specificity

The distributions of the AC values generated by I-TASSER are displayed in Figure 2.[23,24] The AC values range from 0 (buried residue) to 9 (highly exposed residue) which quantifies the degree of the surface area of a given residue that is accessible to the solvent. Near the noncleavage sites, from position P_5–P_5', the distribution is uniform at each position. This is especially true for the residues of degree 0 (buried residues), which account for ∼25% and the residues of degree 4 (moderately exposed residues) which account for ∼23% at each position. However, near the cleavage sites from position P_5–P_5', there are relatively few residues of 0 degree with a maximum of ∼14% at the position P_2. There are a significantly greater number of residues with degree 4 exposures at positions P_1–P_5' near the cleavable sites than near the noncleavable sites, with residues with degree 4 exposure accounting for 31% of all residues near cleavable sites. The distribution also demonstrates that most residues are moderately exposed and only very few residues are highly exposed (9 degree) near the cleavage sites. In the vicinity of the noncleavage sites, the buried residues account for a greater proportion of residues.

The frequencies of SS types occurring at each position from P_{10} to P_{10}' reveal that calpain most frequently cleaves substrates that contain coils or loops (Fig. 3), which is in agreement with the earlier findings.[20] Despite the fact that coil regions account for 45% of the residues

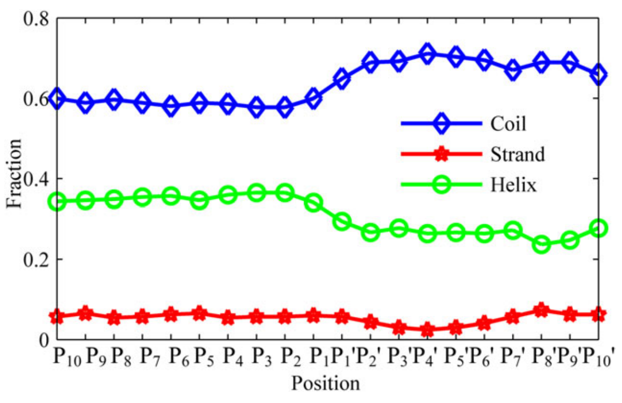

Figure 3

Distributions of predicted SSs from P_{10} to P_{10}' positions in the 129 substrate sequences. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

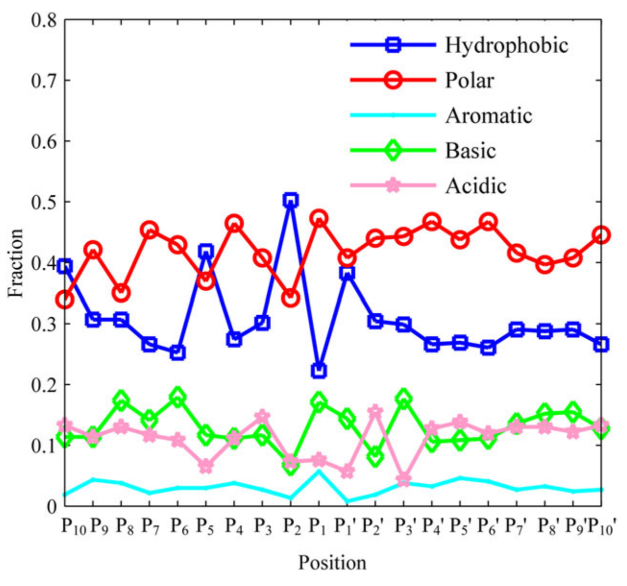

Figure 4

Distributions of different PC categories of AA residues at P_{10}–P_{10}' positions. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

in our sample, the SS composition at position P_{10}–P_1 is approximately 60% coil as shown in Figure 4, which steadily increases up to 70% at position P_1'–P_{10}'. The amount of helix is more than 30% at position P_{10}–P_1 and slightly lower than 30% at position P_1'–P_{10}'. The amount of strand structure is relatively small, accounting for only 6%. These statistics demonstrate that the cleaving is more likely to happen in the flexible regions than the rigid domains, which supports former hypothesis.[6]

The distributions of different PC categories of AA residues are displayed in Figure 4. The region near the calpain cleavage site contains mainly polar residues, but also a significant proportion of hydrophobic residues. It is interesting to find that hydrophobic residues increase to 50% at position P_2 and significantly decrease to 20% at position P_1. In contrast, polar residues, aromatic residues, and basic residues decrease at position P_2, and then increase at position P_1.

Performance affected by imbalanced training samples in two-class classifiers

As we have discussed above, the task of predicting substrate cleavage site was widely formulated as a two-class classification problem, that is, classifying cleavable peptides from nonleavable ones. For example, kernel function based classifiers are applied in previous studies.[15,20] To demonstrate the effects caused by the extremely imbalanced negative and positive training samples of calpain substrate cleavage sites, we apply the widely used SVM as a benchmark algorithm for a test as kernel learning has been adopted in Ref. 15. For the calpain substrate prediction purpose, we then use a similar dataset for a demonstration, which consists of 91 non-redundant substrate sequences with 244 cleavage sites (a slight difference occurred because of the update of CaMPDB database used for constructing dataset). The inputs to the SVM are encoded by the classical binary encoding scheme in the window of P_2–P_3' positions because this local environment shows the most significant discriminative feature. The cost parameter C and kernel parameter γ of SVM were optimized based on the grid-searches on $C \in [2^6, 2^5, \ldots, 2^{-1}, 2^{-2}]$ and $\gamma \in [2^3, 2^2, \ldots, 2^{-4}, 2^{-5}]$. Figure 5 illustrates the average classification performances by the jackknife test at the sequence level under different configurations in the training dataset between negative and positive samples: 2^0:1, 2^1:1, 2^2:1, 2^3:1, 2^4:1, 2^5:1, 2^6:1, and 2^7:1. As revealed by Figure 5, the performance is indeed heavily affected by the ratio between negative and positive samples. Take the case of 2^7:1 as an example that is the closest to the reality, the AUC is 0.634, which is significantly lower than the AUC of other cases.

Performance of single feature input

We assessed the abilities of five individual features to predict the calpain substrate cleavage sites using the CRF model. The corresponding ROC curves are displayed in Figure 6. Among the five features, the AA preference feature (AA) achieves the highest AUC value of 0.821, followed by the peptide pairwise alignment similarity score (BL) which yields an AUC value of 0.811. AA categories based on PC properties yields an AUC value of 0.735 which is higher than the AUC value of 0.700 obtained by the predicted AC. The AUC values for features based on predicted SS is 0.694.

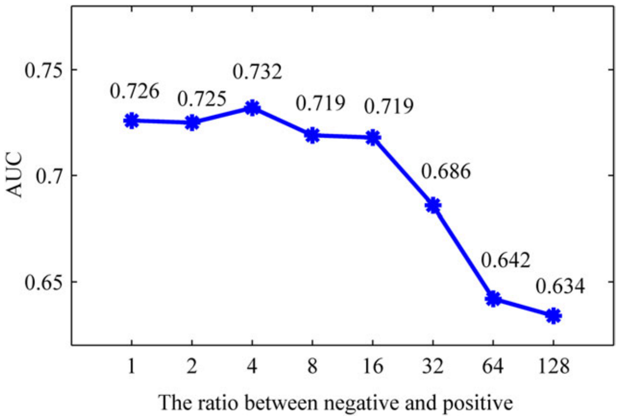

Figure 5

The AUC values of different proportions between negative and positive samples. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

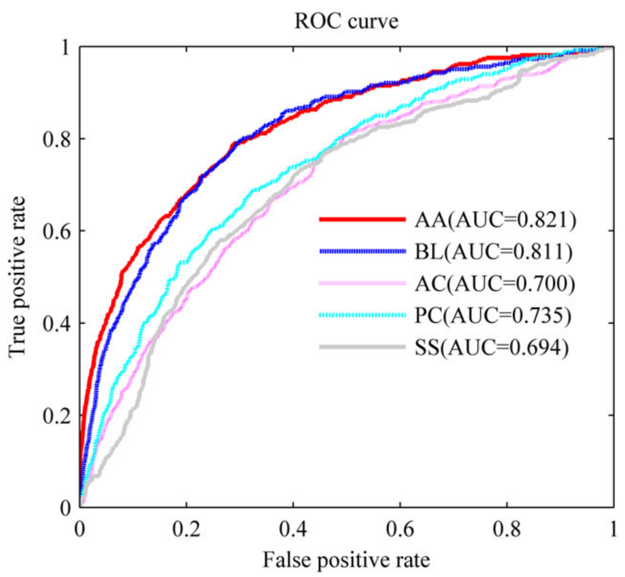

Figure 6

The prediction performances achieved by different single-view feature inputs. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Ensemble prediction achieved by feature level fusion

Obviously, as shown in Figure 6, significant differences exist among different single features, indicating single-view feature can only reflect part of the information of a target. It has been proven in many reports [20,30] that combining different features can improve predictive performance. According to the forward search idea, different features can be added step by step starting from the best single-view feature for inputting to the CRF model, that is, starting from AA feature in this study (Fig. 6). The results are displayed in Figure 7. First, the combination of AA and SS is chosen for the next round because it yields the best AUC value of 0.835 after testing the four different groups BL, SS, AC, and PC with AA. Then, starting from the group of AA and SS, we obtain a triple-features group consisting of AA, SS, and PC with an AUC value of 0.838, and so forth. Finally, an AUC of 0.840 is observed when combining all five types of features. These results have demonstrated that the performance can be improved by the combination of multiview features because different features can be complementary to each other. It is worth pointing out here that it is not always the case that the best performance will be achieved based on the combination of all available features, which is the case of this study. At the same time, the paired t-tests were carried out among the five sub-methods for different features using 129 AUC values generated in the jackknife cross validation tests (one AUC value for one sequence) as tabulated in Table III. If the resulting p-value is below the desired level, for example, 0.05, the differences between the tested features can be considered significant. Taking feature AA as an example as shown in Table III, the differences of AA and SS, AA and PC, AA and AC, are statistically significant, however, the p-value between sub-methods of AA and BL is 0.9507, which is larger than 0.05. It is also interesting to find from Figure 7 that the performances yielded from the combinations of statistically different features are generally better than combinations of features of no significance difference. For example, the combination of AA and SS generates an AUC score 0.835, while fusion of AA and BL gives an AUC value 0.825. This phenomenon is supported by the general acknowledgement in the machine learning field that diversity is closely related with the ensemble model and higher diversity will yield better results.[34]

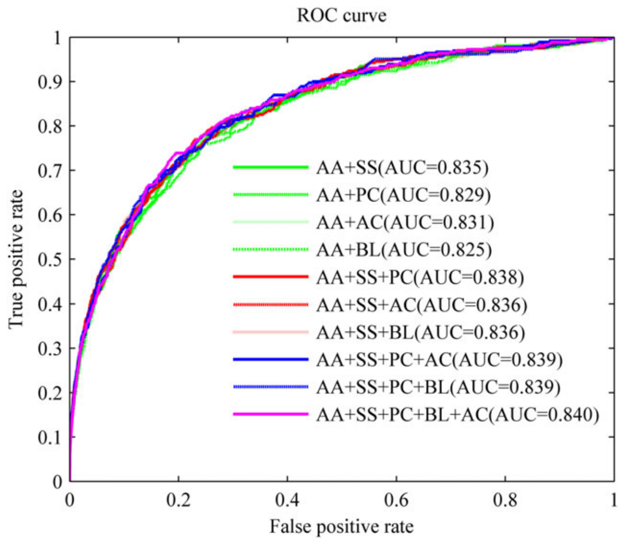

Figure 7

ROC curves in different feature level fusion cases. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table III

The Comparison Results Among the Five Sub-Methods Using the Paired t-Test

	AA	BL	AC	PC	SS
AA	–	0.9507	4.8129e-09	1.2834e-05	2.5705e-08
BL	0.9507	–	7.7942e-10	7.0403e-06	1.7332e-08
AC	4.8129e-09	7.7942e-10	–	0.0056	0.3664
PC	1.2834e-05	7.0403e-06	0.0056	–	0.0225
SS	2.5705e-08	1.7332e-08	0.3664	0.0225	–

Ensemble prediction achieved by decision level fusion

An improved AUC of 0.840 has been achieved by performing ensemble prediction through combination of

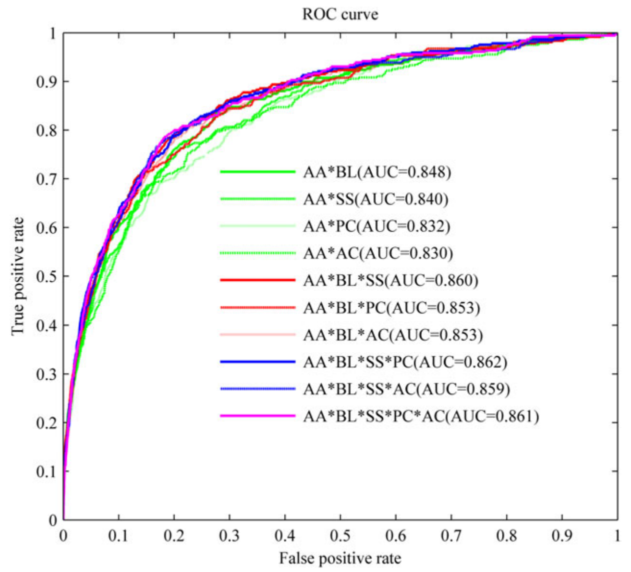

Figure 8

ROC curves in different decision level fusion cases. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

multiple features. In this section, we will try to construct another type of decision level fusion based consensus predictor. Instead of the feature level fusion, five independent base predictors will be trained on the five different single sequence features. The five independent outputs can be used as inputs of a consensus predictor. The product rule is used for the combination of five base predictors [Eq. (7)].[35] The result generated by each single feature is then combined step by step according to the forward search algorithm as illustrated in Figure 8. As shown in Figure 8, a best AUC value of 0.862 is obtained by the combination of four base predictors from AA, BL, SS, and PC features. By comparing the results shown in Figures 7 and 8, we find that performance from decision level fusion is better than the feature level fusion, which has been improved by 2%. The *p*-value of the paired *t*-test to compare the 129 jackknife cross validation AUC values from decision level fusion and feature level fusion approaches is 4.802e-004, which demonstrates the decision level fusion strategy is statistically better than the feature level fusion method. The reason could be simply that a combination of the different views of features will increase the information redundancy although it will represent more knowledge. Hence, based on the analysis above, we finally implemented LabCaS based on the decision fusion protocol.

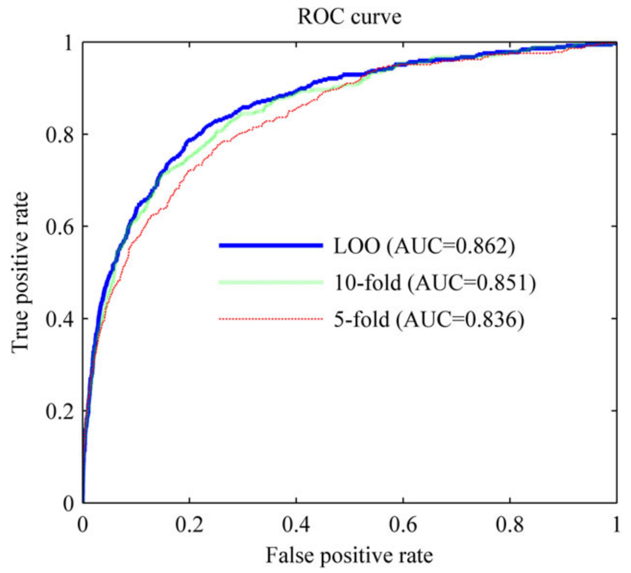

Figure 9

ROC curves in LOO validation, 5-fold and 10-fold cross-validations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table IV

Comparison of LabCaS with GPS-CCD

SP	Method	SN (%)	ACC (%)	MCC	Threshold
95%	LabCaS	49.05	94.82	0.1253	0.0037
	GPS-CCD	45.92	94.87	0.0998	High
90%	LabCaS	63.22	89.89	0.1107	0.0026
	GPS-CCD	60.87	89.98	0.0908	Medium
85%	LabCaS	71.93	84.95	0.0998	0.0020
	GPS-CCD	66.58	84.99	0.0773	Low

In addition to the aforementioned leave-one-out (LOO) jackknife validation, 5-fold and 10-fold cross-validations were also carried out to evaluate the prediction robustness of the constructed LabCaS. According to the results of decision level fusion displayed in Figure 8, the ROC curves for the fusion of four base predictors of AA, BL, SS, and PC were drawn in Figure 9. The AUC values were 0.836 (5-fold) and 0.851 (10-fold), respectively. As demonstrated by Figure 9, the performances of LOO, 10-fold, and 5-fold are decreasing, indicating that the training dataset size affects the prediction models. That is to say, in the 5-fold test, 26 substrate sequences are singled out for tests, where only 103 sequences are left for training the model; while in the 10-fold test, there are 116 training samples, and in the LOO jackknife test, there are a total of 128 training samples. Considering that there are only very limited experimentally verified calpain substrates with known cleavable sites, it is critical to develop much more robust computational approaches in this regard.

Comparison with existing methods

GPS-CCD was developed by Liu *et al.*[16] as a web-tool for calpain substrate cleavage sites prediction. GPS-CCD achieved the prediction of a putative calpain substrate cleavage peptide via similarity scoring. Table IV compares

Table V

Prediction Results for Using SVM (RBF), GPS-CCD, and LabCaS on Rat Microtubule-Associated Protein tau

Rank	SVM (RBF)	GPS-CCD	LabCaS
From 1st to 5th	Nil	Nil	Ser^{120}↓Lys^{121} (4th)
From 6th to 10th	Nil	Nil	Nil
From 11th to 15th	No outputs	Nil	Gly^{147}↓Ala^{148} (15th)
From 16th to 20th	No outputs	Ser^{120}↓Lys^{121} (16th)	Nil

LabCaS with GPS-CCD in the three cases of fixed SP on the same dataset consisting of 129 substrate sequences. LabCaS outperforms GPS-CDD in all tested situations. When the SP is set to the most stringent 95%, the sensitivity of LabCaS is 3% higher than the sensitivity of GPS-CCD; and when we set the SP to 85%, the sensitivity of LabCaS is approaximately 5% higher than the sensitivity of GPS-CCD.

SVM(RBF) is another web-tool for calpain substrate cleavage site prediction which was built by duVerle *et al.*[14,15] Following the steps described in the original paper, we downloaded the 104 calpain substrates from the latest CaMPDB database[14,15] and reduced their homology at the threshold 95% by using CD-HIT.[36] At last, 96 non-redundant calpain sucstrate sequences were obtained. We designed a 10 × 10 cross validation test based on this non-redundant dataset the same as SVM based predictor by using our proposed LabCaS method. The final average AUC value of LabCaS is 0.8440 on the 96 non-redundant sequences, which is higher than 0.7686 reported in SVM(RBF).[15] To further compare the LabCaS with the SVM-based approach, we searched the 129 calpain substrate sequences in the benchmark dataset of this paper against the latest CaMPDB database and found 77 sequences are not included in the 104 records of CaMPDB.[14,15] These 77 calpain substrate sequences are submitted to the web-server of SVM(RBF) for calculations. In accordance to the scores outputted by SVM(RBF), the AUC value is 0.6139 (the probabilities of sites without outputs from SVM(RBF) are set to zeros). The prediction results of these 77 calpain substrates by our LabCaS in the jackknife test are also extracted to calculate the AUC value and 0.8703 is obtained, which is significantly better than the SVM(RBF) approach. All these results demonstrate that the LabCaS is better than the state-of-the-art calpain cleavage site predictors and will play an important complementary role with existing methods.

Rat microtubule-associated protein tau: A case study and comparison

Axonal specific microtubule-associated protein tau plays important roles in complex diseases such as Alzheimer's disease and chronic traumatic encephalopathy. In the living cell, both calpain and caspase-3 are capable of tau processing. Although it has been known that tau protein is a substrate for calpain *in vitro* for a long time,[37] the specific calpain cleavage sites have never been reported until a recent study by Liu *et al.*,[38] which has identified three novel calpain cleavage sites in rat tau, that is, Ser^{120}↓Lys^{121}, Gly^{147}↓Ala^{148}, and Arg^{370}↓Glu^{371}. We then submit the primary sequence to SVM(RBF),[14] GPS-CDD[16] for predictions and compare their outputs with LabCaS's, and the results are tabulated in Table V. LabCaS successfully predicted two cleavable sites for rat tau of Ser^{120}↓Lys^{121} and Gly^{147}↓Ala^{148} with the highest confidence threshold of 0.0037, but missed Arg^{370}↓Glu^{371}. Table V also shows that the top 10 prediction outputs from SVM (RBF)[14] fail to identify any of the three cleavage sites; the top 20 prediction outputs of GPS-CCD[16] target one true cleavage site of Ser^{120}↓Lys^{121}, which is ranked 16th. For LabCaS, Ser^{120}↓Lys^{121} cleavage site is ranked 4th and Gly^{147}↓Ala^{148} ranked 15th. These results demonstrate that LabCaS is more powerful than existing approaches in this example.

In Figure 10, we show the 3D structural model of the rat microtubule-associated tau protein generated by the I-TASSER simulations, one of the best performing protein structure prediction algorithms in the recent community-wide CASP experiments.[23,39] The model has a confidence score (C-score) −1.03 which corresponds to a modest TM-score 0.58 ± 0.14, where a TM-score >0.5 indicates a correct fold of the protein molecule.[40] Nevertheless, all the true positive cleavage sites (red color residues) are located on the surface of the 3D structure in

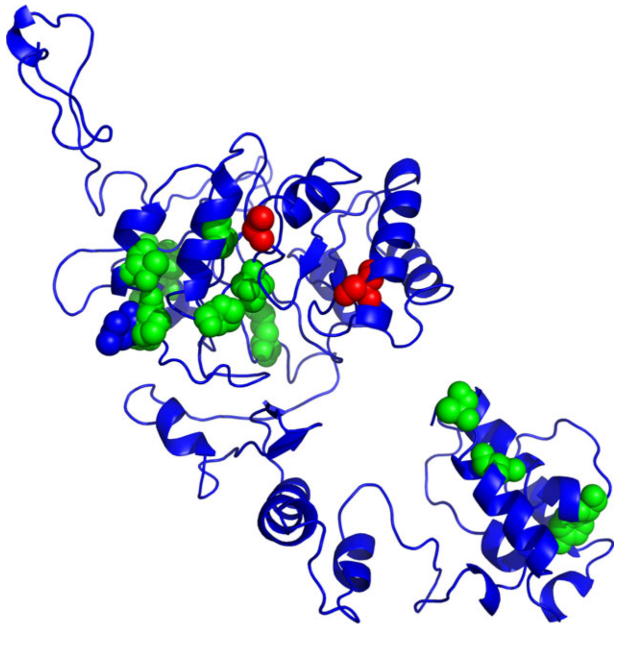

Figure 10

The 3D view of the rat microtubule-associated protein tau with top 15 predicted cleavage sites by LabCaS. The correct predicted cleavage sites are colored red and the incorrect predicted cleavage sites are colored green.

Table VI
The Predicted Cleavage Sites of Potential Calpain Substrates in Lysosomal Membranes Using LabCaS

Number	Calpain Substrates in lysosome membrane[a]	Protein length (aa)	Predicted cleavage site using LabCaS at the highest threshold[b]
1	Long-chain-fatty-acid--CoA ligase 1 (NP_032007.2)	699	694, 111, **607**, 200, **403**, 256, **44**, 619, **448**, 676, 471, 432, 612, 281, 578, 262
2	Serotransferrin precursor (NP_598738.1)	697	304, 348, 460, 618, 100, 690, **525**, 670, 467, 628, 132, 166, 167, 475, 314, 693, 47, 140, 689, 678, 343, **610**, 688, 622, 623, 692, 271, 142, 80, 694, 90
3	Beta-2-glycoprotein 1 precursor (NP_038503.4)	345	**154**, **338**, **49**, 275, 38, 340, 336, 342
4	V-type proton ATPase subunit B, brain isoform (NP_031535.2)	511	311, 12, 11, 355, 393, 504, 18, 154, **4**, 403, 502, **501**, 464, 34, 192, 314, 508, 100, 507
5	D-3-phosphoglycerate dehydrogenase (NP_058662.2)	533	406, 322, 96, 494, 364, **389**, 285, 530, 125, 74, **60**, 379, 469, 165, 454, 344, 95, 126, 526, 250
6	Adipocyte plasma membrane-associated protein (NP_082253.1)	415	186, 22, 155, 135, 215, 411, **323**, **203**, 14, 410, 184
7	N(G),N(G)-dimethylarginine dimethylaminohydrolase 1 (NP_081269.1)	285	**19**, 201, **236**, 113, 42, **50**, **18**, 5, 11, 99, 213
8	Serine/threonine-protein phosphatase PP1-gamma catalytic subunit (NP_038664.2)	323	319, **318**, **315**, **314**, 320, **267**, 294, **20**, 309, 268, 316, 308, **305**, 54, 122, 310, 48
9	60S acidic ribosomal protein P0 (NP_031501.1)	317	312, 252, 142, **70**, 129, 114, 293, **286**, **120**, 228, **291**, 308, 112, 103, 159, **285**, 297, **55**
10	Golgi phosphoprotein 3-like isoform 1 (NP_666245.2)	343	338, 334, 148, 319, 35, 55, 80, 209, 78, **38**, 89, 337, 185

[a] Screened with the 2D-DIGE and mass spectrometry proteomics techniques.[42]

[b] Sites are listed according to their scores from LabCaS; those highlighted with underlined bold faces are consistent with the 10 outputs from SVM (RBF).[14]

this case, consistent with the insight shown in Figure 2. Among the 13 false positives out of the top 15 predictions by LabCaS (green), two sites are buried in the core structure regions. This data demonstrates that we can further improve the SP of the LabCaS algorithm when combining with the state-of-the-art protein structure predictions.

Prediction of calpain substrate cleavage sites in lysosomal membranes

It has been revealed that during mammary gland involution, calpain proteases play important roles in mediating epithelial-cell death.[41,42] It has also been suggested that calpains are involved in both apoptosis and necrotic cell death, where they first cleave substrates on the lysosomal membrane and then induce the intrinsic mitochondrial apoptotic pathway.[42] These findings support the new theory of calpain-mediated cleavage of new substrates from lysosomal membranes being crucial for mammary gland involution.[43] In consideration of this, it is critical to understand the cleavage mechanisms of calpain substrates in lysosomal membrane. Despite its importance, no experimentally verified cleavage sites have been reported for calpain substrates in the lysosomal membrane. To speed up the progress, we apply the LabCaS developed in this paper to predict the cleavage sites for 10 potential calpain targeted substrates in lysosome membrane, which were screened in a large-scale analysis by the 2D-DIGE and mass spectrometry proteomics techniques in lysosomal fraction from lactating mammary gland.[42] The predicted results from LabCaS at the highest threshold are tabulated in Table VI, which serve as a good base for further experimental designs and verifications. Particularly, the predicted sites of underlined bold face in Table VI are those overlapped with the 10 outputs from SVM (RBF).[14]

Large-scale identification of putative calpain substrate cleavage sites

One advantage of automatic prediction tools is the feasibility of large-scale cleavage site prediction. CaMPDB contains a set of potential calpain substrates and their cleavage sites determined using BLAST homology search and a predefined set of rules.[14] We have collected a total of 1973 putative substrates along with 2927 cleavage sites from CaMPDB. It has been noticed that in the 1973 putative substrate sequences, the average number of cleavage sites per sequence is $2927/1973 \approx 1.48$, which is lower than $367/129 \approx 2.85$ in the benchmark dataset of this study. This indicates that the cleavage sites of these 1973 substrates could be under-predicted in the current version of CaMPDB. For example, the calpain substrate of Src substrate cortactin protein (CaMPDB recoded ID XSB0288) is predicted to have one cleavage site of Lys^{336}↓Thr^{337} in CaMPDB by using the BLAST homology search and a defined set of rules, but four cleavage sites of Lys^{336}↓Thr^{337}, Lys^{346}↓Thr^{347}, Arg^{351}↓Ala^{352}, and Ala^{358}↓Lys^{359} were reported by experiments.[44] These observations suggest that more information should be provided on these 1973 substrates. We thus apply LabCaS to predict the potential cleavage sites for the 1973 putative substrates, and the predicted results are available at http://www.csbio.sjtu.edu.cn/bioinf/LabCaS/Data.htm. Here, we further compared the top 5 predicted outputs from Lab-

Table VII
Comparisons Between the Top Five Prediction Outputs for 1973 Calpain Substrates from LabCaS and the Original Records in CaMPDB

	The 1st	The 2nd	The 3rd	The 4th	The 5th	Total
Predicted sites	1973	1973	1973	1973	1973	9865
Overlapped sites with records in CaMPDB	1328	399	223	186	130	2266
Percentages	$\frac{1328}{2927} = 45.37\%$	$\frac{399}{2927} = 13.63\%$	$\frac{223}{2927} = 7.62\%$	$\frac{186}{2927} = 6.35\%$	$\frac{130}{2927} = 4.44\%$	$\frac{2266}{2927} = 77.42\%$

CaS with those cleavage sites recorded in CaMPDB as shown in Table VII. As can be seen from Table VII, there are a total of 1328 sites overlapping with the 1st ranked predicted site with the original CaMPDB records. Taking all the top 5 LabCaS's outputs into consideration, the overlapping rate is 77.42%. These results demonstrate the high confidences of the LabCaS predictions. In addition, they can provide more important complementary information for updating and understanding the knowledge of the 1973 substrates in the current database.

DISCUSSIONS

In order to estimate the false positive rates of the predictors, we create a control dataset by collecting sequences according to following steps: (1) Only the proteins in mitochondrion subcellular location are selected from the Swiss-Prot database since previous reports have shown that calpain proteins are mainly located in the cytoplasm and nucleus localizations.[45,46] (2) Proteins with less than 50 AAs have been removed because they could be fragments. (3) Proteins annotated with keywords of transcription factors, receptors, and enzymes are removed because currently identified calpain substrates mainly belong to these families.[3] (4) The sequence redundancy of in the control dataset and to the training dataset is removed at the cut-off 30% with the CD-HIT method.[36] (5) 100 non-redundant sequences are randomly selected as the final tested control dataset, which consists of 32,947 noncleavable sites and zero cleavage sites.

The final control dataset is respectively submitted to the three web-severs, LabCaS, GPS-CCD[16] and SVM(RBF)[14] for predictions. Table VIII gives the results.

Table VIII
Comparison of False Positive Rates of LabCaS with GPS-CCD and SVM(RBF) on the Control Dataset

Method	Threshold	False positive rates
LabCaS	0.0037	$\frac{1575}{32947} = 4.78\%$
GPS-CCD	High	$\frac{1431}{32947} = 4.34\%$
LabCaS	0.0026	$\frac{2873}{32947} = 8.72\%$
GPS-CCD	Medium	$\frac{2956}{32947} = 8.97\%$
LabCaS	0.0020	$\frac{4238}{32947} = 12.86\%$
GPS-CCD	Low	$\frac{4574}{32947} = 13.88\%$
SVM(RBF)[a]	–	$\frac{1000}{32947} = 3.04\%$

[a] A fixed false positive rate in this test since ten predicted sites will be outputted from SVM(RBF) for every submitted query sequence.

These results show that at the 3 decision thresholds corresponding to specificities of 95, 90, and 85%, the estimated false positive rates of LabCaS are 4.78, 8.72, and 12.86% respectively, and the values of GPS-CCD are 4.34, 8.97, and 13.88%. Though the listed false positive rate of SVM(RBF) is the lowest of 3.04%, the reason is that 10 predicted sites will be outputted from SVM(RBF) for every submitted query sequence, meaning it is a fixed rate in this test. Comparing LabCaS with GPS-CDD, we find that LabCaS predicts a little more false positives than GPS-CDD at the 95% SP cut-off, but performs better at the other two thresholds. Two potential ways are expected to be helpful for lowering the false positive rates in existing predictors: (1) A two-layer model should be developed where the proteolyzed proteins by calpains can be recognized in the first layer before it is fed into the second layer for cleavable residues prediction. (2) We have shown an example in the case study that the modeled protein 3D structure with the I-TASSER software can provide valuable information for screening the false positives. Hence, a hybrid model by combining the sequences and modeled 3D structures is a promising way to enhance the predictions of whether a protein can be proteolyzed by calpains and where the cleaving will happen.

CONCLUSION

In this study, we formulated the prediction of calpain substrate cleavage sites as a sequence labeling problem that was achieved by the CRFs algorithm and presented a novel ensemble method called LabCaS. LabCaS is robust to the extreme imbalance in positive and negative samples in the training dataset. Improvements of the performances by fusing multiple features have been observed demonstrating calpain substrate recognition and proteolysis are not controlled by a single determinant but by multiple ones. As an implementation of our approach, LabCaS is freely available for academic use at http://www.csbio.sjtu.edu.cn/bioinf/LabCaS, which is anticipated to become a powerful tool for *in silico* identification of calpain substrate cleavage sites. One of the important future directions is the investigation of a proper post-processing approach to further screen the false-positive predictions.

ACKNOWLEDGMENTS

The authors thank Dr. Jouko Virtanen and Mr. Brandon Govindarajoo for reading through the manuscript, and the anonymous reviewers for suggestions and comments which helped improving the quality of this paper.

REFERENCES

1. Storr SJ, Carragher NO, Frame MC, Parr T, Martin SG. The calpain system and cancer. Nat Rev Cancer 2011;11:364–374.
2. Bertipaglia I, Carafoli E. Calpains and human disease. Subcell Biochem 2007;45:29–53.
3. Franco SJ, Huttenlocher A. Regulating cell migration: calpains make the cut. J Cell Sci 2005;118:3829–3838.
4. Croall DE, Ersfeld K. The calpains: modular designs and functional diversity. Genome Biol 2007;8:218.
5. Zatz M, Starling A. Calpains and disease. N Engl J Med 2005;352:2413–2423.
6. Ono Y, Shimada H, Sorimachi H, Richard I, Saido TC, Beckmann JS, Ishiura S, Suzuki K. Functional defects of a muscle-specific calpain, p94, caused by mutations associated with limb-girdle muscular dystrophy type 2A. J Biol Chem 1998;273:17073.
7. Horikawa Y. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 2000;26:502–502.
8. Friedrich P, Bozoky Z. Digestive versus regulatory proteases: on calpain action in vivo. Biol Chem 2005;386:609–612.
9. Cuerrier D, Moldoveanu T, Davies PL. Determination of peptide substrate specificity for mu-calpain by a peptide library-based approach: the importance of primed side interactions. J Biol Chem 2005;280:40632–40641.
10. Tompa P, Buzder-Lantos P, Tantos A, Farkas A, Szilagyi A, Banoczi Z, Hudecz F, Friedrich P. On the sequential determinants of calpain cleavage. J Biol Chem 2004;279:20775–20785.
11. Banik NL, Chou CH, Deibler GE, Krutzch HC, Hogan EL. Peptide bond specificity of calpain: proteolysis of human myelin basic protein. J Neurosci Res 1994;37:489–496.
12. Boyd SE, Pike RN, Rudy GB, Whisstock JC. Garcia de la Banda M. PoPS: a computational tool for modeling and predicting protease specificity. J Bioinform Comput Biol 2005;3:551–585.
13. Verspurten J, Gevaert K, Declercq W, Vandenabeele P. SitePredicting the cleavage of proteinase substrates. Trends Biochem Sci 2009;34:319–323.
14. duVerle D, Takigawa I, Ono Y, Sorimachi H, Mamitsuka H. CaMPDB: a resource for calpain and modulatory proteolysis. Genome Inform 2010;22:202–213.
15. duVerle D, Ono Y, Sorimachi H, Mamitsuka H. Calpain Cleavage Prediction Using Multiple Kernel Learning. PloS One 2011;6:e19035.
16. Liu Z, Cao J, Gao X, Ma Q, Ren J, Xue Y. GPS-CCD: a novel computational program for the prediction of calpain cleavage sites. Plos One 2011;6:e19001.
17. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, Williamstown; 2001. p 282–289.
18. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990;18:6097–6100.
19. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–1190.
20. Song J, Tan H, Shen H, Mahmood K, Boyd SE, Webb GI, Akutsu T, Whisstock JC. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics 2010;26:752–760.
21. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33(Web Server issue):W72–W76.
22. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell 2008;134:866–876.
23. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725–738.
24. Xu D, Zhang J, Roy A, Zhang Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 2011;79 (Suppl 10):147–160.
25. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17.
26. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
27. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
28. Granseth E, von Heijne G, Elofsson A. A study of the membrane-water interface region of membrane proteins. J Mol Biol 2005;346:377–385.
29. Mak MW, Wang W, Kung SY. Fusion of Conditional Random Field and SignalP for Protein Cleavage Site Prediction. Annual Summit and Conference, Sapporo; 2009. p 716–721.
30. Fan YX, Song J, Shen HB, Kong X. PredCSF: an integrated feature-based approach for predicting conotoxin superfamily. Protein Pept Lett 2011;18:261–267.
31. Savojardo C, Fariselli P, Alhamdoosh M, Martelli PL, Pierleoni A, Casadio R. Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization. Bioinformatics 2011;27:2224–2230.
32. Hammersley J, Clifford P. Markov field on finite graphs and lattices, 1971. Available at:http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf. Accessed on 1 June 2012.
33. Sutton C, McCallum A. An introduction to conditional random fields for relational learning. Introduction to statistical relational learning; 2006. p 93–128.
34. Zhou ZH, Yu Y. Ensembling local learners through multimodal perturbation. IEEE Trans Syst Man Cybern B Cybern 2005;35:725–735.
35. Xu L, Amari S. Combining classifiers and learning mixture-of-experts. Encyclopedia Artif Intell 2009:318–326.
36. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–1659.
37. Johnson GV, Jope RS, Binder LI. Proteolysis of tau by calpain. Biochem Biophys Res Commun 1989;163:1505–1511.
38. Liu MC, Kobeissy F, Zheng W, Zhang Z, Hayes RL, Wang KK. Dual vulnerability of tau to calpains and caspase-3 proteolysis under neurotoxic and neurodegenerative conditions. ASN Neuro 2011;3:e00051.
39. Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. Proteins 2009;77Suppl 9:100–113.
40. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score= 0.5? Bioinformatics 2010;26:889–895.
41. Orrenius S, Zhivotovsky B, Nicotera P. Regulation of cell death: the calcium-apoptosis link. Nat Rev Mol Cell Biol 2003;4:552–565.
42. Arnandis T, Ferrer-Vicens I, Garcia-Trevijano ER, Miralles VJ, Garcia C, Torres L, Vina JR, Zaragoza R. Calpains mediate epithelial-cell death during mammary gland involution: mitochondria and lysosomal destabilization. Cell Death Differ 2012; 19:1536–48.
43. Kreuzaler PA, Staniszewska AD, Li W, Omidvar N, Kedjouar B, Turkson J, Poli V, Flavell RA, Clarkson RW, Watson CJ. Stat3 controls lysosomal-mediated cell death in vivo. Nat Cell Biol 2011;13:303–309.
44. Perrin BJ, Amann KJ, Huttenlocher A. Proteolysis of cortactin by calpain regulates membrane protrusion during cell migration. Mol Biol Cell 2006;17:239–250.
45. Suzuki K, Hata S, Kawabata Y, Sorimachi H. Structure, activation, and biology of calpain. Diabetes 2004;53 (Suppl 1):S12–S18.
46. Gil-Parrado S, Popp O, Knoch TA, Zahler S, Bestvater F, Felgentrager M, Holloschi A, Fernandez-Montalvan A, Auerswald EA, Fritz H, Fuentes-Prior P, Machleidt W, Spiess E. Subcellular localization and in vivo subunit interactions of ubiquitous mu-calpain. J Biol Chem 2003;278:16336–16346.