ORIGINAL RESEARCH

Procleave: Predicting Protease-specific Substrate Cleavage Sites by Combining Sequence and Structural Information

Fuyi Li [1,2,#], **Andre Leier** [3,#], **Quanzhong Liu** [4], **Yanan Wang** [1,2], **Dongxu Xiang** [1,4], **Tatsuya Akutsu** [5], **Geoffrey I. Webb** [2], **A. Ian Smith** [1,6], **Tatiana Marquez-Lago** [3,*], **Jian Li** [7,*], **Jiangning Song** [1,2,6,*]

[1] Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia
[2] Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia
[3] School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35233, USA
[4] College of Information Engineering, Northwest A&F University, Yangling 712100, China
[5] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[6] ARC Centre of Excellence in Advanced Molecular Imaging, Monash University, Melbourne, VIC 3800, Australia
[7] Biomedicine Discovery Institute and Department of Microbiology, Monash University, Melbourne, VIC 3800, Australia

Abstract Proteases are enzymes that cleave and hydrolyse the peptide bonds between two specific amino acid residues of target substrate proteins. **Protease**-controlled proteolysis plays a key role in the degradation and recycling of proteins, which is essential for various physiological processes. Thus, solving the substrate identification problem will have important implications for the precise understanding of functions and physiological roles of proteases, as well as for therapeutic target identification and pharmaceutical applicability. Consequently, there is a great demand for bioinformatics methods that can predict novel substrate cleavage events with high accuracy by utilizing both sequence and structural information. In this study, we present Procleave, a novel bioinformatics approach for predicting protease-specific substrates and specific cleavage sites by taking into account both their sequence and 3D structural information. Structural features of known cleavage sites were represented by discrete values using a LOWESS data-smoothing optimization method,

---

which turned out to be critical for the performance of Procleave. The optimal approximations of all structural parameter values were encoded in a **conditional random field** (CRF) computational framework, alongside sequence and chemical group-based features. Here, we demonstrate the outstanding performance of Procleave through extensive benchmarking and independent tests. Procleave is capable of correctly identifying most cleavage sites in the case study. Importantly, when applied to the human structural proteome encompassing 17,628 protein structures, Procleave suggests a number of potential novel target substrates and their corresponding cleavage sites of different proteases. Procleave is implemented as a webserver and is freely accessible at http://procleave.erc.monash.edu/.

Introduction

Protease-specific cleavage is a ubiquitous type of irreversible post-translational modification (PTM) that occurs when proteases specifically cleave the peptide bonds between the P1 and P1′ sites of target proteins or peptide substrates [1]. Numerous experimental studies indicate that proteolytic cleavage plays a critical role in a variety of developmental and physiological processes, including cell cycle, pathway regulation, and protein degradation. On the other hand, the dysregulation of proteases is associated with numerous diseases [2]. Thus, it is very important to identify protease-specific substrate cleavage sites, as such knowledge can provide deeper insights into the mechanisms and biological functions of proteases, which in turn might lead to novel therapeutic targets and pharmaceutical applicability. However, current existing experimental methods for protease substrate cleavage site identification are expensive, labour-intensive, and time-consuming. Therefore, the development of cost-effective computational approaches for precise prediction of protease-specific proteolytic events is very important. Such tools can not only provide high-quality predictions of target substrates for a specific protease, but also guide hypothesis-driven experimental efforts to identify substrate specificity and associated biological functions of proteases.

Due to the importance and the benefits of computational predictions of protease-specific target substrates, over the past two decades, more than 20 computational methods have been proposed [3,4]. In our recent review paper, we categorized these methods into two major groups according to the employed methodologies: (i) sequence-scoring function-based methods, such as PoPS [5], SitePrediction [6], and CAT3 [7], and (ii) machine learning methods, such as Pripper [8], Cascleave [9], PROSPER [10], LabCaS [11], ScreenCap3 [12], Cascleave 2 [13], iProt-Sub [14], and PROSPERous [15]. These publicly available computational tools have successfully guided experiments in finding novel cleavage sites and obtaining a better understanding of protease–substrate interactions.

A number of encouraging studies have been done regarding the development of computational methods and tools for predicting protease-specific cleavage sites. However, all of these existing prediction methods are developed based on protein sequences and they are only used for predicting the cleavage sites from substrate sequences. Previous studies have shown that protease cleavage sites are primarily distributed in loop regions of the substrate proteins, while cleavage within other structural regions of substrate proteins, such as α-helices and β-sheets, is also possible [16–18]. These findings indicate protease substrate cleavage specificity at the secondary structure

(SS) level. The majority of existing predictors did not consider the structure-level preference and parameters, which can potentially improve the prediction performance and also help better understand the biological functions of proteases.

In this study, we introduce Procleave to fill the knowledge gap outlined above and enhance protease substrate cleavage site prediction by incorporating 3D structural features of substrate cleavage segments. More specifically, Procleave uses the data curated from the MEROPS database [19] and maps substrate sequences to PDB structures by performing BLAST search, thereby generating an extensive 3D structural substrate dataset. Multi-faceted sequence and structural features are then extracted, which are further integrated into a novel conditional random field (CRF) algorithm with a data-smoothing framework to train cleavage site prediction models. A comprehensive performance test confirms that smoothed structural features combined with sequence-based features can greatly improve the prediction performance. Subsequently, we implement a webserver for 27 major proteases, taking advantage of the findings in this study, and make it publicly accessible.

Method

Overall framework

Figure 1 provides an overview of the Procleave framework. Five major steps are involved in the construction and evaluation of Procleave. At the first step, *i.e.*, data collection and pre-processing, the benchmark training and independent test datasets were collected from MEROPS [19]. At the second step, multi-faceted sequence features and 3D structure features were generated. At the third step, a novel integrative CRF framework was developed for model training and optimization. At the fourth step, the trained CRF models were further evaluated and validated by performing the independent test. A performance comparison with currently existing methods was also conducted. At the final step, the Procleave webserver was implemented to facilitate public use.

Dataset collection and pre-processing

The experimentally verified protein substrate cleavage annotations for training and benchmarking Procleave were extracted from the MEROPS database (Release 9.0) [19]. MEROPS is a public resource and knowledgebase for experimentally validated protease substrates and cleavage sites, which is accessible via https://www.ebi.ac.uk/merops/. To develop reliable predic-
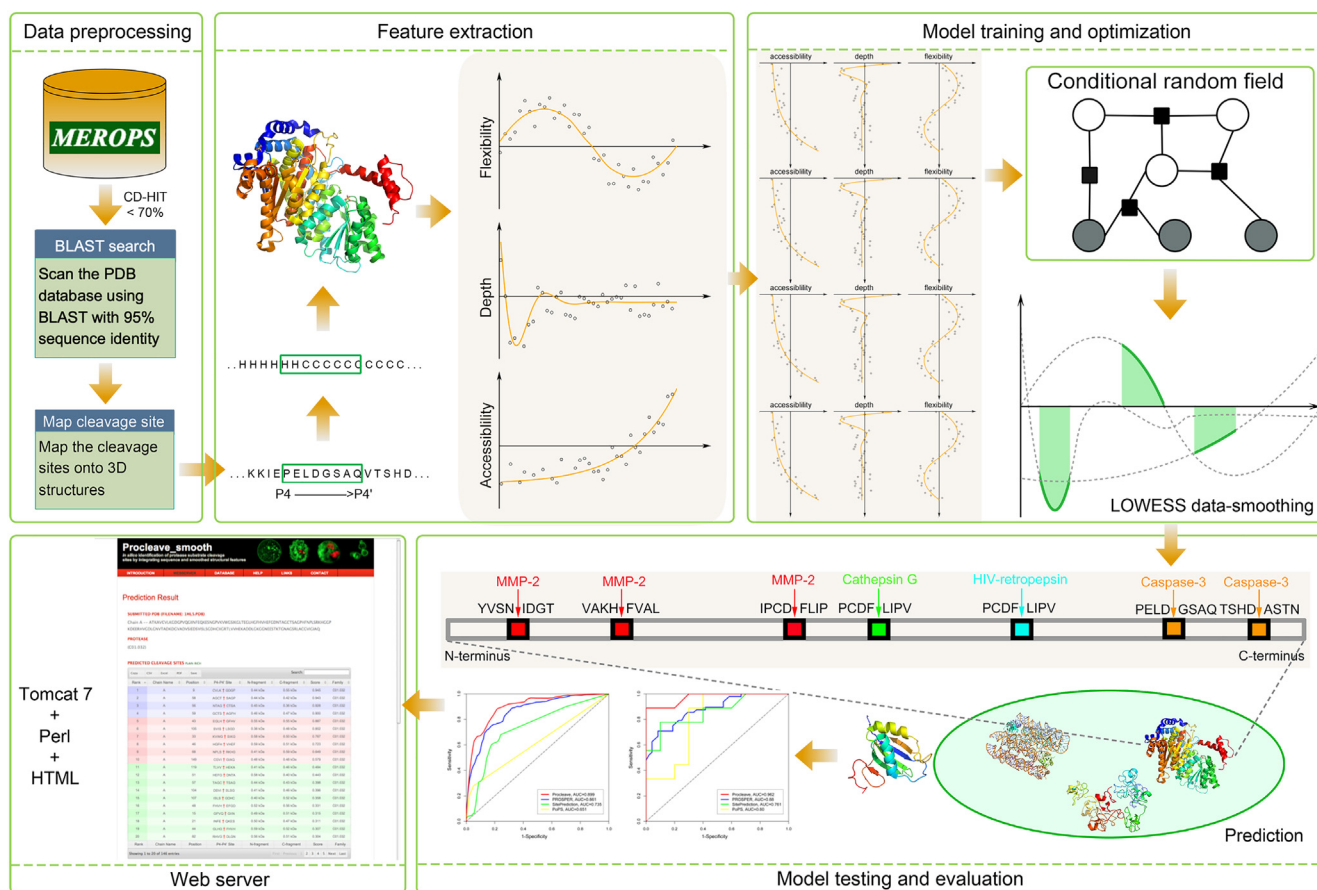
Figure 1   The overall framework of Procleave
There are five major steps in the framework of Procleave, including data pre-processing, feature extraction, model training and optimization, model testing and evaluation, as well as web server development.

tion models and objectively evaluation the model performance, we discarded highly homologous sequences from the initial substrate dataset with a sequence identity (SI) threshold of 70% between any two substrate protein sequences. This avoids overestimating the prediction performance in cross-validation tests. It is noticeable that a number of existing studies used SI cut-off values of 70% [9,14,15] or a higher, e.g., 80% [12]. The MEROPS database was recently updated (Release 12.0, 26-April-2019) and we decided to use all the newly added protease substrates and cleavage sites as the independent test dataset to assess the performance of trained Procleave models and conduct the performance comparison with existing methods. In addition, in order to perform a more fairly independent test, we used a stricter SI threshold (30%) to remove the sequence redundancy in the independent test dataset. CD-HIT [20] was applied to remove the redundant sequences between the independent test datasets and training datasets at the SI threshold of 30%. This ensures that any two substrates in the training and independent test datasets have a SI of < 30%. A statistical summary of both benchmark and independent test datasets is provided in Tables S1 and S2, respectively. Subsequently, the remaining sequences were mapped to PDB [21] by performing PSI-BLAST [22] to search against the PDB sequence database (using the 'pdbaa' file) with three iter-

ations, with an e-value of 10^{-3}, and a SI threshold of 95%. We only retained the X-ray crystallography (X-ray) structures, while nuclear magnetic resonance (NMR) and electron microscopy (EM) structures were discarded. After this procedure, all substrate cleavage sites were mapped onto respective 3D structures using our in-house Perl script and all of these cleavage sites were used as positive samples to train the Procleave models. Sites that have been not annotated as cleavage sites in substrate proteins were considered as negative samples. Accordingly the same number of negative sites was randomly selected as that of the positive samples. In this study, a gallery of all mapped respective 3D structures with visualized cleavage sites can be accessed at http://procleave.erc.monash.edu/gallery.html.

Feature engineering

The substrate cleavage site prediction task can be regarded as a binary classification problem. Each cleavage site is denoted as an N-dimensional feature vector $F = \{f_1, f_2, \ldots, f_N\}$. Three major types of features were extracted, namely structural features, sequence features, and chemical group features. A detailed description of each feature type is presented below.

Structural features

In this study, several different types of 3D structural descriptors were extracted from the P4–P4′ local windows surrounding cleavage sites, which include:

(1) Protrusion and depth index. We calculated the protrusion (cx) index and the depth index by CX [23] and DPX [24] programs, respectively.

(2) Solvent accessibility. Naccess [25] was employed to compute the absolute and relative solvent accessibility features using the default settings. There are five types of solvent accessibility features, including all atoms, total side chain, main chain, non-polar side chain, and all-polar side chain solvent accessibility.

(3) Packing. Packing was calculated using the method proposed previously [26].

(4) Molecular surface accessibility. Molecular surfaces are either solvent-accessible surfaces (SAS) or solvent-excluded surfaces (SES). Both were calculated by the MSMS program [27].

(5) Secondary structure features. The DSSP program [28] was used to calculate the secondary structure features. These encompass hydrogen bonds, secondary structures (eight classes were transformed to three classes, *i.e.*, α-helix, β-sheet, and coil), and backbone torsion angles. The HBPLUS v.3.06 program [29] was used to calculate the hydrogen bond.

(6) Solvent exposure properties. Half-sphere exposure properties were also used as candidate features. They were extracted using the Biopython package [30]. They included contact number (CN), the number of Cα atoms in the upper half-sphere (HSEAU), the number of Cα atoms in the lower half-sphere (HSEAD), the number of Cβ atoms in the upper half-sphere (HSEBU), and the number of Cβ atoms in the lower half-sphere (HSEBD).

(7) B-factor. The B-factor values of all atoms were extracted from PDB files, with the average values being used as the input feature [31].

Sequence features

We employed the binary encoding scheme to extract and encode sequence features. In particular, a sliding window approach (P4–P4′) centred around the potential cleavage sites was used to extract the local sequence features. Each amino acid (AA) residue was encoded by a binary vector with 20 dimensions. Therefore, the total number of dimensions of the obtained vector is $8 \times 20 = 160$.

Chemical group features

Apart from structural and sequence features, the chemical/structural groupings of AAs were also used as candidate features. According to the chemical/structural properties, 20 AAs were clustered into eight chemical groups [32]. These include sulfur-containing (residues C and M), aliphatic 1 (residues A, G, and P), aliphatic 2 (residues I, L, and V), acidic (residues D and E), basic (residues H, K, and R), aromatic (residues F, W, and Y), amide (residues N and Q), and small hydroxy (residues S and T) residues. Then, these eight chemical groups were encoded as input features using the one-hot

encoding. The total number of dimensions of the chemical group features is $8 \times 8 = 64$ (for any 8-AA window).

Model training and optimization

CRFs and LOWESS data smoothing

CRFs are a type of undirected graphical models originally introduced by Lafferty et al. [33] to deal with the segmentation and labelling tasks of text sequences. CRFs have been proven to be effective in a number of applications with structured outputs, such as information extraction, image processing, and parsing. A CRF is an undirected graph, and its nodes can be categorized as two disjoint sets, namely the observed variables X and the output variables Y. Its principle is to define a conditional probability distribution $p(Y|X)$ over label sequences $Y = \{y_1, y_2, \cdots, y_n\}$, given the observational sequence $X = \{x_1, x_2, \cdots, x_n\}$. Y is a sequence of hidden state variables that needs to be inferred given the observation. $y_1, \cdots, y_i, y_{i+1}, \cdots, y_n$ are structured to form a chain, with an edge between each y_i and y_{i+1}. The distribution of the network has the following form:

$$p(Y|X) = \frac{1}{Z(X)} \exp \left(\sum_{k=1}^{K} \lambda_k f_k(y_i, y_{i-1}, x_i) \right), \qquad (1)$$

where $Z(X) = \sum_{y_i} \exp(\sum_{k=1}^{K} \lambda_k f_k(y_i, y_{i-1}, x_i))$, K denotes the number of class labels (*e.g.*, $K = 2$ stands for a two-class classification), λ_k is the weight vector of features, and f_k is the function of features for the clique $\{y_i, y_{i-1}, x_i\}$.

Since a CRF does not have the assumption for the distribution of inputs and, instead, finds the decision boundary directly, it may be considered as an extended version of logistic regression to model sequential data. CRFs have been applied to bioinformatics rather recently and have delivered promising results, such as for gene prediction [34] and phosphorylation sites prediction [32]. CRFs can capture sophisticated dependencies and combine information from different aspects. The specific advantages of CRFs are well-suited for incorporating structural information into a cleavage site prediction algorithm. Many of the structural parameters are closely related, and structural parameters contain important information for determining the potential cleavage site that might be better captured by CRFs.

In this study, our input variables X are the structural, sequence, and chemical group features of a given substrate peptide and the output variables are binary labels corresponding to "cleavage site" or "non-cleavage site". The CRF models were trained by maximizing the likelihood that the positive samples of a training set were cleavage sites, given their structural, sequence, and chemical group features. We used the open source package CRF++ (version 0.54) and, as part of the CRF implementation, used Boolean feature functions to train the models. As the Boolean feature functions evaluate one of the two states of being true or false for a feature appearing at an exact position, all structural features are regarded in the form of discrete instead of continuous values during model training. In addition, considering that the substrate cleavage depends on the overall 3D shape or neighbourhood of multiple AAs, structural features recognized by cleavage sites, *e.g.*, the overall shape of the P4–P4′ segment surrounding

the potential cleavage sites, we combined CRF with a LOW-ESS data-smoothing approach [35] and examined whether cleavage site prediction could be further improved. Specifically, feature optimization first ran the LOWESS smoothing algorithm on the input vectors of each structural feature. Then the resulting vectors were discretized into equally sized bins to group similar values for use by the Boolean feature functions. Algorithm 1 describes the detailed procedures of the LOWESS smoothing algorithm.

Algorithm 1 LOWESS data-smoothing algorithm
Input:
    Range value, *range*; Initial feature array, *@iniArry*;
Output:
    Smoothed feature array, *@smoothedArray*;
1: **for** each $i \in [1, \#iniArry]$ **do**
2:     $avey = avex = norm = weight = 0$;
3:     **if** $Expression_N$
4:       **for** each $j \in [S_N]$ **do**
5:         **calculate** *weight*;
6:         $avey+ = weight \times iniArray[j]$;
7:         $avex+ = weight \times j$;
8:         $norm+ = weight$;
9:       **end for**;
10:      $avey = avey/norm; avex = avex/norm$;
11:      $mtop = mbot = 0$;
12:      **for** each $k \in [S_N]$ **do**
13:        $weight = \left(1 - \left|((i-k)/(2 \times range - i + 1))^3\right|\right)^3$;
14:        $mtop+ = weight \times ((k - avex) \times (array[k] - avey))$;
15:        $mbot+ = weight \times ((k - avex)^2)$;
16:      **end for**;
17:      $smoothedArray[i] = \left(\dfrac{mtop}{mbot) \times i + \left(avey - \left(\frac{mtop}{mbot) \times avex;}\right)\right.}\right.$
18:     **end if**;
19: **end for**;

The input to Algorithm 1 was the smoothing range *range* and the initial feature vector *@iniArry*, which needed to be smoothed and tuned. In this study, each type of structural feature was described by an 8-bit vector, where each bit was associated with the feature value of a local sliding window (P4–P4′) surrounding the potential cleavage site. The output of Algorithm 1 was the 8-bit vector *smoothedArray*. The smoothing procedure was performed in a 'for' loop. At step 1, *#iniArray* was the length of feature vector, which equals to eight. At the second step, four variables, namely *avey*, *avex*, *norm*, and *weight*, were set to 0. These variables represented the average value of *y* (*i.e.*, values of the features), the average value of *x* (positions of the feature vector), the normalization variable, and the weight of the variable, respectively. At step 3, the if statement has three different expressions $Expression_N$, which can be presented as:

$Expression_1$	$i - range < 0$;
$Expression_2$	$range + i > \#iniArray$;
$Expression_3$	Others.

For these three expressions, the range S_N of the parameter *range* in the step 4 and step 12 is different:

$Expression_1$	S_1	$[1, 2 \times range]$;
$Expression_2$	S_2	$[\#iniArray - 2 \times range, \#iniArray]$;
$Expression_3$	S_3	$[i - range, i + range]$.

Then, at step 5, the weight of the variable was calculated. The method used for calculating the variable weight is also different:

S_1	$weight = \left(1 - \left\|((i-j)/(2 \times range - i + 1))^3\right\|\right)^3$;
S_2	$weight = \left(1 - \left\|\left((i-j)/(x - \#iniArry - 2 \times range + 1)\right)^3\right\|\right)^3$;
S_3	$weight = \left(1 - \left\|\left((i-j)/(2 \times range + 1)\right)^3\right\|\right)^3$.

At steps 6 and 7, *weight* was used to calculate the normalized values of *x* and *y*. Then, *avey* and *avex* were updated at step 10 by dividing the normalization variable calculated at step 8. At step 11, the smoothed value of *mtop* and the smoothed bottom value of *mbot* were initialized to 0. At steps 12–16, these two variables were calculated and updated, and at step 17 the final output *smoothedArray* was generated according to these two values.

We set the smoothing range *range* from 1 to 5 and the bin number from 1 to 10, respectively, in this study. The smoothing procedure and the number of bins for each type of structural feature were optimized by maximizing the area under the curve (AUC) of the receiver operating characteristic (ROC) curves on the 5-fold cross validation test using the benchmark dataset. In this way, by optimizing the smoothing range and the number of bins for each of the structural features, the optimal combination of smoothing and discretization that best represented structural features of all samples in the training set could be determined.

Performance evaluation

To assess the performance of the Procleave models and benchmark it with other currently available methods, a set of five commonly used performance measures were applied, including sensitivity (Sn), specificity (Sp), precision, accuracy (Acc), Matthew's correlation coefficient (MCC), and AUC. Sn, Sp, Precision, Acc, and MCC are defined as:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where *TP*, *TN*, *FP*, and *FN* represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. Moreover, we plotted the ROC curves and accordingly calculated the AUCs, as a primary measure to assess the prediction performance of Procleave models and all compared methods.

Results and discussion

Characterization of structural features in the proximity of cleavage sites

To better understand the structural determinants surrounding cleavage sites of different proteases, we examined the structural features of protease cleavage sites using the curated PDB structure datasets. Bar graphs for a total of 27 proteases presented in Figure 2 (9 proteases) and Figure S1 (18 proteases) show the secondary structure preferences of protease-specific substrates across the P4–P4′ sites surrounding the cleavage sites. As shown in these figures, different protease cleavage sites generally have distinctly different secondary structure preferences. However, on the other hand, some proteases also share similar secondary structure preferences. For instance, the P4–P4′ site surrounding cleavage sites of caspase-3, granzyme

B (human) (Figure 2E and H), cathepsin S, caspase-6, meprin α subunit, meprin β subunit, and LAST_MAM peptidase (Figure S1G, J, and L–N) are more likely to be located in loop regions than in helix and strand regions. In addition, the cleavage sites of most proteases can be found in all three types of secondary structures, except for those of necepsin-1, cathepsin L1 (*Fasciola* sp.), falcipain-2, and falcipain-3 (Figure S1D, F, H, and I). The cleavage sites of these four proteases are predominately found in helix and loop regions, but not in strands. The results are in good agreement with the findings of existing studies and suggest that proteases prefer to cleave within loop regions of substrate proteins, while cleavage within helix/sheet regions is also possible [16–18]. In addition, we plotted the boxplots for other structural features of positive samples (cleavage sites) for all 27 proteases. These results are provided in supplementary figures, including protrusion index (Figure S2), depth index (Figure S3), solvent accessibility calculated by Naccess (Figures S4–S13), packing (Figure S14),
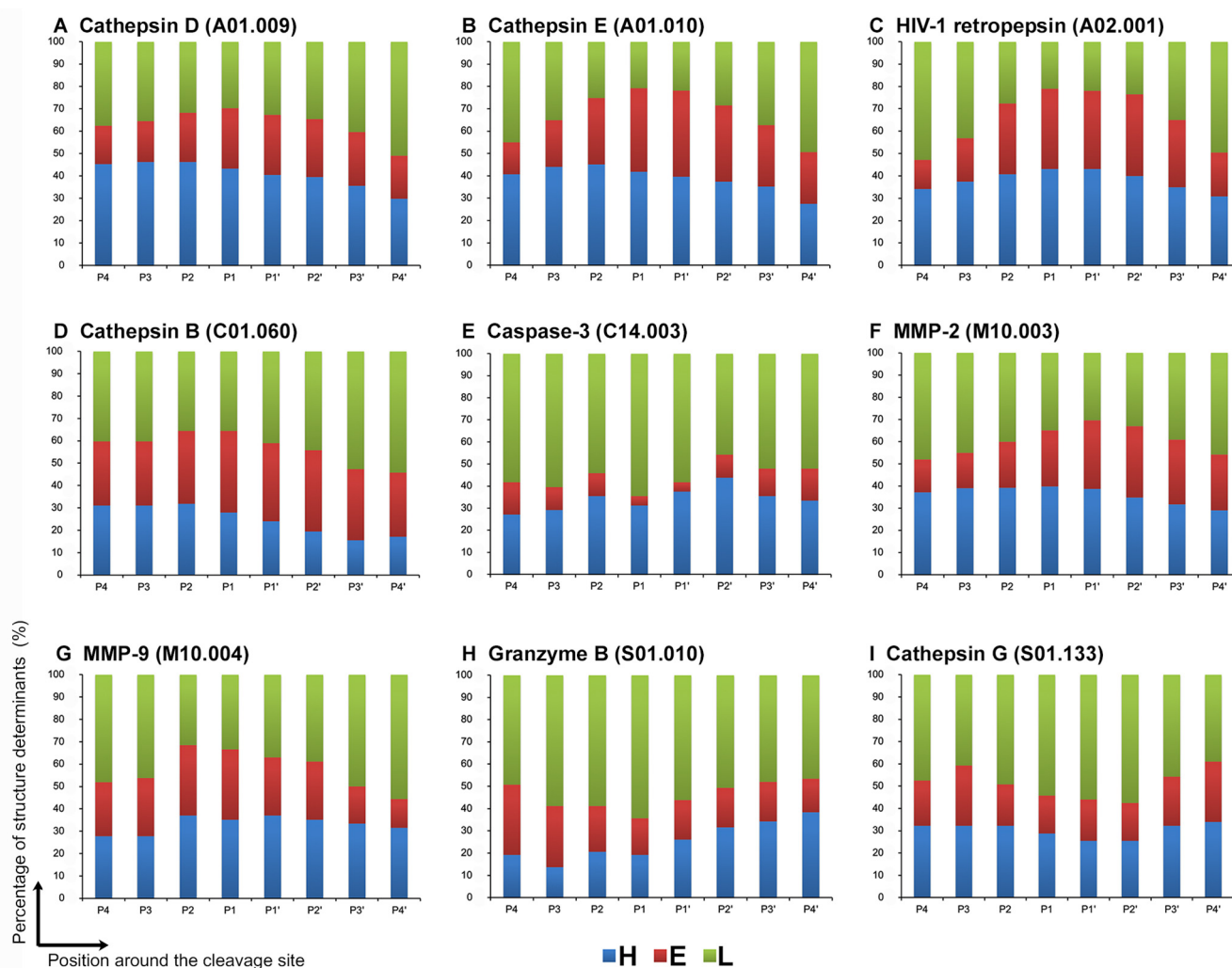
Figure 2 Structural determinants of the substrate specificity of nine proteases across the P4–P4′ cleavage sites
A. Cathepsin D. **B.** Cathepsin E. **C.** HIV-1 retropepsin. **D.** Cathepsin B. **E.** Caspase-3. **F.** MMP-2. **G.** MMP-9. **H.** Granzyme B (human). **I.** Cathepsin G. MMP, matrix metallopeptidase. The secondary structure information was extracted from DSSP results. H, helix; E, strand; L, loop.

solvent exposure properties (Figures S15 and S16), solvent accessibility calculated by DSSP (Figure S17), backbone torsion angles (Figures S18 and S19), solvent exposure properties (Figures S20–24), B-factor (Figure S25), and hydrogen bonds (Figure S26).

Performance assessment

To examine how the structural features help to predict the cleavage sites and how our proposed feature smoothing algorithm improves the prediction performance of trained CRF models, we evaluated the performance of different types of feature combinations. The experiments were conducted by performing 10 times of 5-fold cross-validation tests using the benchmark datasets. The evaluated features/feature combinations include Seq only (using sequence features only), Seq + Chem (using sequence features together with chemical features), Seq + Chem + real structure (using sequence, chemical, and original structural features, without any smoothing), Seq + Chem + smooth DSSP (using sequence, chemical, and smoothed DSSP structural features), and Seq + Chem + smooth structure (using sequence, chemical, and smoothed structural features). Performance comparisons of different feature combinations in terms of AUC values (average AUC values of 10 times of 5-fold cross-validation tests) are shown in Figure 3 and Table S3.

From these results, several important observations can be made. The Seq + Chem + smooth structure models performed the best compared with all other feature combinations in terms of AUC values for 22 of the 27 tested proteases (see Figure 3 and detailed results in Table S3). Also, the Seq + Chem + smooth DSSP models achieved highest AUC values for meprin β and chymotrypsin A (bovine), while the Seq + Chem models achieved highest AUC values for cathepsin B and lysyl peptidase (bacteria). Seq only model performed the best for HIV-1 retropepsin. These results demonstrate that the sequence features and chemical group features are more relevant and important for the three proteases, while the structural features may not be useful for further improving the cleavage site prediction performance for these proteases. Not surprisingly, the Seq + Chem + real structure models performed the worst among all the compared feature combination models, because the Boolean feature functions of the CRF cannot deal properly with continuous values. This not only leads to the loss of some useful feature information, but also affects the model training.

In addition, to test and verify the statistical significance of AUC improvement by the Seq + Chem + smooth structure models, we conducted a student's *t*-test to compare the AUC values of different feature combination models trained with CRF. The *P* values of the student's *t*-test are given in Table 1, indicating that the AUCs of the Seq + Chem + smooth structure models were significantly ($P \leq 0.01$, marked in bold) higher than those of other models according to the pairwise tests. Feature combinations that achieved the best performance during each comparison test are underlined in Table 1. Furthermore, the AUC values of the Seq + Chem + smooth DSSP models were significantly higher than those of the Seq + Chem and the Seq + Chem + real structure models, while inconclusive with the Seq only models. Altogether, both the performance comparisons and pairwise *t*-test comparisons demonstrate that structural features smoothed by the LOWESS data smoothing algorithm can greatly help to boost the performance of CRF models. A possible explanation is that the LOWESS smoothing takes the structural variables defined over the cleavage segment P4–P4′ sites, and flattens the fluctuations of the structural variables over the eight AA residues of the cleavage sites. This makes intuitive sense because the structural variables are defined over the crystal structure of the protein, which represents only one of the many conformations that constitute the equilibrium ensemble of the protein in solution. In particular, the cleavage site is generally located on or near the surface of the protein, where the side chains of residues on the surface are particularly prone to fluctuations due to thermal contact with the water [16]. As such, a single value for the structural variables of a given AA residue will not be a fair representation, especially given that in crystal structures, sidechain conformations on the surface are often flush against symmetric repeats of the protein [16]. As such, the smoothing of the structural parameters provides a way to reduce these effects and a more appropriate representation of the structural determinants of cleavage sites.

Moreover, in order to further illustrate the advantage of CRF, we benchmarked the performance of CRF models with that of the other two popular machine learning algorithms, *i.e.*, support vector machine (SVM) and random forest (RF), on both the training and independent test datasets. The performance results on the 5-fold cross validation and independent tests are provided in Tables S3 and S4, respectively. As a result, the CRF models achieved the best performance across almost all comparative experiments on the training datasets. The only exceptions were the Seq + Chem + real structure feature for matrix metallopeptidase 2 (MMP-2) and the Seq + Chem feature for both astacin and meprin α, for which the RF models achieved the best prediction results. For the performance evaluation on the independent test, we applied the SVM and RF models trained using the Seq + Chem + smooth structure feature combinations, as the SVM and RF models trained on this feature combination performed the best compared to all the other feature combinations. The performance results on the independent tests confirm that the CRF models of Procleave achieve overall a better performance than SVM and RF models, for all 27 proteases examined. Taken together, the performance results on both 5-fold cross validation and independent tests demonstrate the superiority of the CRF framework, making it the model of choice for the development of Procleave.

Therefore, we accordingly built two prediction models for protease cleavage site prediction from both protein sequences and structures. We built the Procleave_sequence based on Seq + Chem feature combination models for protease cleavage site prediction from protein sequences; while the Procleave_smooth based on Seq + Chem + smooth structure feature combination was built for protease cleavage site prediction from protein structures.

Comparison with existing methods

We compared the performance of two variant models 'Procleave_sequence' and 'Procleave_smooth' against five existing tools, including PoPS, SitePrediction, PROSPER,
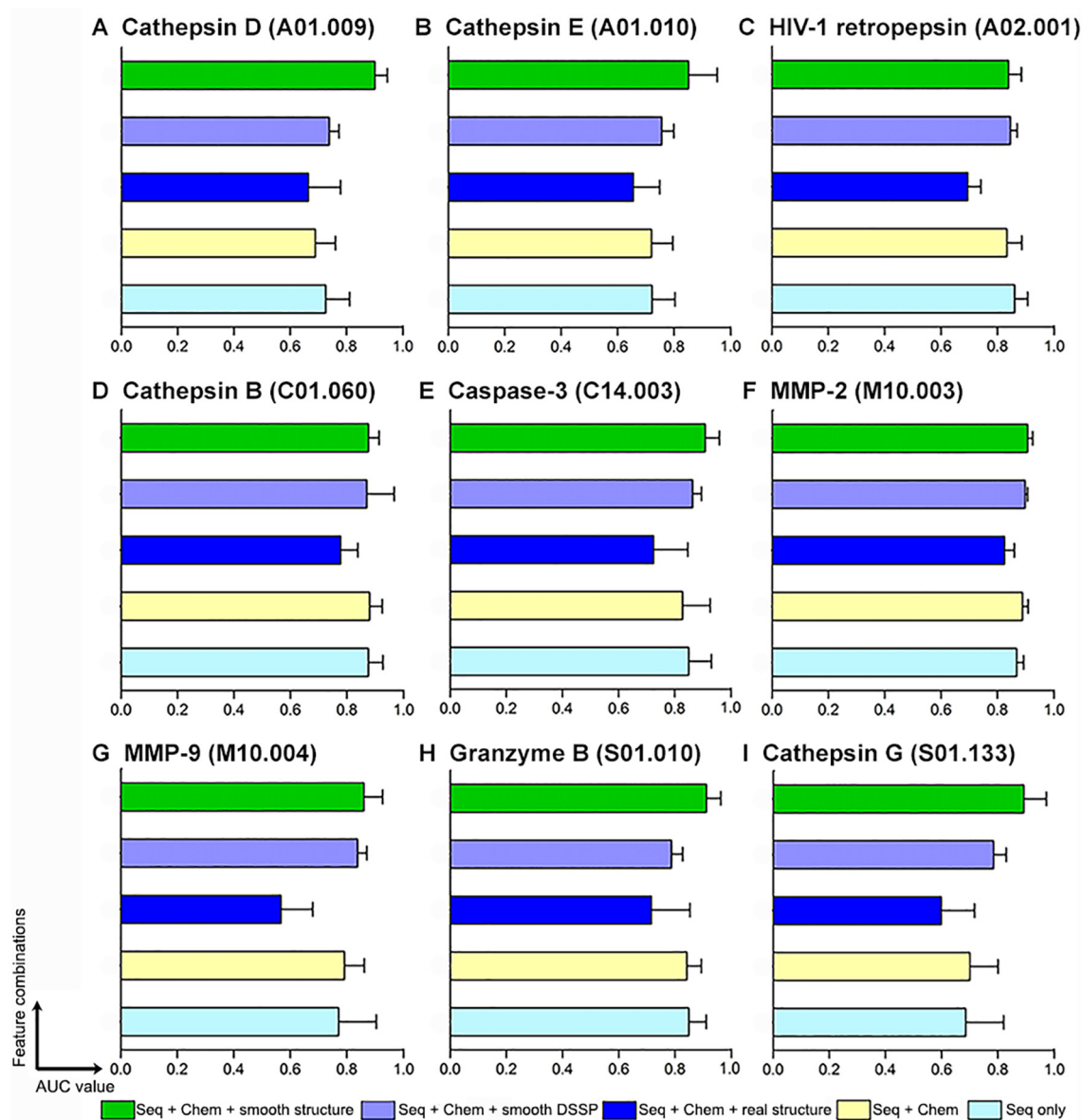
Figure 3 Performance comparison of CRF models trained using different feature combinations in terms of AUC values
A. Cathepsin D. **B.** Cathepsin E. **C.** HIV-1 retropepsin. **D.** Cathepsin B. **E.** Caspase-3. **F.** MMP-2. **G** MMP-9. **H.** Granzyme B (human). **I.** Cathepsin G. The evaluation was based on 10 times of 5-fold cross-validation tests on training datasets.

PROSPERous, and iProt-Sub, by performing the independent test. In order to avoid any potential bias and objectively assess the performance, we submitted the PDB sequences in the FASTA format in the independent test dataset to each of the webservers of these methods. The detailed performance results are summarized in Table S4. In addition to AUC, MCC, Acc, Sn, Sp, and precision are also provided and listed in Table S4, while ROC curves are presented in Figure 4.

Table 1   *P* values for pairwise *t*-test comparisons of prediction performance using different feature combinations

Feature combination	*P* value
Seq + Chem ± smooth DSSP *vs.* Seq only	0.10
Seq + Chem ± smooth DSSP *vs.* Seq + Chem	0.01
Seq + Chem ± smooth DSSP *vs.* Seq + Chem + real structure	1.12E−20
Seq + Chem + smooth structure *vs.* Seq only	1.91E−13
Seq + Chem + smooth structure *vs.* Seq + Chem	2.04E−16
Seq + Chem + smooth structure *vs.* Seq + Chem + real structure	3.94E−49
Seq + Chem + smooth structure *vs.* Seq + Chem + smooth DSSP	5.38E−09

Note: Tests were performed using AUC results of 10 times 5-fold cross-validation tests of all the 27 proteases examined.
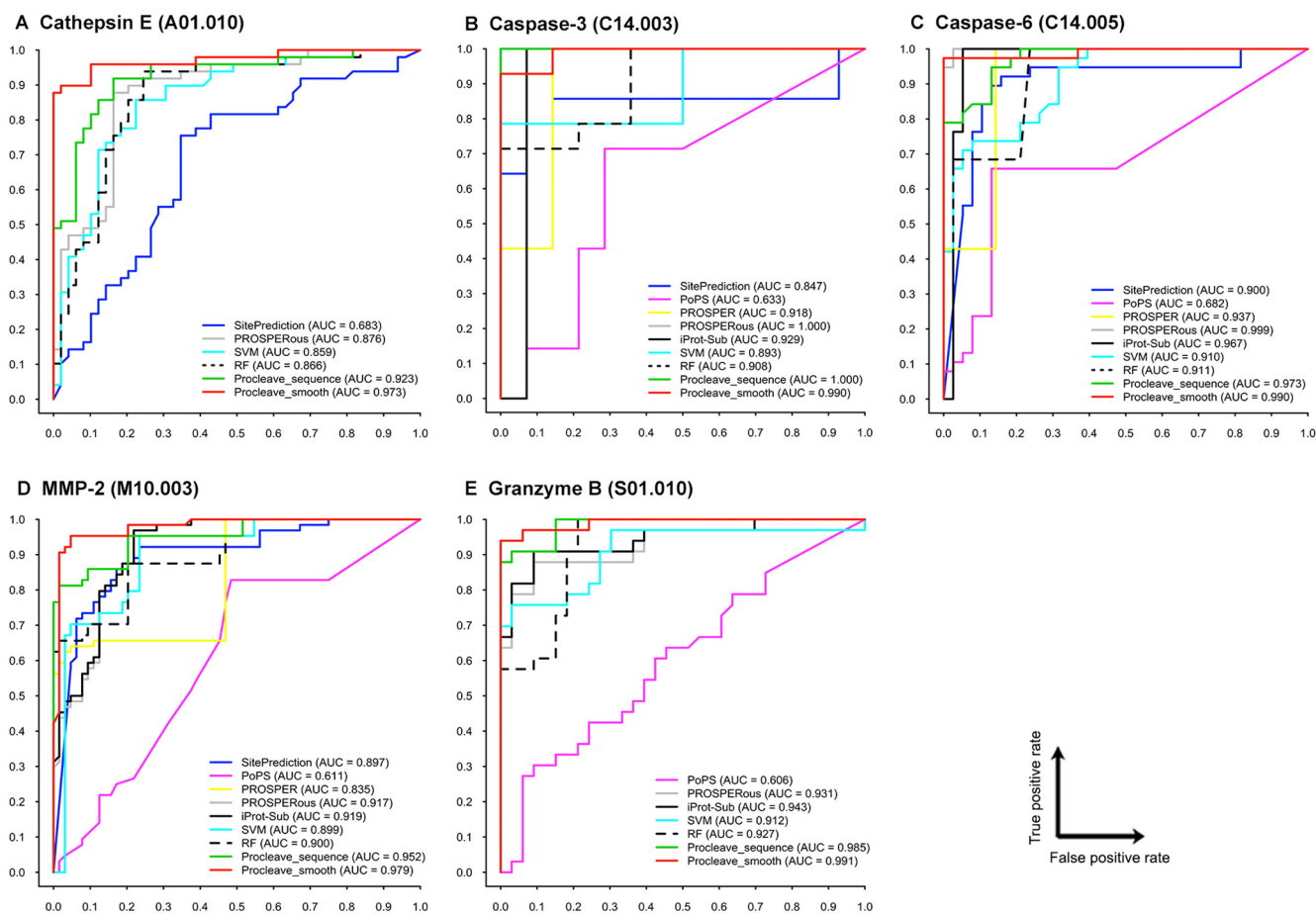
Figure 4   Comparison of cleavage site prediction performance of Procleave and other methods in terms of AUC values for 5 different proteases

A. Cathepsin E. **B.** Caspase-3. **C.** Caspase-6. **D.** MMP-2. **E.** Granzyme B. PoPS, PROSPER, and iProt-Sub cannot predict cleavage sites of cathepsin E; SitePrediction and PROSPER cannot predict cleavage sites of granzyme B. SVM and RF were included to test whether the conditional random field model employed in Procleave provides better performance.

Figure 4 displays the ROC curves of PoPS, SitePrediction, PROSPER, PROSPERous, iProt-Sub, Procleave_smooth, and Procleave_sequence on the independent test dataset. As the entries in the independent test dataset were obtained solely from the newly identified protease substrates and cleavage sites from the most-recent version of MEROPS (12.0) as compared to its previous version (release 9.0), the amount of newly added data was relatively small, and there was even fewer data remaining after mapping onto the PDB 3D structures. Therefore, only five proteases were used for the test, including cathepsin E, caspase-3, caspase-6, MMP-2, and granzyme B (human). As can be seen, Procleave_smooth (red line) performed the best and Procleave_sequence (green line) ranked second in terms of AUC for Cathepsin E (Figure 4A), MMP-2 (Figure 4D), and granzyme B (human) (Figure 4E). For caspase-3, Procleave_sequence and PROSPERous achieved the best performance (AUC = 1) and Procleave_smooth achieved the second highest AUC (0.990) (Figure 4B). While for caspase-6, PROSPERous achieved the highest AUC (0.999) value and Procleave_smooth ranked second (Figure 4C). To summarize, all these results demonstrate that Procleave is a reliable and powerful bioinformatics

approach that improves protease cleavage site prediction. In particular, there are three important factors that account for the good performance of Procleave. First, the high quality and comprehensive 3D structural substrate cleavage data provide solid foundation for the training of Procleave. Second, extracting useful and complementary 3D structural features as calculated by multiple software tools provides a better description of the characteristics of substrate cleavage sites. And lastly, processing initial 3D structural features using the LOWESS data-smoothing algorithm is necessary to enable CRF to learn the underlying rules and characteristics of protease-specific cleavage events.

Webserver implementation

To facilitate bioinformatics analyses of novel protease target substrates and cleavage sites, we implemented the CRF-based Procleave approach and developed a publicly available webserver for the wider research community. The Procleave webserver was implemented using HTML and Perl. The webserver is freely accessible at http://procleave.erc.monash.edu/. Procleave webserver is operated by Tomcat7 and configured

in a Linux server with an eight-core CPU, 500-GB hard disk and 16-GB memory. Both the Procleave_smooth and Procleave_sequence variant models are implemented on the web server. The web server requires two steps of inputs in order to make a prediction of the potential cleavage sites for the given protein. First, Procleave_smooth requires users to supply a protein 3D structure file (*.pdb file is preferred), while for Procleave_sequence models, users are required to input the FASTA formatted protein sequences. Second, users need to specify the PDB chain name and protease type in the case of submitting the 3D structure file. Each submission takes approximately 3–4 min to complete. The prediction outcome for the submitted structure file is returned on the result webpage. The prediction results can be exported in the CSV, Excel, and PDF formats. 3Dmol.js [36] is also employed for protein 3D structure visualization at the webserver. The predicted potential cleavage sites are labelled at their corresponding positions.
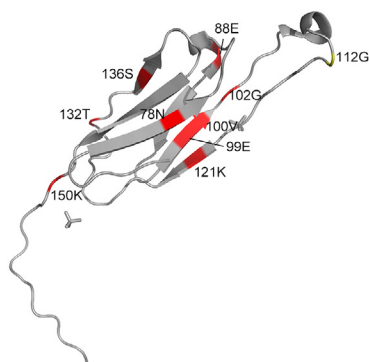
Structural proteome-wide prediction

Furthermore, we conducted a structural proteome-wide prediction of novel protease substrate cleavage sites (containing 17,628 human proteins extracted from the PDB database) by applying the Procleave_smooth model. The results are briefly summarized in this section. We applied an Sp threshold of 99% to all predictions [15,31,37,38] to generate a compendium of high-confidence predicted cleavage sites and then performed the statistical analyses. Statistics of the identified cleavage substrates and the predicted cleavage sites for the 27 different proteases are summarized in Table S5. The results of the identified cleavage substrates and their cleavage sites are also accessible at the Procleave webserver, which can be freely downloaded at http://procleave.erc.monash.edu/.
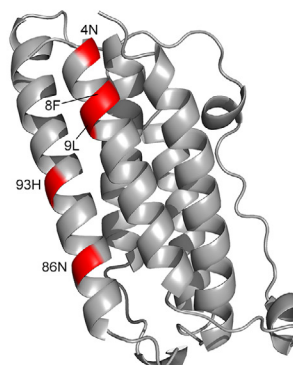
Case study

To illustrate the utility and capacity of Procleave, a case study of the protease-specific cleavage site prediction in four substrate proteins was conducted in this section. The four proteins were selected from the independent test dataset. The first protein is human αB crystalline (PDB ID: 3L1G, chain A), which functions as a chaperone and oligomeric assembly. It serves as a stability sensor and can recognize and bind to destabilized proteins in eye lens and other tissues [39]. The second protein is human interferon β (PDB ID:1AU1, chain A), which is the protein to
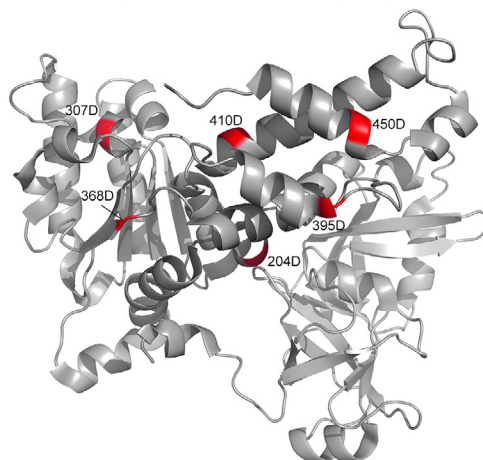

A αB-crystallin cleaved by MMP-9
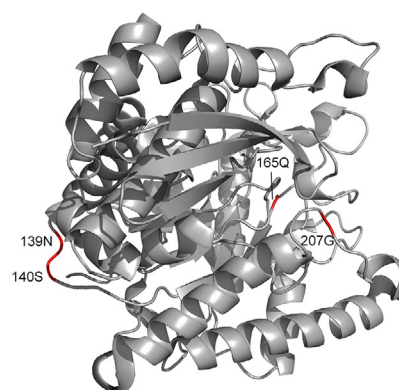**B** Interferon β cleaved by MMP-9
C ATPase p97 mutant cleaved by caspase-6
D Enolase 1 cleaved by meprin β

Figure 5 **Predicted cleavage sites of four substrate protein structures**
A. Human αB crystalline (PDB ID: 3L1G, chain: A) cleaved by MMP-9. **B.** Human Interferon β (PDB ID: 1AU1, chain: A) cleaved by MMP-9. **C.** ATPase p97 mutant (PDB ID: 3HU2, chain A) cleaved by caspase-6. **D.** Human enolase 1 (PDB ID: 3B97, chain A) cleaved by meprin β.

defend the cells from various viruses [40]. The third protein is an ATPase p97 mutant (PDB ID: 3HU2, chain A). ATPase p97 is one of the most abundant cytosolic proteins and can interact with different adaptor proteins involved in many cellular activities, including protein degradation, cell cycle regulation, and membrane fusion [41]. The fourth protein is human enolase 1 (PDB ID: 3B97, chain A), which is a glycolytic enzyme expressed in most tissues. A previous study indicates that this protein is involved in many diseases, including metastatic cancer, ischaemia, autoimmune disorders, and bacterial infection [42]. Structure scanning results and the predicted cleavage sites are shown in **Figure 5** and Table S6. All correctly predicted cleavage sites are highlighted in red. These prediction results of demonstrate that Procleave could correctly identify all the experimentally verified cleavage sites. These results suggest that Procleave is a useful tool and can be used to identify cleavage sites based on the 3D structural information of the substrate proteins.

Conclusion

In the present work, we developed Procleave, a new CRF approach, which combines both sequence and structural information to enhance the protease-specific cleavage site prediction. Procleave employs multi-faceted 3D structure-based features, in combination with a LOWESS smoothing optimization algorithm to train and optimize the CRF-based cleavage site prediction models for a protease. We conducted a comprehensive set of empirical benchmarking tests to benchmark the performance of CRF models built based on different combinations of sequence, chemical, and structural features. We also assessed the performance of Procleave with several state-of-the-art approaches. The comparison results demonstrate that Procleave outperforms these methods, and the LOWESS smoothing optimization is critical to the performance of Procleave. The aim of this study is to systematically investigate whether both sequence-derived and real 3D structural information can be integrated in a machine learning framework to improve the substrate cleavage site prediction for 27 major proteases. A user-friendly webserver of Procleave has been made available as an implementation of the proposed approach. All predicted cleavage sites of the human proteome with 3D the structure data available are provided for further protease biology research. We envisage that Procleave will become a useful tool in the future, facilitating community-wide hypothesis-driven experimental design and functional characterization studies. As a generally useful framework, the CRF-based methodology combined with the LOWESS smoothing optimization algorithm can be readily extended and applied to develop useful methods for predicting other important types of PTM sites [43–46] and functional sites that utilize 3D structural information in future work.

Data availability

The datasets and proteome-wide prediction results are publicly accessible at http://procleave.erc.monash.edu/gallery.html/.

Authors' contributions

JS, JL, and TML conceived the project and supervised the study. JS, FL, and AL designed the algorithm and drafted the manuscript. FL performed the machine learning experiments and analyzed the results. FL and YW analyzed the performance comparison results. FL, QL, and DX implemented the online webserver. TA, GIW, and AIS revised the manuscript critically for important intellectual content. All authors read, revised, and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2019.08.002.

ORCID

0000-0001-5216-3213 (Li F)
0000-0002-2647-2693 (Leier A)
0000-0002-0611-0556 (Liu Q)
0000-0002-3820-8443 (Wang Y)
0000-0002-2937-2786 (Xiang D)
0000-0001-9763-797X (Akutsu T)
0000-0001-9963-5169 (Webb GI)
0000-0002-4143-2892 (Smith AI)
0000-0003-3279-5592 (Marquez-Lago T)
0000-0001-7953-8230 (Li J)
0000-0001-8031-9086 (Song J)

References

[1] Overall CM, Blobel CP. In search of partners: linking extracellular proteases to substrates. Nat Rev Mol Cell Biol 2007;8:245–57.
[2] Turk B. Targeting proteases: successes, failures and future prospects. Nat Rev Drug Discov 2006;5:785–99.

[3] Li F, Wang Y, Li C, Marquez-Lago TT, Leier A, Rawlings ND, et al. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. Brief Bioinform 2018;20:2150–66.

[4] Li F, Chen J, Leier A, Marquez-Lago T, Liu Q, Wang Y, et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. Bioinformatics 2020;36:1057–65.

[5] 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;3:551–85.

[6] Verspurten J, Gevaert K, Declercq W, Vandenabeele P. SitePredicting the cleavage of proteinase substrates. Trends Biochem Sci 2009;34:319–23.

[7] Ayyash M, Tamimi H, Ashhab Y. Developing a powerful *in silico* tool for the discovery of novel caspase-3 substrates: a preliminary screening of the human proteome. BMC Bioinformatics 2012;13:14.

[8] Piippo M, Lietzen N, Nevalainen OS, Salmi J, Nyman TA. Pripper: prediction of caspase cleavage sites from whole proteomes. BMC Bioinformatics 2010;11:320.

[9] Song J, Tan H, Shen H, Mahmood K, Boyd SE, Webb GI, et al. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics 2010;26:752–60.

[10] Song J, Tan H, Perry AJ, Akutsu T, Webb GI, Whisstock JC, et al. PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. PLoS one 2012;7:e50300.

[11] Fan YX, Zhang Y, Shen HB. LabCaS: labeling calpain substrate cleavage sites from amino acid sequence using conditional random fields. Proteins 2013;81:622–34.

[12] Fu SC, Imai K, Sawasaki T, Tomii K. ScreenCap3: improving prediction of caspase-3 cleavage sites using experimentally verified noncleavage sites. Proteomics 2014;14:2042–6.

[13] Wang M, Zhao XM, Tan H, Akutsu T, Whisstock JC, Song J. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. Bioinformatics 2014;30:71–80.

[14] Song J, Wang Y, Li F, Akutsu T, Rawlings ND, Webb GI, et al. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief Bioinform 2018;20:638–58.

[15] Song J, Li F, Leier A, Marquez-Lago TT, Akutsu T, Haffari G, et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics 2018;34:684–7.

[16] Kazanov MD, Igarashi Y, Eroshkin AM, Cieplak P, Ratnikov B, Zhang Y, et al. Structural determinants of limited proteolysis. J Proteome Res 2011;10:3642–51.

[17] Belushkin AA, Vinogradov DV, Gelfand MS, Osterman AL, Cieplak P, Kazanov MD. Sequence-derived structural features driving proteolytic processing. Proteomics 2014;14:42–50.

[18] Timmer JC, Zhu W, Pop C, Regan T, Snipas SJ, Eroshkin AM, et al. Structural and kinetic determinants of protease substrates. Nat Struct Mol Biol 2009;16:1101.

[19] Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res 2018;46:D624–32.

[20] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.

[21] Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 2011;39:D392–401.

[22] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[23] Pintar A, Carugo O, Pongor S. CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics 2002;18:980–4.

[24] Pintar A, Carugo O, Pongor S. DPX: for the analysis of the protein core. Bioinformatics 2003;19:313–4.

[25] Hubbard SJ. NACCESS: a program for calculating accessibilities. Department of Biochemistry and Molecular Biology, University College of London. 1992. http://wolf.bms.umist.ac.uk/naccess.

[26] Nishikawa K, Ooi T. Radial locations of amino acid residues in a globular protein: correlation with the sequence. J Biochem 1986;100:1043–7.

[27] Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 1996;38:305–20.

[28] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.

[29] McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–93.

[30] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3.

[31] Li F, Li C, Revote J, Zhang Y, Webb GI, Li J, et al. GlycoMine (struct): a new bioinformatics tool for highly accurate mapping of the human *N*-linked and *O*-linked glycoproteomes by incorporating structural features. Sci Rep 2016;6:34595.

[32] Dang TH, Van Leemput K, Verschoren A, Laukens K. Prediction of kinase-specific phosphorylation sites using conditional random fields. Bioinformatics 2008;24:2857–64.

[33] Lafferty J, McCallum A, Pereira FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc18th Int Conf Mach Learn 2001;2001:282–9.

[34] DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE. Conrad: gene prediction using conditional random fields. Genome Res 2007;17:1389–98.

[35] Meysman P, Dang TH, Laukens K, De Smet R, Wu Y, Marchal K, et al. Use of structural DNA properties for the prediction of transcription-factor binding sites in *Escherichia coli*. Nucleic Acids Res 2010;39 e6.

[36] Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. Bioinformatics 2015;31:1322–4.

[37] Li F, Li C, Wang M, Webb GI, Zhang Y, Whisstock JC, et al. GlycoMine: a machine learning-based approach for predicting *N*-, *C*- and *O*-linked glycosylation in the human proteome. Bioinformatics 2015;31:1411–9.

[38] Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou KC, et al. PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. J Theor Biol 2018;443:125–37.

[39] Laganowsky A, Benesch JL, Landau M, Ding L, Sawaya MR, Cascio D, et al. Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. Protein Sci 2010;19:1031–43.

[40] Karpusas M, Nolte M, Benton CB, Meier W, Lipscomb WN, Goelz S. The crystal structure of human interferon beta at 2.2-A resolution. Proc Natl Acad Sci U S A 1997;94:11813–8.

[41] Tang WK, Li D, Li CC, Esser L, Dai R, Guo L, et al. A novel ATP-dependent conformation in p97 N-D1 fragment revealed by crystal structures of disease-related mutants. EMBO J 2010;29:2217–29.

[42] Kang HJ, Jung SK, Kim SJ, Chung SJ. Structure of human alpha-enolase (hENO1), a multifunctional glycolytic enzyme. Acta Crystallogr D Biol Crystallogr 2008;64:651–7.

[43] Li F, Fan C, Marquez-Lago TT, Leier A, Revote J, Jia C, et al. PRISMOID: a comprehensive 3D structure database for post-

translational modifications and mutations with functional impact. Brief Bioinform 2020;21:1069–79.

[44] Li F, Zhang Y, Purcell AW, Webb GI, Chou KC, Lithgow T, et al. Positive-unlabelled learning of glycosylation sites in the human proteome. BMC Bioinformatics 2019;20:112.

[45] Li F, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW, et al. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics 2018;34:4223–31.

[46] Jia C, Zuo Y, Zou Q. O-GlcNAcPRED-II: an integrated classification algorithm for identifying *O*-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. Bioinformatics 2018;34:2029–36.