OXFORD

Briefings in Bioinformatics, 20(2), 2019, 638–658

doi: 10.1093/bib/bby028
Advance Access Publication Date: 19 April 2018
Review Article

iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites

Jiangning Song,* Yanan Wang,* Fuyi Li, Tatsuya Akutsu, Neil D. Rawlings, Geoffrey I. Webb and Kuo-Chen Chou

*The authors contributed equally to this work.
Corresponding authors: Jiangning Song, Monash Centre for Data Science and Monash Biomedicine Discovery Institute, Monash University, Melbourne, Victoria 3800, Australia. Tel.: +61-3-9902-9304. E-mail: Jiangning.Song@monash.edu; Geoffrey I. Webb, Monash Centre for Data Science, Monash University, Melbourne, Victoria 3800, Australia. Tel.: +61-3-9905-3296. E-mail: Geoff.Webb@monash.edu; Kuo-Chen Chou, Gordon Life Science Institute, Boston, Massachusetts, USA. E-mail: kcchou@gordonlifescience.org

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

Abstract

Regulation of proteolysis plays a critical role in a myriad of important cellular processes. The key to better understanding the mechanisms that control this process is to identify the specific substrates that each protease targets. To address this, we have developed iProt-Sub, a powerful bioinformatics tool for the accurate prediction of protease-specific substrates and their cleavage sites. Importantly, iProt-Sub represents a significantly advanced version of its successful predecessor, PROSPER. It provides optimized cleavage site prediction models with better prediction performance and coverage for more species-specific proteases (4 major protease families and 38 different proteases). iProt-Sub integrates heterogeneous

Jiangning Song received his BEng and DEng degrees from Jiangnan University, China. He is affiliated with the Monash Centre for Data Science, Faculty of Information Technology, Monash Biomedicine Discovery Institute and ARC Centre for Excellence in Advanced Molecular Imaging, Monash University, Melbourne, Australia. His research interests include bioinformatics, computational biology, machine learning, data mining and pattern recognition.
Yanan Wang received his MEng degree from Shanghai Jiao Tong University, China. His research interests are bioinformatics, machine learning, data mining and pattern recognition.
Fuyi Li received his BEng and MEng degrees in Software Engineering from Northwest A&F University, Yangling, China. He is currently a PhD student at the Biomedicine Discovery Institute, Monash University, Melbourne, Australia. His research interests are computational biology, bioinformatics, machine learning and data mining.
Tatsuya Akutsu received his DEng degree in Information Engineering in 1989 from University of Tokyo, Japan. Since 2001, he has been a professor in the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. His research interests include bioinformatics and discrete algorithms.
Neil D. Rawlings received his bachelor degree in Biological Sciences from Aston University, Birmingham, UK and his PhD from the Open University, Milton Keynes, UK. Since 1996, he and Alan J. Barrett have been the originators, developers and curators of the MEROPS database of proteolytic enzymes, their inhibitors and substrates, which provides the globally accepted classification of proteolytic enzymes and their inhibitors. He is one of the editors of the *Handbook of Proteolytic Enzymes* (Elsevier, 2013), and he is currently a curator for the InterPro database at the European Bioinformatics Institute, Cambridge, UK.
Geoffrey I. Webb received his PhD degree in 1987 from La Trobe University, Australia. He is Director of the Monash Centre for Data Science and Professor in Faculty of Information Technology at Monash University, Australia. He is a leading data scientist and the only Australian to have been Program Committee Chair of the two leading Data Mining conferences, ACM SIGKDD and IEEE ICDM. He received the 2016 Australian Computer Society's ICT Researcher of the Year Award, the 2016 Australasian Artificial Intelligence Distinguished Research Contributions Award, a 2014 Australian Research Council Discovery Outstanding Researcher Award and the 2013 IEEE ICDM Service Award and was elevated to IEEE Fellow in 2015. His research interests include machine learning, data mining, computational biology and user modeling.
Kuo-Chen Chou received his DSc degree in 1984 from Kyoto University, Japan. He is the founder and chief scientist of Gordon Life Science Institute. He is also a Distinguished High Impact Professor and Advisory Professor of several Universities. His research interests are in computational biology and biomedicine, protein structure prediction, low-frequency internal motion of protein/DNA molecules and its biological functions, diffusion-controlled reactions of enzymes, as well as graphic rules in enzyme kinetics and other biological systems.
Submitted: 2 January 2018; **Received (in revised form):** 2 March 2018

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

sequence and structural features and uses a two-step feature selection procedure to further remove redundant and irrelevant features in an effort to improve the cleavage site prediction accuracy. Features used by iProt-Sub are encoded by 11 different sequence encoding schemes, including local amino acid sequence profile, secondary structure, solvent accessibility and native disorder, which will allow a more accurate representation of the protease specificity of approximately 38 proteases and training of the prediction models. Benchmarking experiments using cross-validation and independent tests showed that iProt-Sub is able to achieve a better performance than several existing generic tools. We anticipate that iProt-Sub will be a powerful tool for proteome-wide prediction of protease-specific substrates and their cleavage sites, and will facilitate hypothesis-driven functional interrogation of protease-specific substrate cleavage and proteolytic events.

Key words: protease; substrate; cleavage site; sequence analysis; machine learning; five-step rule

Introduction

Proteolytic cleavage is one of the few irreversible posttranslational modifications. It plays a key role in numerous developmental and physiological processes, including digestion, protein degradation, endocrine signaling and cell division [1]. This process is controlled by proteases (also known as peptidases or proteinases) that selectively cleave the peptide bonds between amino acids in specific protein or peptide substrates. Proteases have central roles in 'life or death' processes. Through the highly selective proteolytic processing, proteases can precisely regulate a myriad of biological processes across all living organisms [1]. In addition, these are also many other proteases involved in protein degradation rather than processing, for example cathepsin D and cathepsin B. Pepsin, trypsin and chymotrypsin also come into this category, even though they have a defined specificity because they degrade so many substrates, most of which are foreign to the body [2]. The malfunction or deregulation of proteases results in many pathological conditions [3]. For example, proteases are often associated with cancer invasion and metastasis because of their ability to degrade the extracellular matrix [4–8]. Intriguingly, proteases can function as part of an extensive network of proteolytic interactions through interacting with other important signaling pathways involving other protein substrates and enzymes, termed the 'protease web' [9].

Our knowledge of the mechanisms that regulate and control the proteolytic processing of proteases remains limited. The precise understanding of the biological function of a protease requires the identification of the complete repertoire of its natural substrates and corresponding substrate cleavage sites [10, 11]. The specificity of proteases can vary significantly, depending on the protease and the active sites, with the cleavage site selectivity ranging from preferences for limited and specific amino acids at specific positions, to more general preferences with little discrimination. Current experimental methods for proteolytic cleavage characterization include one-dimensional and two-dimensional gel-based methods (used for identifying the substrates) [12], N-terminal peptide identification methods (for identifying both substrates and cleavage sites), methods using mass spectrometry, as well as quantitation methods of proteolysis to better understand the dynamics and extent of proteolytic events such as the TAILS method [13]. Despite the advances of these experimental methods, they are labor intensive, expensive and time-consuming, and are often limited to the investigation of one protease each time. In this context, it is highly desirable to develop cost-effective computational methods that can be used to identify the target substrates for a specific protease and to facilitate the characterization of substrate specificity and the function of proteases.

The importance and value of the *in silico* identification of protease target substrates and cleavage sites has led to the development of a variety of computational methods for predicting protease-specific substrates and cleavage sites. A number of computational studies have suggested that substrate cleavage sites (sites surrounding the cleavage P1 sites) targeted by proteases present unique structural and physicochemical properties that vary across different proteases, which can be exploited to predict potential cleavage sites [11, 14–24]. However, the most successful computational methods use a combination of these features with other complementary features [10, 14–16, 25–31], achieving overall accuracies of 70–90% for most of the proteases under investigation. A consensus resulting from such computational methods is that machine learning algorithms that take into consideration integrated heterogeneous information can be used to build more accurate predictive models for most of the proteases under investigation. Often a consensus scoring mechanism is performed using scoring function-based techniques or machine learning techniques. The former includes PeptideCutter [32], PoPS [33] and SitePrediction [34]. The latter has gained significant interest in recent years and includes CASVM [35], Cascleave [15], Pripper [36], Cascleave 2.0 [30], PROSPER [29, 31] and PROSPERous [37]. For the substrate cleavage site prediction of specific proteases, some *ad hoc* consensus schemes can also be effective, including GraBCas [38], CaSPredictor [39] and GPS-CCD [40].

At the end of 2012, we published PROSPER (PROtease substrate SPecificity servER), a bioinformatic tool for predicting target substrates and their specific cleavage sites for 23 proteases [29]. It represented the first comprehensive server capable of predicting cleavage sites of multiple proteases within a single substrate sequence using machine learning techniques. To date, the PROSPER server has attracted >25 000 unique users worldwide and has processed >60 000 job submissions since its inception. Here, we build on this previous work to develop a new computational method, termed iProt-Sub to address the problem of identifying the most probable protease-specific substrates and their detailed cleavage sites from the substrate sequence information. According to the well-known Chou's five-step rule [41] in developing a useful predictor, we need to accomplish the following: (1) benchmark data set construction, (2) protein sample formulation, (3) operating algorithm, (4) evaluating expected accuracy and (5) Web server establishment. In this work, we have considerably improved the design of iProt-Sub package for each of the five procedures.

More specifically, using a well-prepared benchmark data set, iProt-Sub extracts a wide range of sequence-derived structural, physicochemical and evolutionary information, which is further integrated into a common machine learning framework in the form of support vector machine (SVM) classifiers, to identify and rank potential substrate cleavage sites in a protease-specific manner. The cleavage site prediction models are

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

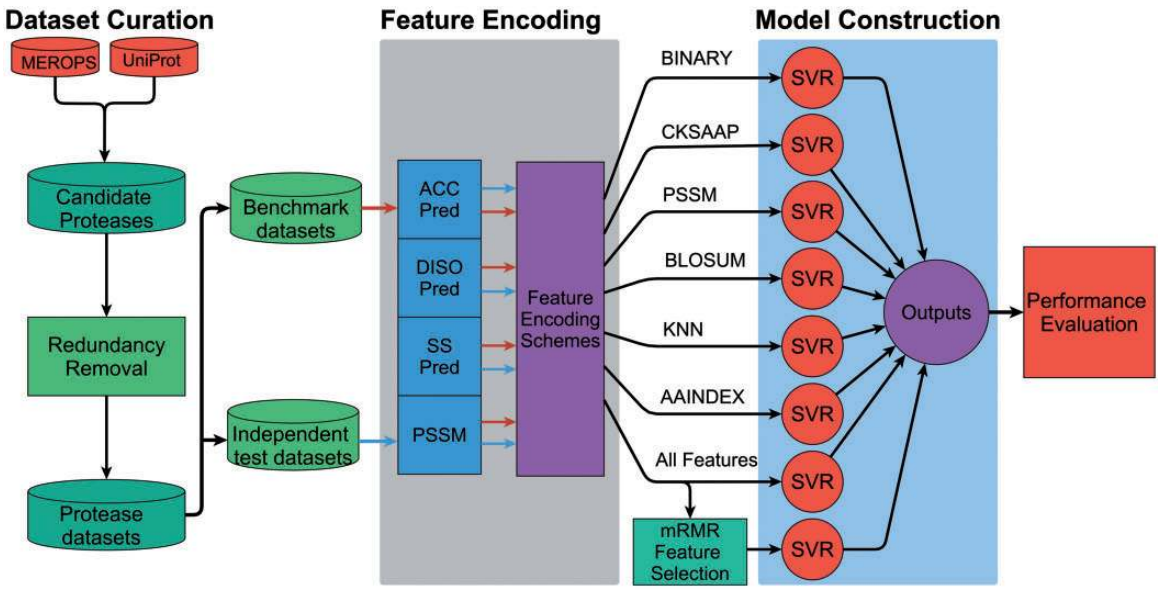

Figure 1. The workflow of the iProt-Sub methodology. There exist four major stages during the development of iProt-Sub, including Data set curation, Feature extraction and encoding, Model construction and Performance evaluation. Refer to the main text for a detailed description of each of the major stages. 'All features' included all the 11 types of extracted features (a detailed list is shown in Table 2).

trained and optimized to achieve best-performing prediction by performing a 5-fold cross-validation test. Benchmarking experiments indicated that the iProt-Sub method compared favorably with recently published methods. Moreover, mapping of the protease-specific cleavage target substrates at the proteome-wide scale was highly accurate and selective. iProt-Sub is accessible through a user-friendly Web application available at http://iProt-Sub.erc.monash.edu/. The Web application of iProt-Sub features a powerful and convenient graphic interface that allows the visualization and analysis of the predicted cleavage site within the same protein by different proteases simultaneously. The implemented iProt-Sub server thus represents a centralized Web resource for accurate *in silico* prediction of protease-specific substrates and their cleavage sites.

Materials and methods

Overall workflow of iProt-Sub

iProt-Sub represents an advanced version of PROSPER [29]. Importantly, the improvement of iProt-Sub over PROSPER is reflected by the following: (1) larger coverage of more proteases. iProt-Sub can be used to predict protease-specific substrates and cleavage sites for 38 different proteases, whereas PROSPER covered 23 proteases; (2) use of a wider range of sequence-derived features. iProt-Sub uses 11 diverse types of sequence-based features (4562-dimensional); (3) application of a more effective feature selection technique to filter out irrelevant and noisy features. iProt-Sub uses the mRMR (minimum redundancy maximum relevance) [42] algorithm to identify more informative features to enhance the predictive performance: (4) improved predictive performance. Through an effective feature extraction, selection and model learning strategy, iProt-Sub consistently achieves improved predictive performance for predicting the substrate cleavage sites for all tested proteases and (5) completely redesigned interface. The new iProt-Sub Web server now provides a more user-friendly and interactive interface that enhances user experience. The overall flowchart of the iProt-Sub methodology is shown in Figure 1.

Data sets

Numerous studies have suggested that a high-quality, well-established data set is crucial for training a robust and reliable prediction model of protease cleavage sites [37, 43–45]. In this study, we constructed a well-prepared benchmark data set for assessing the predictive performance of our method and other existing methods. For this purpose, we used the MEROPS database [46], which is a comprehensive information resource for proteases, their substrates and inhibitors. Only experimentally verified substrate sequences and cleavage sites were retrieved. The annotations of experimentally verified cleavage sites and the corresponding proteases that cleave the target substrates were extracted from MEROPS, while the annotations of protein identifiers of the substrates and their sequence information were extracted from UniProt [47]. In particular, exopeptidases (aminopeptidases, carboxypeptidases, etc.) and oligopeptidases were generally not included, which is consistent with our previous study [29]. As we are more interested in predicting cleavages within native proteins, peptidases that work at pH extremes and are likely to degrade only denatured proteins were also excluded [29].

To avoid potential model bias and overfitting, we performed sequence clustering and homology reduction using the CD-HIT program [48]. We removed sequence redundancy in the retrieved data set, so that any two sequences in the benchmark data set and independent test data set have a sequence identity of $<$70%, which is in accordance with previous studies [15, 30, 37, 49, 50]. After this procedure, we only retained those proteases that had $\geq$50 experimentally verified cleavage sites. Finally, we ended up with 38 proteases with a total of 3688 substrates and 6637 cleavage sites. A complete list of these substrate sequences and their cleavage sites can be found at the iProt-Sub website. A statistical summary of the curated data sets in this study is shown in Table 1.

In this study, five of the six of the substrate sequences in the resulting data set obtained above were randomly selected as the benchmark training data set, while the remaining one of the six of the data set was used as the independent test data set. The

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

Table 1. Statistical summary of the substrate data sets curated in this study

No	MEROPS ID	Protease name	Number of substrates	Number of cleavage sites	Number	MEROPS ID	Protease name	Number of substrates	Number of cleavage sites
1	A01.009	Cathepsin D	23	59	20	M16.002	Insulysin	6	50
2	C01.032	Cathepsin L	17	63	21	S01.010	Granzyme B (human-type)	410	515
3	C02.001	Calpain-1	30	61	22	S01.017	Kallikrein-related protease 5	31	59
4	C02.002	Calpain-2	17	66	23	S01.131	Elastase-2	45	133
5	C14.003	Caspase-3	251	373	24	S01.135	Granzyme A	44	57
6	C14.004	Caspase-7	48	64	25	S01.136	Granzyme B (rodent-type)	143	157
7	C14.005	Caspase-6	58	165	26	S01.139	Granzyme M	491	707
8	C14.009	Caspase-8	37	56	27	S01.233	Plasmin	42	89
9	M10.001	Matrix metallopeptidase-1	21	52	28	S01.251	Kallikrein-related peptidase 4	78	80
10	M10.002	Matrix metallopeptidase-8	23	85	29	S08.071	Furin	56	75
11	M10.003	Matrix metallopeptidase-2	35	115	30	A01.009 (mouse)	Cathepsin D	342	579
12	M10.004	Matrix metallopeptidase-9	43	290	31	A01.010 (mouse)	Cathepsin E	655	1216
13	M10.005	Matrix metallopeptidase-3	44	132	32	C14.001 (mouse)	Caspase-1	47	53
14	M10.008	Matrix metallopeptidase-7	42	142	33	S01.010	Granzyme B (human-type)	77	88
15	M10.009	Matrix metallopeptidase-12	23	178	34	S01.136	Granzyme B (rodent-type)	168	201
16	M10.013	Matrix metallopeptidase-13	23	90	35	S08.073	PCSK2 peptidase (mouse)	21	68
17	M10.014	Membrane-type matrix metallopeptidase-1	36	92	36	S08.109	KPC2-type peptidase (*Caenorhabditis elegans*)	34	115
18	M12.221	ADAMTS4 peptidase	13	50	37	S26.001	Signal peptidase I (*Escherichia coli*)	141	141
19	M13.001	Neprilysin	19	67	38	S26.001	Signal peptidase I (*Salmonella typhimurium*)	54	54

purpose of constructing a benchmark training data set was to optimize the parameters of machine learning algorithms, train the prediction model and evaluate model performance in an *n*-fold cross-validation manner, whereas the purpose of constructing the independent test data set was to validate the generalization ability of trained prediction models and compare them with other existing tools. None of the substrate sequences in this constructed independent test data set appeared in the benchmark data set, which ensures that a fair assessment of model performance can be achieved.

Positive and negative samples

In this study, the number of negative samples (i.e. non-cleavage sites) in the data set of protease-specific substrate cleavage sites greatly dominates the number of the positive samples (i.e. cleavage sites). This leads to a class imbalance problem. If not addressed, this can result in models that favor negative predictions over positive [29, 51]. To address this data imbalance issue, we used a down-sampling strategy, randomly discarding from the overrepresented negative samples, to impose a ratio of 1 positive to every 3 negatives, as previously suggested [15, 25, 29, 51, 52].

To extract the sequence-based features of positive and negative samples, we used a local sliding window approach, with a fixed window size of P8–P8′ sites (i.e. eight residues in the upstream and another eight residues in the downstream to surround the cleavage site). The overall size of the sliding window was 16 sites. With regard to the selection of reliable negative samples, several previous studies have indicated that a few cleavage sites at the P1 position were predicted to be solvent inaccessible [14–16]. In view of these studies and for the purpose of extracting reliable non-cleavage sites, we randomly selected those negative samples with P1 sites predicted as solvent inaccessible by the SABLE program [53] for constructing the prediction models.

Sequence encoding schemes

We formulate cleavage site prediction as a classification problem and solve it using machine learning techniques. Each potential cleavage site (or non-cleavage site) of an amino acid sequence is represented by a feature vector x with D-dimensional feature components $\{x_1, \ldots, x_D\}$. The problem is to predict the label y of the site of interest that is represented and encoded by D-dimensional features. The y will be defined as '1' if the site is a cleavage site for a protease, and '0' otherwise.

The representation form of a potential site is determined based on the so-called 'sequence encoding scheme', which is used for extracting the potentially useful information from the amino acid sequence (often combined with predicted structural information) and converting the sequence data into numerical feature vectors [54, 55]. Accordingly, the sequence-encoding scheme plays a crucial role in determining the predictive performance of the machine learning-based model. In this study, we derived a great variety of features organized into 11 different types. We evaluated the relative predictive performance to identify effective combinations of features that could lead to the overall best predictive performance for a given protease. In addition to sequence-derived features, we also integrated evolutionary, physicochemical properties and predicted structural features. Detailed information on the software or databases we used to extract these different types of features is listed in Table 2, along with the feature category, annotations, dimension and references. Below we will first describe the different

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

Table 2. A complete list of sequence-derived structural, physicochemical and evolutionary features used

Number	Category	Feature type	Annotation	Dimension	Tool/database	Reference
1	Sequence-derived	BINARY	Binary sequence profile features	336	–	[37, 51]
2		CKSAAP	Composition of k-spaced amino acid pair	2400	–	[54, 56–59]
3		KNN	k-nearest neighbor features of local sequences	5	–	[58, 60]
4		AAC	Composition of 20 amino acid types	20	–	–
5	Evolutionary	PSSM	Position-specific scoring matrix	320	PSI-BLAST	[61]
6		BLOSUM	BLOSUM62 matrix	336	–	[62]
7	Physicochemical property	AAIndex	Numerical indices representing various physicochemical and biochemical properties of amino acids and pairs	1024	AAIndex	[63]
8		CHR	Charge/hydrophobicity ratio	9	AAIndex	[63]
9	Structural	SS	Predicted secondary structure	48	SABLE	[53]
10		SA	Predicted solvent accessibility	32	SABLE	[53]
11		DISO	Predicted natively disordered region	32	DISOPRED2	[64]
Total				4562		

Note: A local window size of 16 amino acid residues was used to extract the features. The last row shows the total number of features used.

types of sequence encoding schemes in detail, and then describe the SVM learning algorithm that is used to train the prediction models to predict y given x.

Sequence or sequence-derived features

With the avalanche of protein sequences generated in the post-genomic era, one of the most challenging problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine learning algorithms can only handle vectors but not sequences, as elucidated in [65]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid this for proteins, the pseudo amino acid composition (AAC) [66] or PseAAC [67] was proposed. Ever since the concept of PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics (see, e.g., [68–70] as well as a long list of references cited in [71]). According to the concept of general PseAAC [41], a protein sequence can be formulated as:

$$P = [\Psi_1 \ \Psi_2 \ \cdots \ \Psi_u \ \cdots \ \Psi_\Omega]^T, \tag{1}$$

where T is a transpose operator, while the subscript Ω is an integer and its value as well as the components Ψ_u $(u = 1, 2, \cdots, \Omega)$ depend on the way to extract the desired information from the amino acid sequence of P, as done in a series of recent publications (see, e.g., [72–78]).

Encouraged by the success of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (pseudo K-tuple nucleotide composition) [79, 80] was developed for generating various feature vectors for DNA/RNA sequences [81, 82] that have proved useful as well [83–90]. Particularly, recently a powerful Web server called 'Pse-in-One' [91] and its updated version 'Pse-in-One2.0' [92] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

Here, we used a variety of sequence-derived features to generate various different modes of general PseAAC that have proven useful in our previous studies. These include:

1. Binary sequence profile feature (termed as BINARY), which refers to the encoding of amino acid sequences using the 21-bit (20 amino acid types plus a 21-th gap-filling residue 'X') binary encoding method, as previously described [37, 51]. For a local sliding window of 16 amino acids to encode a potential cleavage site, the dimensions of this feature type are $21 \times 16 = 336$.
2. Composition of k-spaced amino acid pair (CKSAAP) [54, 56–59], which was originally termed as collocated amino acid pair encoding [56, 57, 93]. This encoding reflects the short-range interactions of residues within the sequence surrounding potential cleavage sites [59].
   Taking $k = 0$ as an example, there are 400 distinct types of 0-spaced amino acid pairs (i.e. AA, AC, AD, …, YY). Then, a feature vector can be defined as:

$$\left(\frac{N_{AA}}{N_{total}}, \frac{N_{AC}}{N_{total}}, \frac{N_{AD}}{N_{total}}, \ldots, \frac{N_{YY}}{N_{total}}\right)_{400}. \tag{2}$$

   The value of each descriptor denotes the composition of the corresponding amino acid pair in the protein or peptide sequence. For example, if the amino acid pair AA appears n times in the sequence, the composition of the amino acid pair AA is equal to n divided by the total number of 0-spaced amino acid pairs (N_{total}) in the local sliding window. We defined L to be the length of the local sliding window of cleavage site. In this case, $L = 16$, and the value of N_{total} is L—$(k+1)$. In this study, the CKSAAP encoding was performed over $k = 0$, 1, 2, 3, 4 and 5. Thus, the dimension of the CKSAAP feature vector is $400 \times 6 = 2400$.
3. K-nearest neighbor (KNN) features, which describe the cluster information of local sequences for predicting potential sites [58, 60]. This feature type ranks the top K peptides by computing the similarity scores between the query peptide and all peptides in both the positive and negative sets [58]. The similarity score between two peptide sequences P_1 and P_2 is defined as:

$$Score = \sum_{i=1}^{n} S(P_{1,i}, P_{2,i}), \tag{3}$$

$$S(a,b) = \begin{cases} BLOSUM62(a,b), & if(BLOSUM62) > 0 \\ 0, & if(BLOSUM62) \le 0 \end{cases}, \tag{4}$$

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

where P is the peptide with n amino acids, i is the amino acid position in the sequence and $BLOSUM62(a, b)$ is the corresponding element value for amino acids a and b in the BLOSUM62 matrix. Then, the ratio of positive samples in the top K peptides will be calculated. In this study, we set $K = 1$, 3, 5, 7 and 9% of the total numbers of positive and negative samples.

4. AAC, which is based on the calculation of the occurrence frequency of each of the 20 amino acid types in a local window. The frequencies of all 20 natural amino acids (i.e. 'ACDEFGHIKLMNPQRSTVWY') can be calculated as:

$$f(a) = \frac{N(a)}{L}, \quad a \in \{A, C, D, \ldots, Y\}, \tag{5}$$

where $N(a)$ is the number of occurrences of amino acid a, while L is the local window length. The dimension of the AAC feature vector is 20.

Physicochemical property features

These include: (i) charge/hydrophobicity ratio (termed as CHR) [59], which describes the charge and hydrophobicity ratio of the sequences surrounding cleavage sites; (ii) AAIndex features. AAindex [63] is a database of amino acid indices and amino acid mutation matrices. In the current version of the AAindex database (Version 9.2), 566 amino acid indices can be retrieved. Using the AAindex database, we extracted AAindex features that reflected the physicochemical properties of the sequences surrounding potential cleavage sites.

Evolutionary features

These include: (i) position-specific scoring matrix (PSSM) [61], which reflects the evolutionary information of the amino acids surrounding the cleavage sites; (ii) BLOSUM62. The BLOSUM62 matrix [62] is used to represent the sequence information surrounding a potential cleavage site, which reflects the similarity of two sequence fragments.

Structural features

In addition to the above features, we also incorporated structural information predicted from protein sequences, which include: (i) protein secondary structures predicted by SABLE [53]; (ii) solvent accessibility predicted by SABLE [53]; and (ii) natively disordered region predicted by DISOPRED2 [64].

Altogether, using a sliding window of 16 amino acids to encode and represent each potential cleavage site, we generated a 4562-dimensional feature vector based on the 11 types of features described above. Accordingly, each candidate cleavage site was represented by a feature vector x with 4562 feature components $\{x_1, \ldots, x_{4562}\}$.

Feature selection

To improve the feature representation ability and identify a subset of optimal features that contribute the most to the prediction of substrate cleavage sites, we used a two-step feature selection strategy, which combined mRMR [42] with forward feature selection (FFS) as described in our previous work [30, 44, 45, 94].

In this two-step feature selection strategy, the first step is to characterize the relative importance and contribution of each initial feature in the extracted feature set using the mRMR algorithm, which is able to rank all the initial features according to both their relevance to the response variables and the redundancy between the features themselves. Features that were assigned with higher ranking by mRMR were considered as having a better trade-off between their maximum relevance and minimum redundancy. After the first step, we selected the top 100 features as the optimal feature candidates (OFCs).

The second step is to apply the FFS method to sequentially select the most representative subset of optimal features from the 100 OFCs identified above. FFS adds a feature each time (usually starting with the feature that had the highest index assigned by mRMR, all the way to the feature that had the lowest index) and reconstruct the SVM model by performing the 5-fold cross-validation test. As a consequence, FFS resulted in a feature subset that led to the best predictive performance [measured by the area under the receiver operating characteristic (ROC) curve, AUC] of SVM models. The feature subset that resulted was then recognized as the optimal feature set. Finally, we obtained 38 protease-specific SVM models optimized by this two-feature selection strategy based on the benchmark training substrate data set for each protease.

Machine learning methods

SVM is an efficient machine learning algorithm suitable for solving binary classification, multiple classification or regression problems. The version of SVM best suited to predicting numerical outcomes is support vector regression (SVR). In this application, we used SVR to construct the prediction model to estimate the cleavage probability of substrate cleavage sites for a given protease. Owing to its excellent generalization capabilities, SVR has recently been applied in a growing number of applications in bioinformatics and computational biology, including cleavage site prediction [15, 29, 30], residue accessible surface area [95], protein B-factor [96, 97], half sphere exposure [98], disulfide connectivity [99], residue depth [54], torsion angles [29] and protein expression-level prediction [100]. It demonstrates competitive performance compared with other machine learning approaches, especially when dealing with real-valued prediction tasks.

The SVR classifier is able to find a linear discriminative function of the form:

$$f(\mathbf{x}) = \mathbf{W}^T \Phi(\mathbf{x}) + w_0, \tag{6}$$

where Φ is a basis function that maps the D-dimensional feature vector to a higher dimension. It is noteworthy that although $f(x)$ is a linear function of $\Phi(x)$, it can itself be a nonlinear function of $\mathbf{x}$, which reflects an attractive advantage of using kernel methods [101]. SVR assumes that the best discriminative function is the one that represents the largest separation or margin between the two classes of samples.

For implementation of the SVR algorithm, we used the LibSVM software package [102] with the regression mode. The model performance was fully evaluated by using 5-fold cross-validation and independent tests. The model parameters were optimized using the benchmark training data set, and the predictive performance of the SVR models for each protease was evaluated by performing 5-fold cross-validation using the benchmark data set and independent tests using the independent test data set. In particular, for each major sequence encoding scheme, we trained a corresponding SVR model. In addition, we have also concatenated all the initial features and generated an all feature-based model (referred to as ALL-Fea). We also performed feature selection experiments to identify a subset of optimal features for the cleavage site prediction of each protease,

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

and accordingly trained a selected feature-based model (denoted mRMR-FS).

Performance evaluation

To quantitatively evaluate the performance of a model, a set of four metrics is usually used in the literature. They include: (1) overall accuracy (Acc), (2) Mathew's correlation coefficient (MCC), (3) sensitivity (Sn) and (4) specificity (Sp), as given below (see, e.g., [103]):

$$\begin{cases} Sn = \dfrac{TP}{TP+FN} & (7) \\ Sp = \dfrac{TN}{TN+FP} & (8) \\ Acc = \dfrac{TP+TN}{TP+TN+FP+FN} & (9) \\ MCC = \dfrac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\ (TP+FN)(TN+FP)(TN+FN)}} & (10) \end{cases}$$

where TP, TN, FP and FN denote the numbers of true positives, true negatives, false positives and false negatives, respectively.

However, the above four metrics copied from math books lack intuitiveness and are not easy-to-understand for biologists, particularly the MCC, which is an important metric used for describing the stability of a predictor. Further, based on the Chou's symbols introduced in the study of protein signal peptides [104, 105], a set of four intuitive metrics was derived [106, 107], which are given below:

$$\begin{cases} Sn = 1-\dfrac{N_-^+}{N^+} & 0 \le Sn \le 1 & (11) \\ Sp = 1-\dfrac{N_+^-}{N^-} & 0 \le Sp \le 1 & (12) \\ Acc = \Lambda = 1-\dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le Acc \le 1 & (13) \\ MCC = \dfrac{1-\left(\dfrac{N_-^+}{N^+}+\dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1+\dfrac{N_+^- - N_-^+}{N^+}\right)\ \left(1+\dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le MCC \le 1 & (14) \end{cases}$$

where N^+ represents the total number of positive samples being investigated, while N_-^+ is the number of positive samples incorrectly predicted to be negatives; N^- denotes the total number of negative samples being investigated, while N_+^- denotes the number of the negative samples incorrectly predicted to be positives.

According to Equations (11)–(14), we can easily see the following: when $N_-^+ = 0$, $Sn = 1$; while when $N_-^+ = N^+$, we have $Sn = 0$. Likewise, when $N_+^- = 0$, $Sp = 1$; while when $N_+^- = N^-$, $Sp = 0$. When $N_-^+ = N_+^- = 0$, we have $Acc = MCC = 1$; while when $N_-^+ = N^+$ and $N_+^- = N^-$, $Acc = 1$ and $MCC = -1$; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, $MCC = 0$.

As we can see, based on the definition of Equations (11)–(14), the meanings of Sn, Sp, Acc and MCC have become much more intuitive and easier to understand, as concurred in a series of recent publications (see, e.g., [80, 84, 86, 88, 90, 106, 108–113]). It is instructive to point out that the performance metrics as defined in Equations (7)–(10) or Equations (11)–(14) are valid only for single label systems; whereas for multi-label systems (see, e.g., [114–117]), a set of more complicated metrics should be used as discussed in [118].

In addition, the value of AUC (under ROC curve) [119] was also used to quantitatively measure the quality of the predictors in this package via the 5-fold cross-validations and independent data set tests.

Results and discussion

Performance evaluation based on different sequence encoding schemes

In this section, we investigate the predictive performance of SVR models using different sequence encoding schemes and their combinations for cleavage site prediction of multiple proteases, by performing 5-fold cross-validation. The compared sequence encoding schemes include 'BINARY', 'PSSM', 'BLOSUM', 'KNN', 'CKSAAP' and 'AAIndex'. In addition, we also compared the performance of SVR models that were trained using all the initial features (referred to as 'ALL-Fea') and optimal selected features (termed 'mRMR-FS') after the two-step mRMR-FS feature selection. The ROC curves of these SVR models for cleavage site prediction of eight proteases [caspase-3, -6, -7, -8, MMP-2, -3, granzyme-B (human) and granzyme-B (mouse)] on the 5-fold cross-validation are shown in Figure 2.

Several important observations can be made. First, we can see that the 'ALL-Fea' model and 'mRMR-FS' model generally outperformed the other six models trained based on individual encoding schemes, with the AUC values ranging between 0.89 and 1.0. Second, the 'mRMR-FS' model achieved the overall best performance, after the two-step feature selection, compared with the other models for the MMP cleavage site prediction. For example, the 'mRMR-FS' model achieved an AUC of 0.968 for MMP-2 cleavage site prediction, while the second best 'ALL-Fea' model achieved an AUC of 0.892. Third, the accuracy of protease-specific cleavage site prediction varies substantially between different proteases and different protease families. The difficult cases include cleavage site prediction of the MMP family and other proteases (e.g. thrombin) whose activities are also regulated by confounding factors such as the presence of exosites (sites that are located outside the active sites) [120–122]. Compared with the caspases and granzyme B, the performance of cleavage site prediction with the MMP family achieved by a model using the same encoding scheme is much worse in terms of the AUC score. For example, the CKSAAP model only achieved an AUC of 0.502 and 0.581 for the cleavage site predictions of MMP-2 and MMP-3, respectively, compared with that of 0.914 and 0.922 for the cleavage site prediction of caspase-3 and caspase-7, respectively (Figure 2). Future studies should investigate incorporation of other relevant features that might prove useful for improving the predictive performance of cleavage sites for proteases with a requirement for allosteric regulation to cleave their target substrates.

Amino acid distributions in substrate cleavage site

To better understand informative features surrounding a cleavage site that may define protease-specific substrate cleavage, we examined the flanking sequences of protease-specific substrate cleavages with the pLogo program [123], a probabilistic approach to identifying the presence and visualization of sequence motifs. The generated sequence logo diagrams for caspase-3, -7, -6, -8, MMP-2, MMP-3, granzyme B (human) and granzyme B (mouse) are shown in Figure 3. To perform the

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

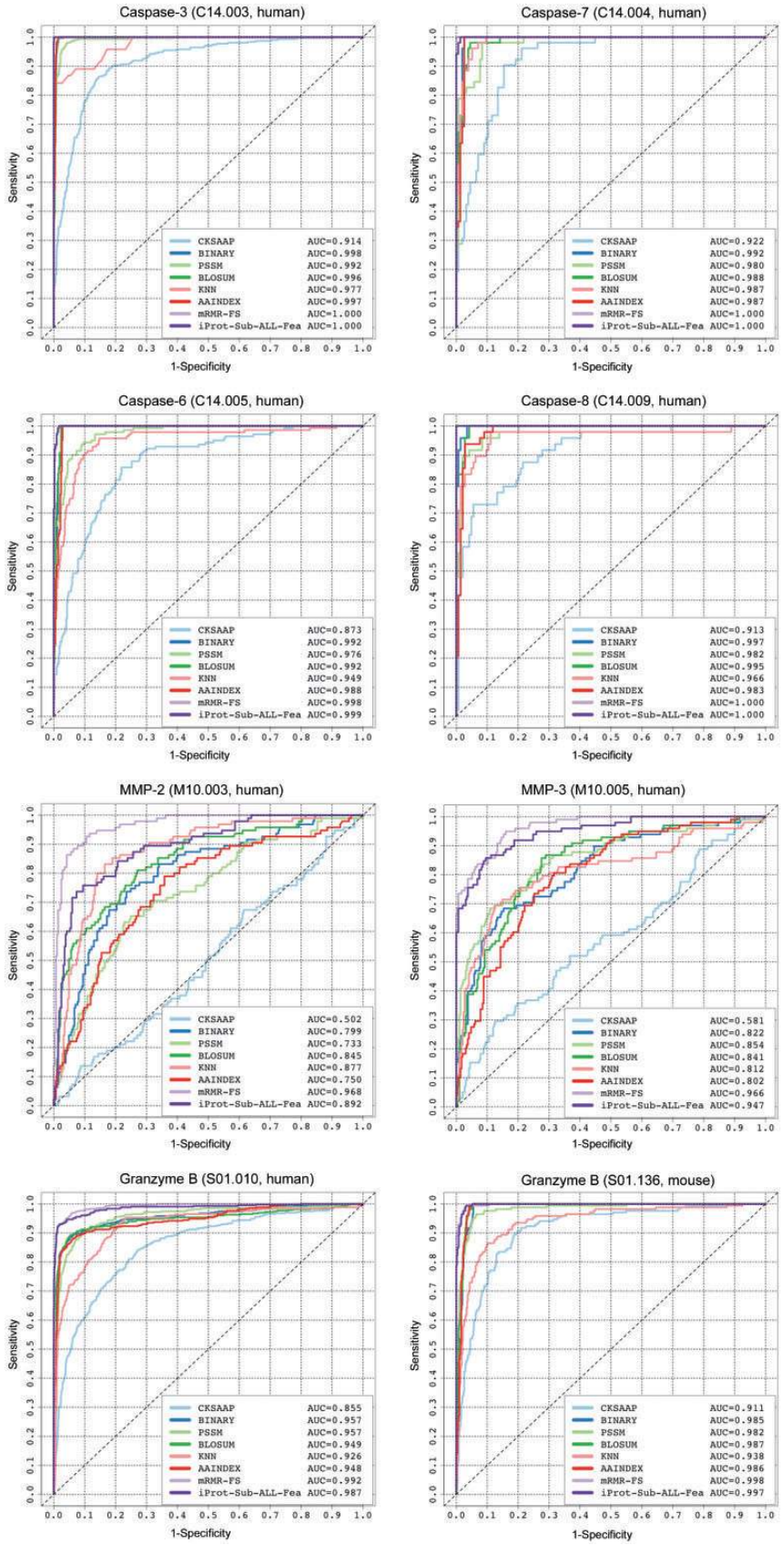

Figure 2. ROC curves of iProt-Sub models trained using different encoding schemes and their combinations for cleavage site prediction of eight proteases on the 5-fold cross-validation test.

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

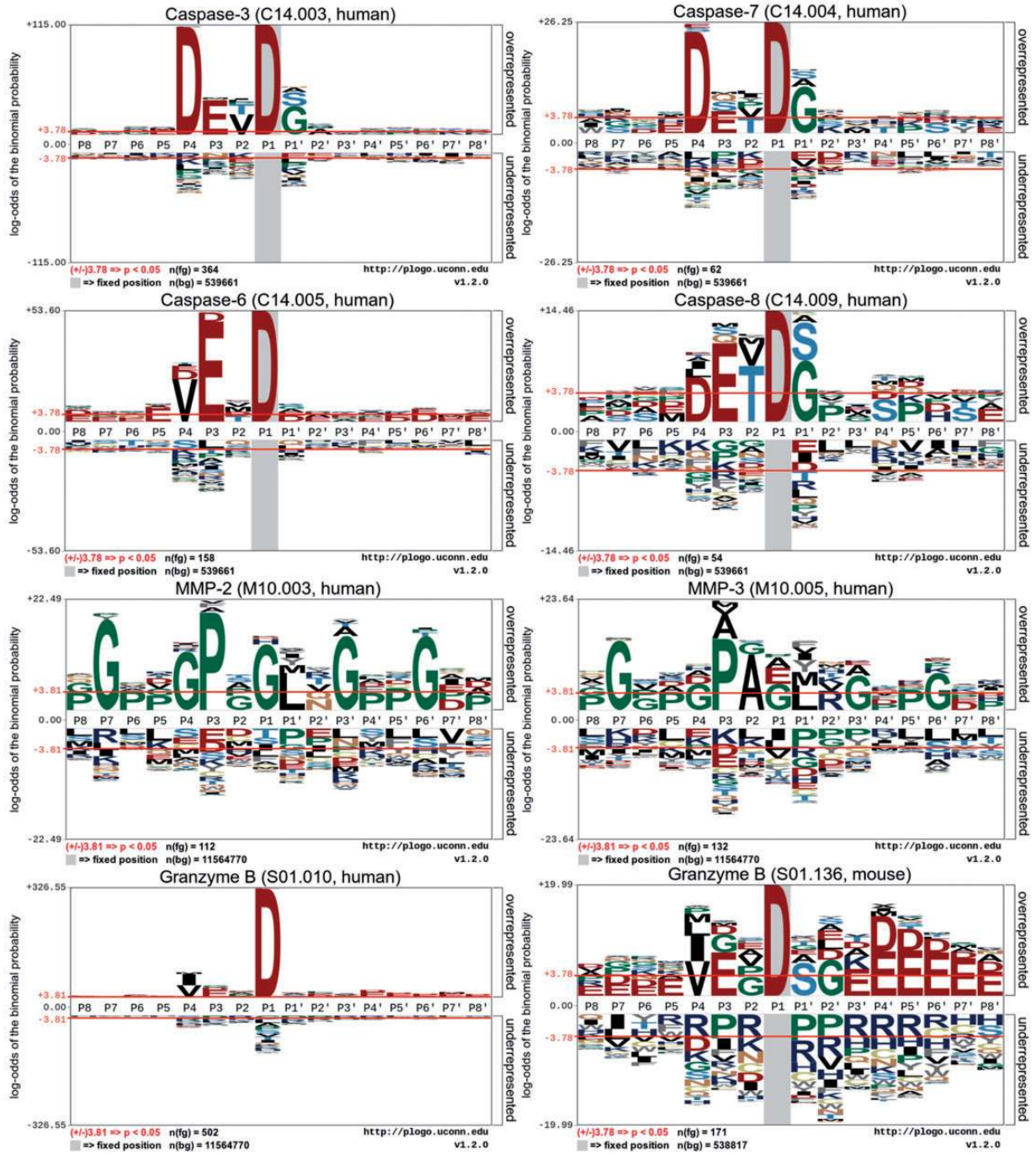

Figure 3. Sequence logo representations of experimentally verified cleavage sites (P8–P8′) of eight proteases caspase-3, -6, -7, -8, MMP-2, 3, granzyme-B (human) and granzyme-B (mouse). Sequence logos were generated using pLogo and scaled to the height of the largest column within the sequence visualization. The red horizontal lines on the pLogo graph denote the threshold of $P = 0.05$.

sequence logo analysis, we examined the P8–P8′ sites according to the Schechter–Berger nomenclature [124]).

Indeed, the sequence logos in Figure 3 show that there exist conserved sequence motifs or distinctive sequence patterns surrounding protease-specific substrate cleavage sites that may potentially be used to differentiate between different proteases. Notably, it can be seen that a predominant characteristic of substrate cleavage sites of caspases (caspase-3, -6, -7 and -8) is the requirement of Asp residue at the P1 position [125]. For certain caspases (e.g. caspase-3 and -7), there is also a lesser selectivity for Asp residue at the P4 position, thereby constituting the canonical DXXD motif [125]. A commonality of the cleavage selectivity of granzyme B, compared with that of caspases, is that they primarily recognize and cleave after the Asp at the P1 position as well. On a closer look, we can see that there exist subtle differences in the substrate cleavage selectivity between granzyme B (human) and granzyme B (mouse) [126]. Apparently, granzyme B (mouse) has a more complicated preference favoring a number of residue types across different positions surrounding the cleavage sites, including Val, Pro and Gly at the P1 position; Ser at the

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

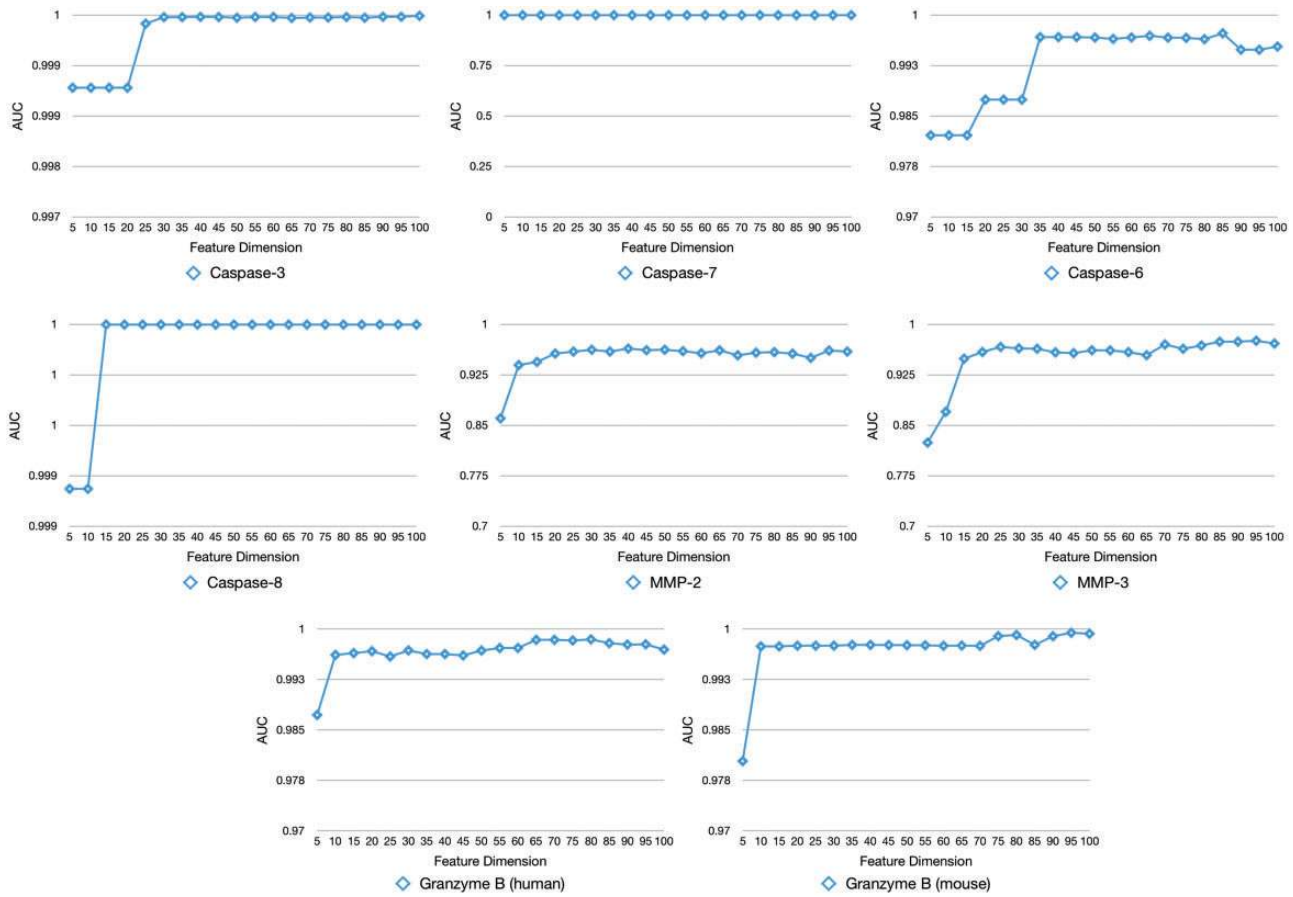

Figure 4. The feature selection curve in stepwise feature selection describes the performance change (in terms of AUC) as a function of the number of gradually increased OFCs.

P1′ position; and Gly at the P2′ position, respectively, while granzyme B (human) has much less selectivity at these positions.

However, different from caspases, matrix metallopeptidases (e.g. MMP-2 and MMP-3) have distinctive substrate specificities (Figure 3). Specifically, Gly was significantly overrepresented at the P7, P4, P2, P1, P3′ and P6′ positions surrounding the cleavage sites ($P < 0.05$, Figure 3). Owing to the intriguing selectivity of multiple residue types across different positions, it is much more difficult to clearly define distinctive sequence motifs for the MMPs. These results highlight the importance and need to improve the substrate cleavage site prediction by developing more accurate machine learning-based predictors, especially for proteases for which canonical sequence motif-based methods fail to perform well.

Feature contribution analysis

We used a two-step feature selection strategy by combining the mRMR algorithm [36] with FFS to characterize a subset of optimal features that contributed the most to the the prediction of substrate cleavage sites of each protease. Figure 4 shows the performance change (in terms of the AUC value) of the trained SVR models by gradually adding the selected features in a stepwise manner. As can be seen, all the feature selection curves started with quickly increasing the AUC value and then settled into the plateau after reaching their maximum at the peak, while in some cases, adding more features will lead to a drop in the AUC value (Figure 4).

Because 11 different types of features were originally extracted and used for training the models, it is of particular interest to characterize their relative importance and contribution to cleavage site prediction performance. In the 'ALL-Fea' sequence encoding scheme that encoded all the initial features, 11 different types of features were included. After the two-step feature selection based on mRMR and FFS, seven types of features remained in the final optimal feature subset of cleavage site prediction for eight proteases. To evaluate the contribution of these seven different feature types to the classification performance for individual proteases, the performance difference can be measured using the AUC value when a particular feature type is removed from the classifier. This measure thus represents the additional value of such feature type in cleavage site prediction, accounting for both interaction and compensatory effects between features [127]. Here, we define this measure as the 'contribution percentage' for a feature type by calculating the percentage of AUC decrease relative to other feature types after removing the feature from the classifier.

From Figure 5, we can see that the three most important types of features are KNN features with a contribution percentage ranging from 6.67% (for caspase-8) to 98.55% (for caspase-7), AAC features with a contribution percentage ranging from 1.45% (for caspase-7) to 54.54% (for caspase-6) and BINARY features with a contribution percentage ranging from 1.45% (for caspase-7) to 54.54% (for caspase-6). Among these three feature types, KNN features appear to be essential and thus most important for the predictive performance, as it was included in

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

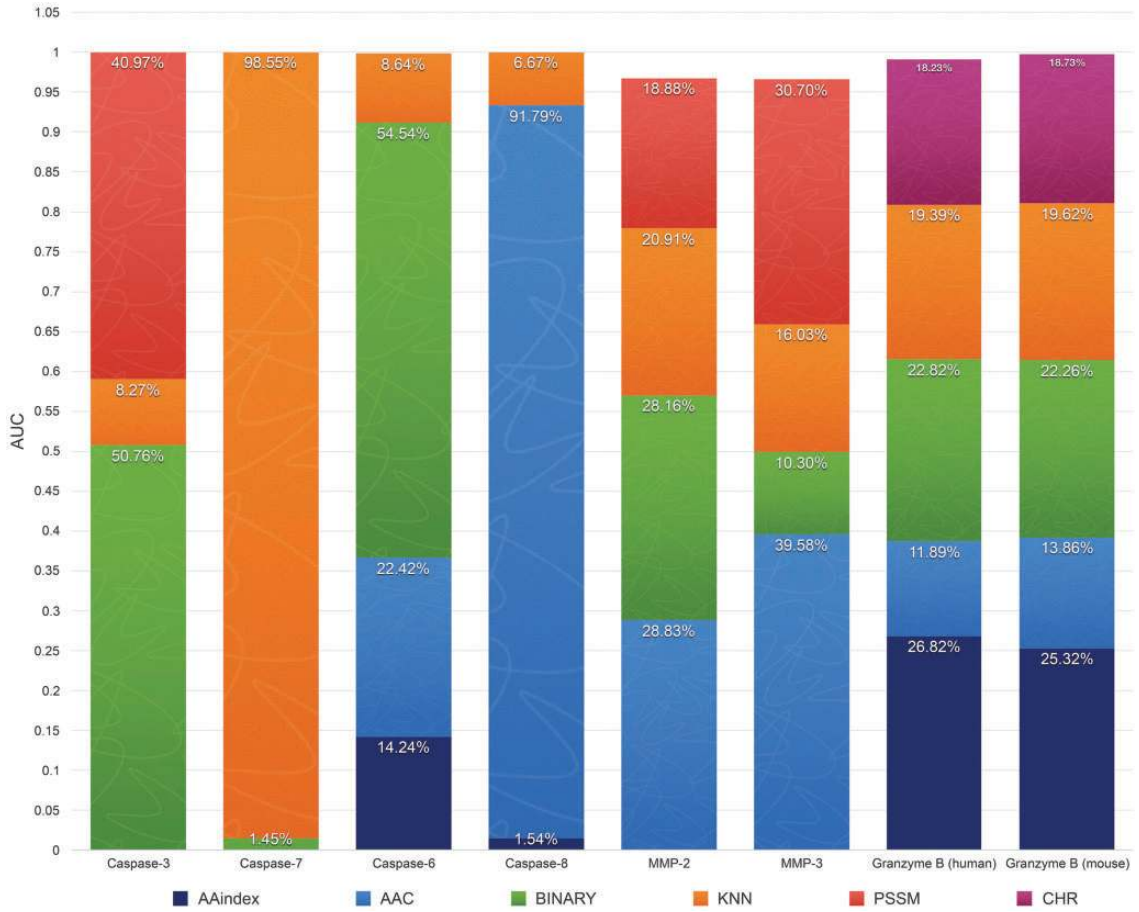

Figure 5. Importance of different feature types to the improvement of cleavage site prediction performance for eight proteases. The height of each bar for a feature type represents the proportional 'contribution percentage' that represents the AUC value of the feature selection model for one protease, and the uniformed AUC drop rate for each type of feature is represented with different colors. The AUC drop rate was obtained by comparing with the model after removing this feature from the input.

the feature sets of all proteases under investigation. In addition, the feature importance also varies depending on the protease, for example KNN features made an exclusive contribution to caspase-7 cleavage site prediction, but only made a moderate contribution to caspase-3 cleavage site prediction, secondary to BINARY and PSSM features. This is also the case for AAC features, which made a predominant contribution to caspase-8 cleavage site prediction (with a contribution percentage of 91.79%), but a marginal contribution to granzyme B (human and mouse) (with contribution percentages of 11.89 and 13.86%, respectively) cleavage site predictions, and completely dropped out of the final optimal feature subsets in the case of caspase-3 and -7 (Figure 5).

Performance comparison between iProt-Sub and other general prediction tools

In this section, we performed an independent test and compared the performance of iProt-Sub with three state-of-the-art general prediction tools that can be used to predict the substrate cleavage sites for multiple proteases: PoPS [33], SitePrediction [34] and Cascleave [15]. As a number of other tools were developed for specific proteases *per se*, they were not included in this comparative analysis. In addition, as the compared tools use different training data and algorithms to develop their respective prediction rules/models, the predictive capability of these tools differs from each other. Thus, to avoid any potential bias, for a protease, we only compared with tools that could provide valid prediction results after submitting the sequences of the independent test data set to each of the online Web servers. As a result, the ROC curves and calculated AUC values of cleavage site prediction for caspase-3, -6, -7, -8, MMP-2, MMP-3, granzyme B (human) and granzyme B (mouse) are shown in Figure 6.

It is of particular interest to compare the performance of iProt-Sub with Cascleave, which also uses the SVR algorithm and sequence-derived features (such as BINARY, predicted secondary structure and native disorder information) to train the prediction models. The online Web server of Cascleave provides three model options: Cascleave-BEAA, Cascleave-BEAA + BPBAA and Cascleave-BEAA + BPBAA + BPBDISO, which were trained using three different sequence encoding schemes [15]. PoPS is one of the most popular bioinformatics tools for modeling and predicting substrate specificity. It creates a simple matrix-based specificity model with different weights for amino acid residues at different positions, built from experimental data or expert knowledge and available to the user [33]. The specificity model can be used to score, predict and rank likely cleavage sites within a given substrate sequence for the designated protease of interest. SitePrediction is also a general prediction tool for predicting cleavage sites in candidate substrates. To make an accurate prediction,

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

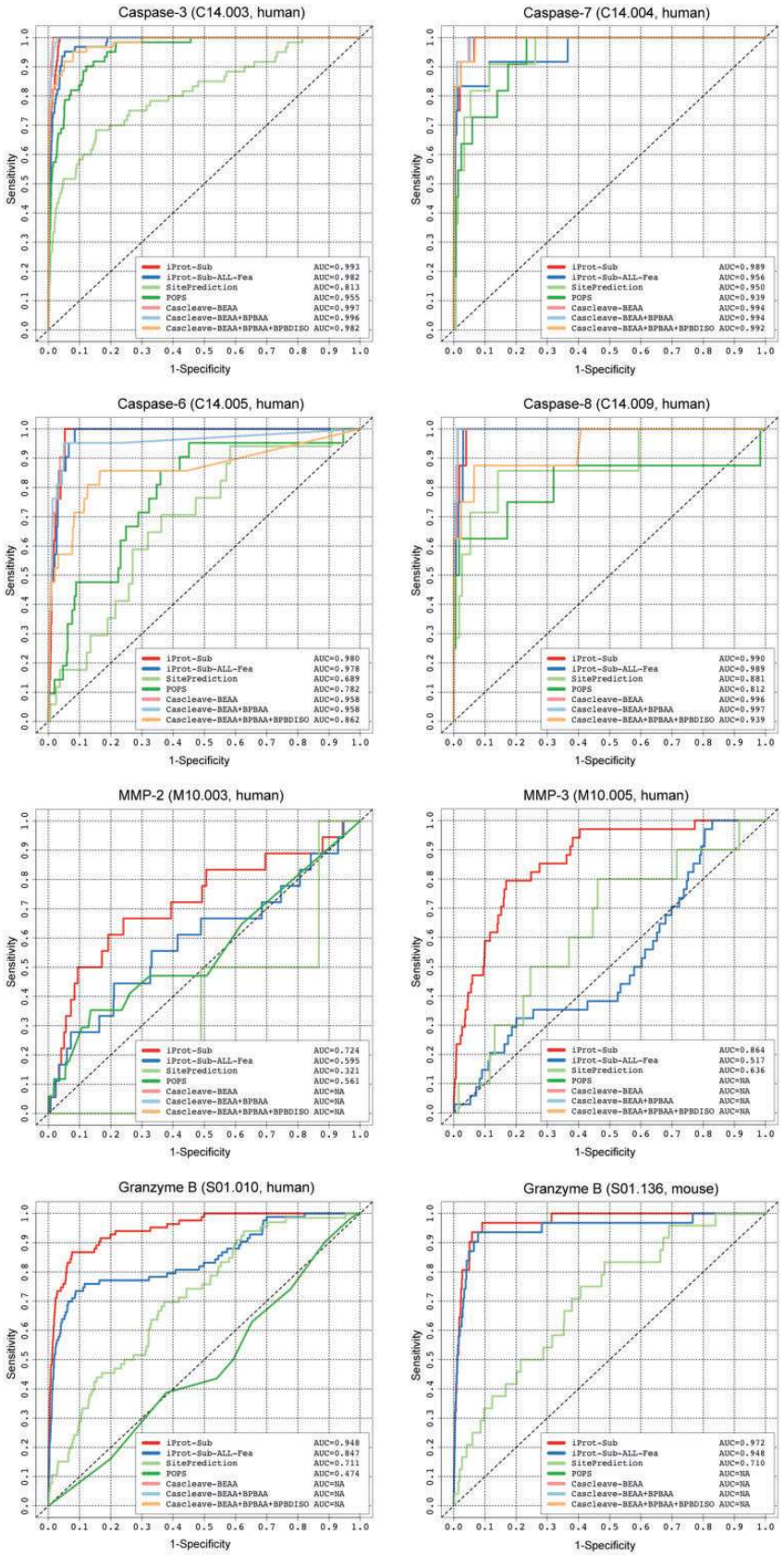

Figure 6. Performance comparison between iProt-Sub and other existing methods for cleavage site prediction for different proteases based on the independent test data sets, evaluated using ROC curves.

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

SitePrediction also considers other additional features that describe the environment of potential cleavage sites (including solvent accessibility, secondary structure and sequence similarity to the known cleavage sites) in conjunction with the amino acid frequency scores [34]. Both PoPS and SitePrediction are regarded as statistical scoring function-based tools, while Cascleave and iProt-Sub are considered as machine learning-based tools.

As can be seen from Figure 6, iProt-Sub achieved the overall best predictive performance compared with the other three tools PoPS, SitePrediction and Cascleave (with three models) for six proteases [including caspase-3, 6, MMP-2, MMP-3, granzyme B (human) and granzyme B (mouse)], with the only exception of caspase-7 and -8, for which iProt-Sub performed the second best, with an AUC of 0.989 and 0.990, respectively, in contrast to the best-performing tool Cascleave, which achieved an AUC of 0.994 and 0.997, respectively. For those proteases that iProt-Sub achieved the best performance, its performance gains over the other compared tools are apparent, which is particularly the case for MMP-2, MMP-3, granzyme B (human) and granzyme B (mouse) (Figure 6).

In addition, we note that although the strategy incorporating additional features generally improved the cleavage site prediction performance for some proteases [e.g. caspase-7, MMP-3, granzyme B (human),] in combination with feature selection, it decreased the performance for other proteases [e.g. caspase-3, -6, -8 and granzyme B (mouse)]. This can be observed by comparing the ROC curves and AUC values between the iProt-Sub models and iProt-Sub-ALL-Fea models in Figure 6. The underlying reason for this outcome is not entirely clear but might be associated with the size of the cleavage site data sets and the presence of other confounding factors that influence the cleavage outcome.

Overall, the results of the independent test indicate that by integrating heterogeneous informative features selected by an effective two-step feature selection strategy coupled with the SVR algorithm, iProt-Sub is able to provide a competitive predictive performance of substrate cleavage site prediction when compared with three existing prediction tools.

The implementation of iProt-Sub Web server

To facilitate high-throughput prediction and analyses of novel protease-specific substrates and cleavage sites, we have implemented an online Web server of iProt-Sub for the wider research community to use. The Web server was designed with a user-friendly interface and modern data visualization functionality and is freely available at http://iProt-Sub.erc.monash.edu/. It was implemented using Java Server Pages running Tomcat7 and configured in the Linux environment on a 16-core server machine with 50 GB memory and a 4 TB hard disk. To submit a prediction job, the server requires protein amino acid sequences (the submission of up to 50 sequences is permitted simultaneously) in FASTA format as the input. Users are also required to provide their e-mail addresses to receive a notification e-mail that contains a link to the prediction output Web page after the submitted job is completed. For a protein sequence with 500 amino acid residues, the prediction task will generally take approximately 3 min to calculate the features and return the final prediction results. A step-by-step guideline of how to use the iProt-Sub Web server can be found at http://iProt-Sub.erc.monash.edu.au/help.html.

Figure 7 provides an example output of the Web server. As can be seen, there are two main sections of the prediction output involving graphical visualization output (Figure 7A) and ranking output (Figure 7B) of the predicted cleavage sites in a protease family-specific manner. In terms of the graphical output, all the predicted cleavage sites are indicated by vertical lines with different colors (different colors indicate a different protease family, such as aspartic, cysteine, metallo, and serine). When hovering the mouse cursor over each differentially colored vertical line, a window pops up displaying detailed information associated with the predicted cleavage site/outcome, including the P4–P4′ sequence segment, cleavage site P1 position and the estimated sizes of N- and C-fragment cleavage products (Figure 7A). This graphical visualization function can greatly facilitate the quick identification of the predicted cleavage site(s) of interest by scanning from the N-terminus to C-terminus and visually comparing the cleavage profiling within the same substrate sequence across different proteases. The ranking output provides a tab-style view of the predicted cleavage sites according to the protease family (Figure 7B). Each tab contains the residue position of the predicted cleavage P1 site, the sequence ID, P4–P4′ sequence segment (with the predicted cleavage site indicated by '|'), the estimated N- and C-fragment sizes and the cleavage probability score.

Features used by the Web server for predicting cleavage sites include 11 previously mentioned feature encoding schemes, such as BINARY, AAC, PSSM, AAIndex, BLOSUM, CHR, CKSAAP, SS, SA, DISO and KNN. Based on the functionalities mentioned above, iProt-Sub offers important advantages over existing prediction servers in its ability to identify potential substrates and achieves a greater coverage and accuracy than previous predictors. To our knowledge, iProt-Sub is the most comprehensive server that is capable of predicting substrate cleavage sites of multiple proteases within a single substrate sequence using machine learning techniques. It is anticipated to be a valuable tool for cost-effective *in silico* identification of novel protease-specific substrates and cleavage sites.

Proteome-wide prediction and Gene Ontology enrichment analysis of protease-specific substrates at the proteome level

We applied the developed iProt-Sub tool to scan the human proteome (149 730 proteins) with a high stringency at the 100% specificity level in an effort to provide an overview of the substrate repertoire of several important proteases and gain insights into the significantly enriched Gene Ontology (GO) [128] terms and biological pathways of these 'computational' substrates at the entire human proteome level. Seven protease-specific models [caspase-3, -6, -7, -8, MMP-2, -3 and granzyme B (human)] were used to conduct the human proteome-wide scan, and >20 200 reviewed human protein sequences downloaded from the UniProt database were involved in this analysis. Note that to generate highly accurate mapping, we applied the prediction models that were trained using the final optimal features based on the complete training data sets to perform the proteome-wide substrate scanning. The statistics for the predicted substrate proteins and cleavage sites are shown in Table 3. A complete list of the predicted substrates for each protease and their corresponding substrate cleavage sites is available from the iProt-Sub website.

Based on the proteome-wide scanning results, we further conducted an enrichment analysis using the DAVID online server [129], including GO analysis and KEGG pathway analysis. The top five significantly overrepresented biological process (BP), cellular component (CC), and molecular function (MF)

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

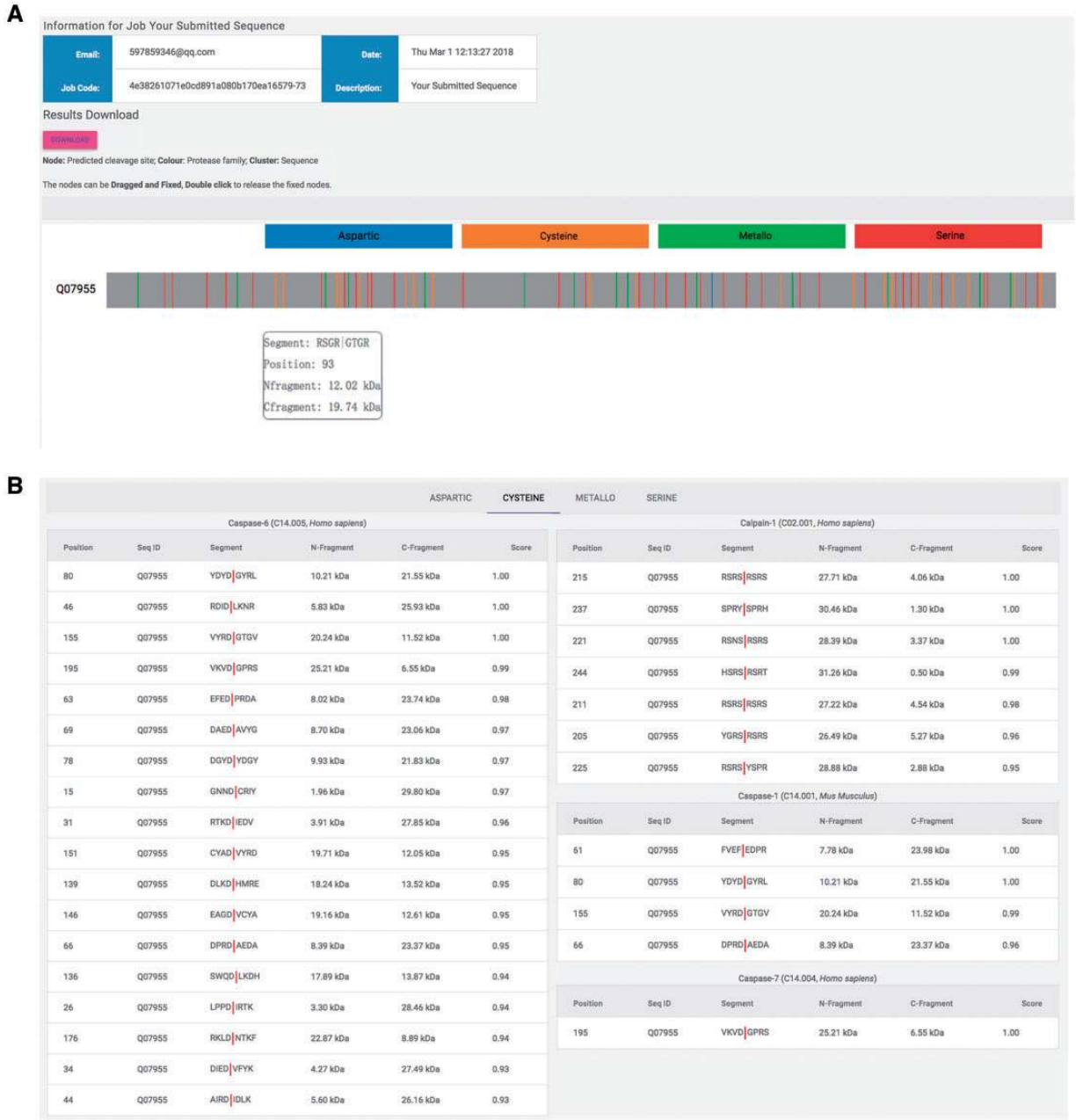

Figure 7. Example output of the iProt-Sub Web server.

terms, and KEGG pathways of the predicted substrate proteins for caspase-3, caspase-6, MMP-2 and MMP-3 at the proteome scale are highlighted in Figures 8 and 9, respectively. The sectorial area for a GO term represents the number of proteins with this term, while the different colors of the sectorial area indicate the statistical significance of the enrichment for the corresponding GO term. In general, substrate proteins targeted by different protease families tend to be associated with different GO terms, but substrate proteins within the same family share similar GO terms. For example, both caspase-3 and caspase-6 substrates were found to be enriched with the GO BP terms 'Cell adhesion' and 'Biological adhesion' and with the GO MF terms 'ATP binding' and 'adenyl ribonucleotide binding'. Similar tendencies can also be observed between the MMP-2 and MMP-3 substrates. With regard to the CC terms, many of the predicted substrates were found to be located in different components, including 'Nuclear lumen', 'Intracellular organelle lumen', 'Organelle lumen' and 'Cytoskeleton', where many apoptotic morphological changes and cellular signaling activities often occur [130].

In terms of pathway enrichment analysis, we observed that caspase and MMP substrates were highly enriched in several KEGG pathways that involve 'Focal adhesion', 'ECM-receptor interaction', different types of signaling pathways and 'Pathways in cancer' (Figure 9), highlighting the functional roles of these protease–substrate interactions in cancer-related biological processes. In addition, there also exist specific signaling pathways and cancer-related pathways that were specifically

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

Table 3. Statistical summary of the predicted substrates and cleavage sites with the 100% specificity at the proteome scale

MEROPS ID	Protease name	Number of predicted substrates	Number of predicted cleavage sites
C14.003	Caspase-3	10 645	26 929
C14.004	Caspase-7	12 288	34 355
C14.005	Caspase-6	5936	10 156
C14.009	Caspase-8	18 532	152 609
M10.003	MMP-2	9805	22 985
M01.005	MMP-3	402	425
S01.010	Granzyme B (human-type)	13 995	47 092

enriched for caspase-3 substrates (small cell lung cancer), caspase-6 substrates (hypertrophic cardiomyopathy), MMP-3 substrates (both small and non-small cell lung cancer) and MMP-6 substrates (endometrial cancer and chronic myeloid leukemia). These results highlight the functional roles of these protease–substrate interactions in cancer-related biological processes [1, 4, 5].

Case study

To illustrate the predictive power of iProt-Sub, we performed a case study where the targeted cleavage of the protein calpastatin by caspase-3 [131] and MMP-2 [132] was examined in detail. Calpastatin (UniProt ID: P20810) is an endogenous calpain (calcium-dependent cysteine protease) inhibitor, which is encoded by the CAST gene in humans. It consists of an N-terminal domain

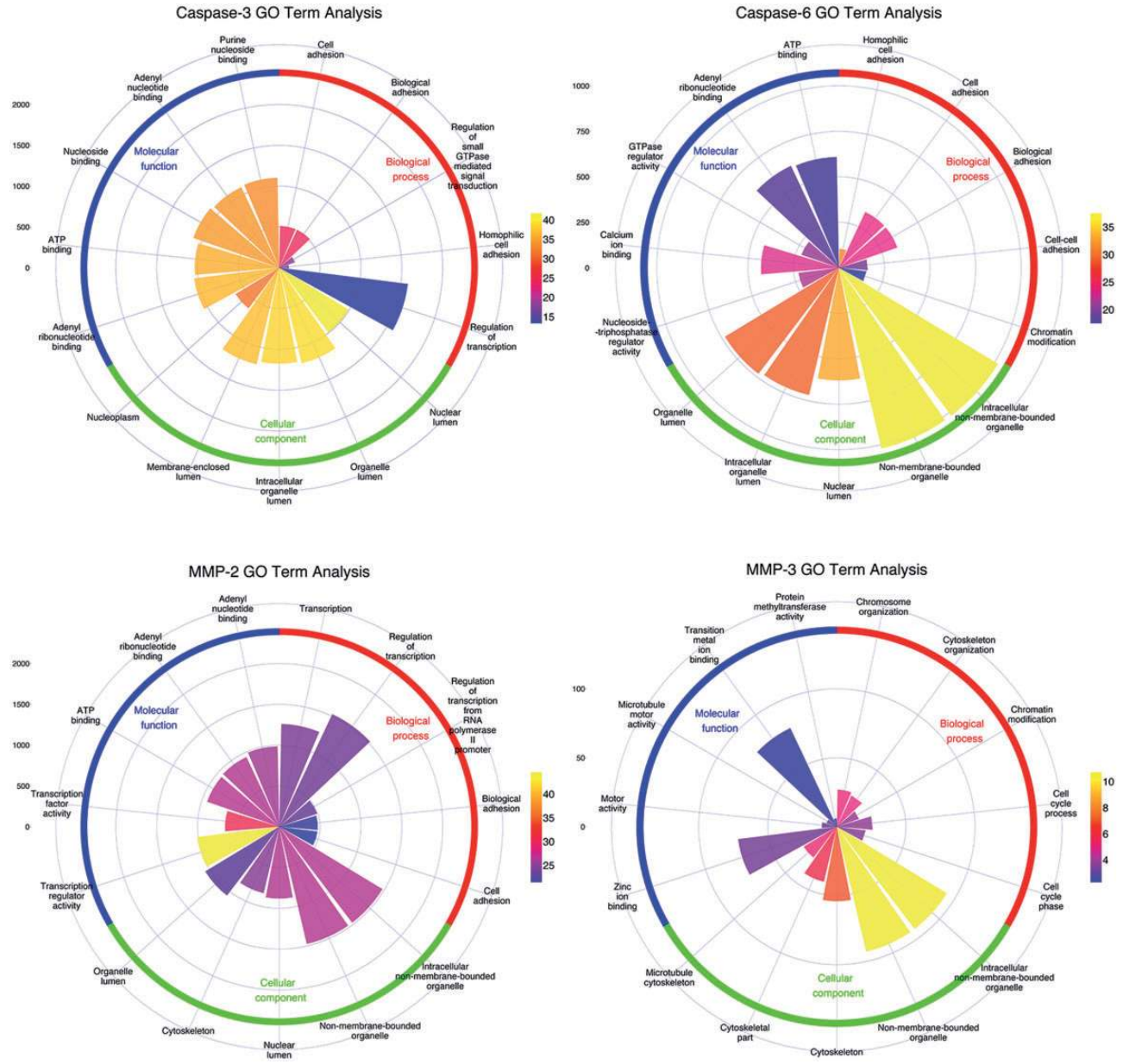

Figure 8. Functional enrichment analysis of the predicted substrates for caspase-3, -6, MMP-2 and -3 at the proteome scale, according to the BP, CC and MF classifications of GO terms. The statistical enrichment analyses of GO terms for predicted substrates were performed with the hypergeometric distribution.

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

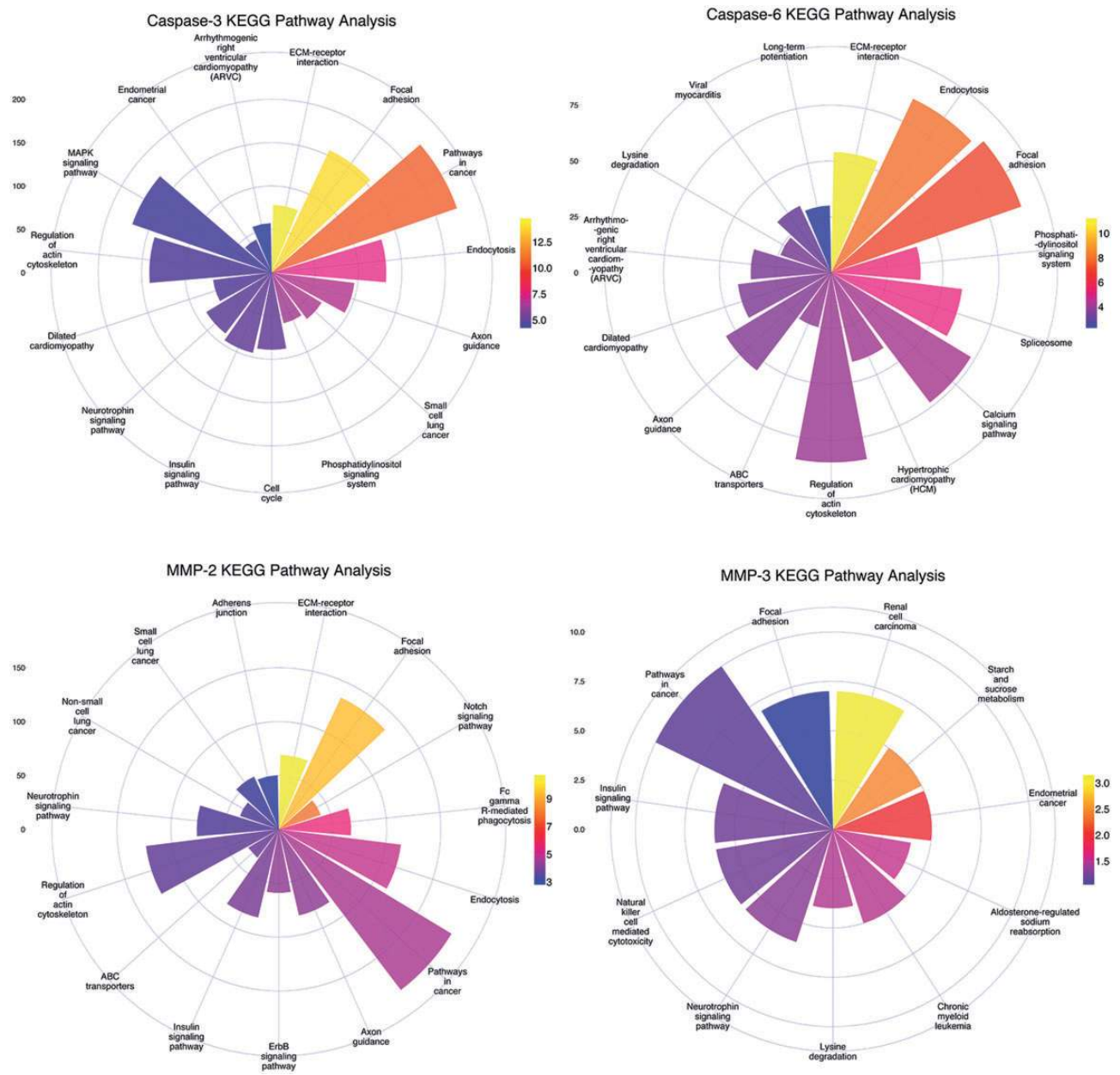

Figure 9. The KEGG pathway enrichment analysis of the predicted substrates for caspase-3, -6, MMP-2 and -3 at the human proteome scale. The statistical enrichment analyses of GO terms for predicted substrates were performed with the hypergeometric distribution.

and four repetitive calpain-inhibitory domains (Inhibitory Domains 1–4). It has been suggested that calpastatin is involved in the control of proteolysis of amyloid precursor protein and also in muscle protein degradation in living tissue [133].

Applying iProt-Sub to perform the substrate sequence scanning to calpastatin, we correctly identified all the three experimentally verified cleavage sites for caspase-3 [131]: ALDD|LIDT, DAID|ALSS and LSSD|FTGG (Figure 10). In terms of MMP-2 cleavage sites, we also identified an additional cleavage site for MMP-2: SVAG|ITAI [132]. Note that all these experimentally verified cleavage sites were on the top-ranking list of hits according to the predicted probability score generated by iProt-Sub, and all were above the threshold of 0.95. Moreover, iProt-Sub-based substrate sequence scanning also led to the prediction of several other high-confidence novel potential cleavage sites for both caspase-3 and MMP-2 (Figure 10). These predicted cleavage sites may represent novel sites targeted for cleavage under different conditions, and require follow-up experimental validation and hypothesis-driven studies. All the results above highlight the usefulness and value of using iProt-Sub as an *in silico* tool for identifying novel putative cleavage targets and unraveling the protease–substrate interaction relationship.

Limitations and future work

Despite the strong performance of our developed computational approach for predicting the substrate cleavage sites of multiple proteases, it has the following limitations:

The first limitation is that iProt-Sub is a machine learning-based approach and as such, its predictive power derives from the machine learning models that are trained based on different forms of sequence encoding schemes. The performance of

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

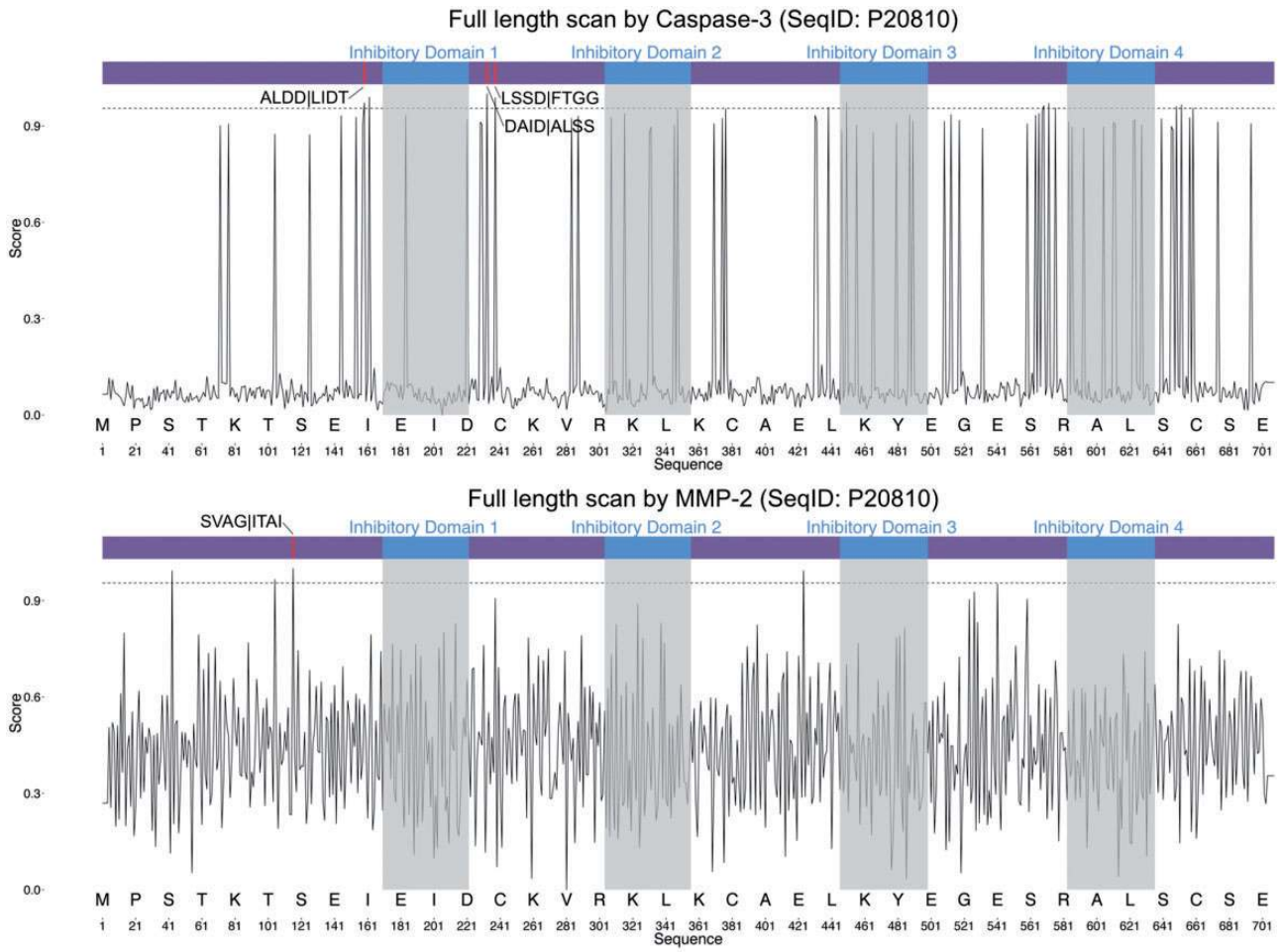

Figure 10. Full-length sequence scanning of calpastatin by iProt-Sub for caspase-3 (above) and MMP-2 cleavage sites (below). The horizontal axis denotes the amino acid residue position, while the vertical axis denotes the cleavage probability score generated by iProt-Sub. A higher threshold value of 0.95 is applied to identify the high-confidence cleavage site predictions, denoted by the dashed line. P4–P4′ sites surrounding the predicted cleavage P1 position are given.

machine learning-based models primarily depends on the effective representation of such feature encoding schemes. Accordingly, it remains a significant challenge in the future to identify further useful encoding schemes. There is much promise in this aspect from the availability of some recently developed powerful toolkits and Web servers for extracting a wide range of features, including Pse-Analysis [87], Bio-Seq Analysis [85], Pse-in-One [91], repDNA [81] and iFeature [134]. These toolkits could enable us to consider a much greater combination of different types of feature encoding schemes and explore the possibility of evolving iProt-Sub to a more robust framework while preserving or enhancing its model accuracy.

The second limitation is that the cleavage site prediction performance of iProt-Sub varies greatly between proteases. Its accuracy is lowest for matrix metalloproteases, including MMP-2 and MMP-3. The current study and several previous studies [29, 52, 135] confirmed the prediction of cleavage sites for these proteases to be an especially challenging problem and highlighted the need to develop specialized methods for improved MMP cleavage site prediction.

The third limitation is that most of the substrate cleavage sites used for training the prediction models of iProt-Sub were identified by high-throughput mass spectrometry methods, which might introduce a potential bias in terms of representing the global proteolytic events [9] and hence might impact on the predictive performance of the trained models [136]. Therefore, when sufficient heterogeneous cleavage site data sets identified by other different experimental approaches are available in the future, it will be important to characterize their potential influence on the predictive performance of cleavage sites.

The fourth limitation is that iProt-Sub only used the SVR algorithm to build the probabilistic cleavage site prediction models. In the future work, we plan to consider using other advanced machine learning techniques such as deep learning (DL), which can model high-level abstraction in the data [137], using significantly enlarged benchmark data sets (the next MEROPS release) to evaluate the performance of the DL models against other popular machine learning classifiers.

Conclusions

We have developed the iProt-Sub tool and constructed protease-specific prediction models for 38 proteases. iProt-Sub substantially upgrades the PROSPER Web server, includes a user-friendly interface and provides users with easy-to-understand data visualization techniques to better serve the wider research community. We have conducted a comprehensive set of experiments to benchmark the performance of different sequence encoding schemes and compare the models with other previously

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

proposed state-of-the-art methods. Our experimental evaluations indicate that the proposed iProt-Sub method outperforms those previously developed methods. iProt-Sub's improved performance can be attributed to several important aspects. First, we have compiled a comprehensive experimentally verified cleavage site data set in this work. Second, protease-specific models have been constructed, optimized and validated to achieve a better performance than other available tools using the powerful SVR algorithm coupled with a two-step feature selection procedure. Third, iProt-Sub allows high-throughput prediction of potential substrate cleavage sites for follow-up experimental validation and hypothesis-driven functional studies. A unique feature of iProt-Sub is that, unlike previously developed tools that require users to designate the protease of interest to make the prediction, iProt-Sub for a given substrate sequence will identify which, if any, of its 38 proteases will cleave that substrate. This unique feature makes iProt-Sub an attractive tool for proteomic research, especially in cases where there is insufficient knowledge about the protease(s) responsible for such cleavage to occur. We expect that iProt-Sub will be used as a valuable and powerful tool by the protease community and can deliver vital functional clues regarding the protease–substrate interactivity relationship in a cost-effective manner.

Key Points

- In this work, we present iProt-Sub, a powerful bioinformatics tool for the accurate prediction of protease-specific substrates and their cleavage sites.
- It provides optimized cleavage site prediction models with better predictive performance and coverage for four major protease families and 38 proteases.
- iProt-Sub integrates heterogeneous sequence and structural features derived from multiple levels in combination with an effective two-step feature selection procedure.
- Benchmarking experiments using cross-validation and independent tests showed that iProt-Sub was able to achieve a better performance than several existing generic tools. It is publicly accessible at http://iProt-Sub.erc.monash.edu.au/.
- Application of iProt-Sub to scan the entire human proteome provides an insightful overview of the substrate repertoire of several important proteases and significantly enriched GO terms and biological pathways of the 'computational' substrates at the proteome level.

Funding

This work was financially supported by grants from the Australian Research Council (ARC) (grant numbers LP110200333 and DP120104460), the National Health and Medical Research Council of Australia (NHMRC) (grant number 4909809), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (grant number R01 AI111965) and a Major Inter-Disciplinary Research (IDR) project awarded by Monash University.

References

1. López-Otín C, Overall CM. Protease degradomics: a new challenge for proteomics. *Nat Rev Mol Cell Biol* 2002;**3**(7):509–19.
2. Goldberg AL. Protein degradation and protection against misfolded or damaged proteins. *Nature* 2003;**426**(6968):895–9.
3. Sternlicht MD, Werb Z. How matrix metalloproteinases regulate cell behavior. *Annu Rev Cell Dev Biol* 2001;**17**(1):463–516.
4. Turk B, Turk D, Turk V. Protease signalling: the cutting edge. *EMBO J* 2012;**31**(7):1630–43.
5. Sevenich L, Joyce JA. Pericellular proteolysis in cancer. *Genes Dev* 2014;**28**(21):2331–47.
6. Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J Biol Chem* 1993;**268**:16938–48.
7. Chou KC. Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal Biochem* 1996;**233**(1):1–14.
8. Du QS, Sun H, Chou KC. Inhibitor design for SARS coronavirus main protease based on "distorted key theory". *Med Chem* 2007;**3**(1):1–6.
9. Fortelny N, Cox JH, Kappelhoff R, *et al*. Network analyses reveal pervasive functional regulation between proteases in the human protease web. *PLoS Biol* 2014;**12**(5):e1001869.
10. Song J, Tan H, Boyd SE, *et al*. Bioinformatic approaches for predicting substrates of proteases. *J Bioinform Comput Biol* 2011;**09**(1):9: 149–178.
11. Timmer JC, Zhu W, Pop C, *et al*. Structural and kinetic determinants of protease substrates. *Nat Struct Mol Biol* 2009;**16**(10):1101–8.
12. Agard NJ, Wells JA. Methods for the proteomic identification of protease substrates. *Curr Opin Chem Biol* 2009;**13**(5–6):503–9.
13. Kleifeld O, Doucet A, Prudova A, *et al*. Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat Protoc* 2011;**6**(10):1578–611.
14. Barkan DT, Hostetter DR, Mahrus S, *et al*. Prediction of protease substrates using sequence and structure features. *Bioinformatics* 2010;**26**(14):1714–22.
15. Song J, Tan H, Shen H, *et al*. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* 2010;**26**(6):752–60.
16. Kazanov MD, Igarashi Y, Eroshkin AM, *et al*. Structural determinants of limited proteolysis. *J Proteome Res* 2011;**10**(8):3642–51.
17. Shen HB, Chou KC. Identification of proteases and their types. *Anal Biochem* 2009;**385**(1):153–60.
18. Shen HB, Chou KC. HIVcleave: a web-server for predicting human immunodeficiency virus protease cleavage sites in proteins. *Anal Biochem* 2008;**375**(2):388–90.
19. Chou JJ. A Formulation for correlating properties of peptides and its application to predicting human-immunodeficiency-virus protease-cleavable sites in proteins. *Biopolymers* 1993;**33**(9):1405–14.
20. Chou KC, Zhang CT, Kezdy FJ. A vector projection approach to predicting HIV protease cleavage sites in proteins. *Proteins* 1993;**16**(2):195–204.
21. Zhang CT, Chou KC. An alternate-subsite-coupled model for predicting HIV protease cleavage sites in proteins. *Protein Eng* 1994;**7**(1):65–73.
22. Thompson TB, Chou KC, Zheng C. Neural network prediction of the HIV-1 protease cleavage sites. *J Theor Biol* 1995;**177**(4):369–79.
23. Chou KC, Tomasselli AG, Reardon IM, *et al*. Predicting human immunodeficiency virus protease cleavage sites in

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

proteins by a discriminant function method. *Proteins* 1996; **24**(1):51–72.
24. Chou KC, Shen HB. ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem Biophys Res Commun* 2008;**376**(2):321–5.
25. Wee LJ, Tan TW, Ranganathan S. SVM-based prediction of caspase substrate cleavage sites. *BMC Bioinformatics* 2006; **7**(Suppl 5):S14.
26. Wee LJ, Tong JC, Tan TW, *et al*. A multi-factor model for caspase degradome prediction. *BMC Genomics* 2009;**10**(Suppl 3): S6.
27. Ono Y, Sorimachi H, Mamitsuka H. Calpain cleavage prediction using multiple kernel learning. *PLoS One* 2011;**6**:e19035.
28. duVerle DA, Mamitsuka H. A review of statistical methods for prediction of proteolytic cleavage. *Brief Bioinform* 2012; **13**(3):337–49.
29. Song J, Tan H, Perry AJ, *et al*. PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One* 2012;**7**(11):e50300.
30. Wang M, Zhao X-M, Tan H, *et al*. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics* 2014;**30**(1):71–80.
31. Singh O, Su EC-Y. Prediction of HIV-1 protease cleavage site using a combination of sequence, structural, and physicochemical features. *BMC Bioinformatics* 2016;**17**(S17):478.
32. Gasteiger E, Hoogland C, Gattiker A, *et al*. Protein identification and analysis tools on the ExPASy server. In: *The Proteomics Protocols Handbook*. Springer, Totowa, NJ, USA, 2005, 571–607.
33. Boyd SE, Pike RN, Rudy GB, *et al*. PoPS: a computational tool for modeling and predicting protease specificity. *J Bioinform Comput Biol* 2005;**3**(3):551–85.
34. Verspurten J, Gevaert K, Declercq W, *et al*. SitePredicting the cleavage of proteinase substrates. *Trends Biochem Sci* 2009; **34**(7):319–23.
35. Wee LJ, Tan TW, Ranganathan S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics* 2007;**23**(23):3241–3.
36. Piippo M, Lietzén N, Nevalainen OS, *et al*. Pripper: prediction of caspase cleavage sites from whole proteomes. *BMC Bioinformatics* 2010;**11**:320.
37. Song J, Li F, Leier A, *et al*. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 2018;**34**:684–7.
38. Backes C, Kuentzer J, Lenhof H-P, *et al*. GraBCas: a bioinformatics tool for score-based prediction of Caspase-and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Res* 2005;**33**:W208–13.
39. Garay-Malpartida HM, Occhiucci JM, Alves J, *et al*. CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics* 2005;**21**(Suppl 1):i169–76.
40. Liu ZX, Cao J, Gao XJ, *et al*. GPS-CCD: a novel computational program for the prediction of calpain cleavage sites. *PLoS One* 2011;**6**(4):e19001.
41. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol* 2011;**273**(1): 236–47.
42. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 2005;**27**:1226–38.
43. Baldi P, Brunak S, Chauvin Y, *et al*. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;**16**(5):412–24.
44. Li F, Li C, Wang M, *et al*. GlycoMine: a machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome. *Bioinformatics* 2015;**31**(9): 1411–9.
45. Li F, Li C, Revote J, *et al*. GlycoMinestruct: 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**(1):34595.
46. Rawlings ND, Barrett AJ, Finn R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* 2015;**44**:D343–50.
47. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
48. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012; **28**(23):3150–2.
49. Song J, Tan H, Wang M, *et al*. TANGLE: two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences. *PLoS One* 2012;**7**(2):e30361.
50. Song J, Wang H, Wang J, *et al*. PhosphoPredict: a bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci Rep* 2017;**7**(1):6862.
51. Song J, Burrage K, Yuan Z, *et al*. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinformatics* 2006;**7**:124.
52. Wang Y, Song J, Marquez-Lago TT, *et al*. Knowledge-transfer learning for prediction of matrix metalloprotease substrate-cleavage sites. *Sci Rep* 2017;**7**(1):5755.
53. Wagner M, Adamczak R, Porollo A, *et al*. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol* 2005;**12**(3):355–69.
54. Shen HB, Song JN, Chou KC. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *J Biomed Sci Eng* 2009;**02**(03):136–43.
55. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr Proteomics* 2009;**6**(4):262–74.
56. Chen K, Kurgan LA, Ruan JS. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct Biol* 2007;**7**(1):25.
57. Chen K, Kurgan LA, Ruan JS. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J Comput Chem* 2008;**29**(10):1596–604.
58. Chen Z, Zhou Y, Song JN, *et al*. hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim Biophys Acta* 2013;**1834**(8):1461–7.
59. Chen Z, Zhou Y, Zhang Z, *et al*. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Brief Bioinform* 2015; **16**(4):640–57.
60. Gao J, Thelen JJ, Dunker AK, *et al*. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics* 2010;**9**(12):2586–600.
61. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;**292**(2): 195–202.
62. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;**89**(22): 10915–9.
63. Kawashima S, Pokarowski P, Pokarowska M, *et al*. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2007;**36**:D202–5.

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

64. Ward JJ, Sodhi JS, McGuffin LJ, *et al*. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004;**337**(3):635–45.
65. Chou K-C. Impacts of bioinformatics to medicinal chemistry. *Med Chem* 2015;**11**(3):218–34.
66. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition (vol 43, pg 246, 2001). *Proteins* 2001;**43**(3):246–60.
67. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;**21**(1):10–9.
68. Kumar R, Srivastava A, Kumari B, *et al*. Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *J Theor Biol* 2015;**365**: 96–103.
69. Ahmad K, Waris M, Hayat M. Prediction of protein submitochondrial locations by incorporating dipeptide composition into chou's general pseudo amino acid composition. *J Membrane Biol* 2016;**249**(3):293–304.
70. Meher PK, Sahu TK, Saini V, *et al*. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci Rep* 2017;**7**:42362.
71. Chou KC. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Cur Topics Med Chem* 2017;**17**:2337–58.
72. Xu Y, Ding J, Wu LY, *et al*. iSNO-PseAAC: predict Cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* 2013;**8**(2):e55844.
73. Xu Y, Wen X, Shao XJ, *et al*. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int J Mol Sci* 2014;**15**(5):7594–610.
74. Xu Y, Wen X, Wen LS, *et al*. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One* 2014;**9**(8):e105018.
75. Behbahani M, Mohabatkar H, Nosrati M. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. *J Theor Biol* 2016;**411**:1–5.
76. Khan M, Hayat M, Khan SA, *et al*. Unb-DPC: identify mycobacterial membrane protein types by incorporating unbiased dipeptide composition into Chou's general PseAAC. *J Theor Biol* 2017;**415**:13–9.
77. Xu Y, Wang Z, Li CH, *et al*. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med Chem* 2017;**13**:544–51.
78. Zhang SL, Duan X. Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. *J Theor Biol* 2018;**437**:239–50.
79. Chen W, Lei TY, Jin DC, *et al*. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal Biochem* 2014;**456**:53–60.
80. Chen W, Zhang XT, Brooker J, *et al*. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 2015;**31**(1): 119–20.
81. Liu B, Liu FL, Fang LY, *et al*. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics* 2015; **31**(8):1307–9.
82. Liu B, Liu FL, Fang LY, *et al*. repRNA: a web server for generating various feature vectors of RNA sequences. *Mol Genet Genomics* 2016;**291**(1):473–81.
83. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol Biosyst* 2015;**11**(10):2620–34.
84. Feng PM, Ding H, Yang H, *et al*. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol Ther Nucleic Acids* 2017;**7**:155–63.
85. Liu B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx165.
86. Liu B, Wang SY, Long R, *et al*. iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 2017;**33**(1):35–41.
87. Liu B, Wu H, Zhang DY, *et al*. Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget* 2017;**8**(8):13338–43.
88. Liu B, Yang F, Chou KC. 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol Ther Nucleic Acids* 2017;**7**:267–77.
89. Feng P, Yang H, Ding H, *et al*. iDNA6mA-PseKNC: identifying DNA N 6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 2018, doi: 10.1016/j.ygeno.2018.01.005.
90. Liu B, Yang F, Huang DS, *et al*. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 2018;**34**(1):33–40.
91. Liu B, Liu FL, Wang XL, *et al*. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res* 2015;**43**(W1): W65–71.
92. Liu B, Wu H, Chou K-C. Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat Sci* 2017;**9**:67.
93. Chen K, Jiang YF, Du L, *et al*. Prediction of integral membrane protein type by collocated hydrophobic amino acid pair. *J Comput Chem* 2009;**30**(1):163–72.
94. Wang M, Zhao X-M, Takemoto K, *et al*. FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS One* 2012;**7**(8):e43847.
95. Yuan Z, Huang BX. Prediction of protein accessible surface areas by support vector regression. *Proteins* 2004;**57**(3):558–64.
96. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. *Proteins* 2005;**58**(4):905–12.
97. Guruge I, Taherzadeh G, Zhan J, *et al*. B-factor profile prediction for RNA flexibility using support vector machines. *J Comput Chem* 2018;**39**(8):407–11.
98. Song J, Tan H, Takemoto K, *et al*. HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics* 2008;**24**(13):1489–97.
99. Song J, Yuan Z, Tan H, *et al*. Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics* 2007;**23**(23): 3147–54.
100. Chang CCH, Li C, Webb GI, *et al*. Periscope: quantitative prediction of soluble protein expression in the periplasm of *Escherichia coli*. *Sci Rep* 2016;**6**(1):21844.
101. Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. *Ann Stat* 2008;**36**(3):1171–220.
102. Chang CC, Lin CJ. LIBSVM. A library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**(3):1.

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022

103. Chen J, Liu H, Yang J, *et al*. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* 2007; **33**(3):423–8.
104. Chou KC. Using subsite coupling to predict signal peptides. *Protein Eng* 2001;**14**(2):75–9.
105. Chou KC. Prediction of signal peptides using scaled window. *Peptides* 2001;**22**(12):1973–9.
106. Chen W, Feng PM, Lin H, *et al*. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res* 2013;**41**(6):e68.
107. Xu Y, Shao XJ, Wu LY, *et al*. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ* 2013;**1**: e171.
108. Lin H, Deng EZ, Ding H, *et al*. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res* 2014;**42**(21):12961–72.
109. Jia J, Liu Z, Xiao X, *et al*. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J Theor Biol* 2016;**394**:223–30.
110. Zhang CJ, Tang H, Li WC, *et al*. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* 2016;**7**(43):69783–93.
111. Chen W, Ding H, Feng PM, *et al*. IACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 2016;**7**(13): 16895–909.
112. Chen W, Feng PM, Yang H, *et al*. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget* 2017;**8**(3):4208–17.
113. Ehsan A, Mahmood K, Khan YD, *et al*. A novel modeling in mathematical biology for classification of signal peptides. *Sci Rep* 2018;**8**(1):1039.
114. Cheng X, Xiao X, Chou KC. pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. *Mol Biosyst* 2017;**13**(9):1722–7.
115. Cheng X, Xiao X, Chou KC. pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. *Gene* 2017;**628**:315–21.
116. Xiao X, Cheng X, Su S, *et al*. pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. *Nat Sci* 2017;**9**:330.
117. Cheng X, Zhao SG, Lin WZ, *et al*. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics* 2017;**33**(22):3524–31.
118. Chou KC. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol Biosyst* 2013;**9**(6):1092–100.
119. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006;**27**(8):861–74.
120. Hauske P, Ottmann C, Meltzer M, *et al*. Allosteric regulation of proteases. *ChemBioChem* 2008;**9**(18):2920–8.
121. Rana S, Pozzi N, Pelc LA, *et al*. Redesigning allosteric activation in an enzyme. *Proc Natl Acad Sci USA* 2011;**108**(13):5221–5.
122. Song J, Tan H, Mahmood K, *et al*. Prodepth: predict residue depth by support vector regression approach from protein sequences only. *PLoS One* 2009;**4**(9):e7072.
123. O'shea JP, Chou MF, Quader SA, *et al*. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 2013; **10**:1211–12.
124. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* 1967;**27**(2):157–62.
125. Timmer J, Salvesen G. Caspase substrates. *Cell Death Differ* 2007;**14**(1):66–72.
126. Kaiserman D, Bird CH, Sun J, *et al*. The major human and mouse granzymes are structurally and functionally divergent. *J Cell Biol* 2006;**175**(4):619–30.
127. Lobley A, Swindells MB, Orengo CA, *et al*. Inferring function using patterns of native disorder in proteins. *PLoS Comput Biol* 2007;**3**(8):e162.
128. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25–9.
129. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocols* 2008;**4**:44–57.
130. 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**(1):14.
131. Pörn-Ares MI, Samali A, Orrenius S. Cleavage of the calpain inhibitor, calpastatin, during apoptosis. *Cell Death Differ* 1998;**5**(12):1028–33.
132. Schilling O, Overall CM. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat Biotechnol* 2008;**26**(6):685–94.
133. Orrenius S, Zhivotovsky B, Nicotera P. Regulation of cell death: the calcium-apoptosis link. *Nat Rev Mol Cell Biol* 2003; **4**(7):552–66.
134. Chen Z, Zhao P, Li F, *et al*. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 2018, doi:10.1093/bioinformatics/bty140.
135. Kumar S, Ratnikov BI, Kazanov MD, *et al*. CleavPredict: a platform for reasoning about matrix metalloproteinases proteolytic events (vol 10, e0127877, 2015). *PLoS One* 2015;**10**(5): e0127877.
136. Chen X, Qiu JD, Shi SP, *et al*. Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics* 2013; **29**(13):1614–22.
137. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**(7553):436–44.

Downloaded from https://academic.oup.com/bib/article/20/2/638/4979587 by guest on 21 August 2022