BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 1 2014, pages 71–80
doi:10.1093/bioinformatics/btt603

Structural bioinformatics

Advance Access publication October 21, 2013

Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets

Mingjun Wang[1,†], Xing-Ming Zhao[2,†], Hao Tan[3], Tatsuya Akutsu[4], James C. Whisstock[3,5,*] and Jiangning Song[1,3,5,*]

[1]National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, [2]Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, [3]Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria 3800, Australia, [4]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan and [5]ARC Centre of Excellence for Structural and Functional Microbial Genomics, Monash University, Melbourne, Victoria 3800, Australia

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Caspases and granzyme B (GrB) are important proteases involved in fundamental cellular processes and play essential roles in programmed cell death, necrosis and inflammation. Although a number of substrates for both types have been experimentally identified, the complete repertoire of caspases and granzyme B substrates remained to be fully characterized. Accordingly, systematic bioinformatics studies of known cleavage sites may provide important insights into their substrate specificity and facilitate the discovery of novel substrates.

Results: We develop a new bioinformatics tool, termed Cascleave 2.0, which builds on previous success of the Cascleave tool for predicting generic caspase cleavage sites. It can be efficiently used to predict potential caspase-specific cleavage sites for the human caspase-1, 3, 6, 7, 8 and GrB. In particular, we integrate heterogeneous sequence and protein functional information from various sources to improve the prediction accuracy of Cascleave 2.0. During classification, we use both maximum relevance minimum redundancy and forward feature selection techniques to quantify the relative contribution of each feature to prediction and thus remove redundant as well as irrelevant features. A systematic evaluation of Cascleave 2.0 using the benchmark data and comparison with other state-of-the-art tools using independent test data indicate that Cascleave 2.0 outperforms other tools on protease-specific cleavage site prediction of caspase-1, 3, 6, 7 and GrB. Cascleave 2.0 is anticipated to be used as a powerful tool for identifying novel substrates and cleavage sites of caspases and GrB and help understand the functional roles of these important proteases in human proteolytic cascades.

Availability and implementation: http://www.structbioinfor.org/cascleave2/.

Contact: Jiangning.Song@monash.edu or James.Whisstock@monash.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 2, 2013; revised on September 30, 2013; accepted on October 16, 2013

1 INTRODUCTION

Proteases are proteolytic enzymes that catalyze the breakdown of protein or peptide substrates by hydrolysis of peptide bonds. They represent ~2% (at least 500–600 proteases) of all gene products in human and are involved in the functional regulation of a large number of important physiological processes such as cell cycle, cell proliferation, programmed cell death, DNA replication, tissue remodeling and immune response (Rawlings *et al.*, 2008; Turk, 2006). Most proteases carry out either highly or moderately selective cleavage of the scissile bonds within the cleavage site of their substrates, with the substrate specificity ranging from preferences for a number of specific amino acids at defined positions to more generic sites with limited or no discrimination at one position (Chen *et al.*, 2008; Song *et al.*, 2011, 2012).

We are particularly interested in bioinformatic studies of the substrate specificity of human caspases and the proapoptotic protease granzyme B (GrB), a family of cysteine and serine proteases that are key regulators of apoptosis initiation and execution (Dix *et al.*, 2008). Both types are involved in a number of fundamental cellular processes and play a key role in apoptotic cell death. They also are involved in other cellular processes, including cell cycle, cell proliferation, development, cell migration and receptor internalization (Dix *et al.*, 2008; Los *et al.*, 2001). Caspases are known to have restricted substrate specificity for an aspartic acid residue (D) at position P1 of the target substrate. This primary specificity is similar with that of the serine protease GrB, which is delivered by natural killer cells into virally infected and tumor cells (Pardo *et al.*, 2009; Russell and Ley, 2002).

To date, at least 15 mammalian caspases have been identified (Chowdhury *et al.*, 2008) and they can be categorized into three groups based on their substrate specificities: group I caspases (caspase-1, 4, 5 and 13) prefer bulky hydrophobic amino acids at the P4 site and cleave the peptide sequence (W/L)EHD, group II caspases (capspase-2, 3 and 7) preferentially cleave the sequence motif DEXD, whereas group III (caspase-6, 8, 9 and 10) cleaves the motif (I/V/L)E(H/T)D. In contrast to the caspases, GrB prefers to cleave the sequence motif IEXD.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Identification of native substrates of caspases and GrB is the key to understanding of their physiological roles that have been implicated in the pathological processes including cancer, neurodegenerative and immunological diseases, contributing to our improved knowledge of proteolytic cascades lead to apoptotic cell death and which potential substrates can serve as potent therapeutic targets. Although the application of advanced large-scale high-throughput proteomic techniques has significantly increased the number of experimentally verified caspase and GrB substrates (Bredemeyer *et al*., 2005; Turk *et al*., 2001), the complete repertoire of the native substrates remains to be discovered, and furthermore, many other cleavage sites within the known substrates are not fully experimentally identified (Fischer *et al*., 2003). Moreover, experimental identification and characterization of protease substrates are often time-consuming, expensive and difficult (Enoksson and Salvesen, 2008; Enoksson *et al*., 2007; Schilling and Overall, 2008). Therefore, bioinformatic prediction of caspase and GrB substrates may provide valuable and experimentally testable information regarding novel potential cleavage sites or putative substrates.

A number of bioinformatic approaches have been developed to address the difficult task of predicting caspase and GrB cleavage sites [see (duVerle and Mamitsuka, 2012; Song *et al*., 2011) for a review]. For example, PeptideCutter used a limited experimental dataset to predict the substrate cleavage sites for a variety of proteases including several caspases (Backes *et al*., 2005; Gasteiger *et al*., 2005) developed GraBCas to provide position-specific scoring prediction of the potential cleavage sites of caspases 1–9 and GrB. Garay-Malpartida *et al*. built CasPredictor to predict caspase cleavage sites by using position-dependent amino acid matrices together with a PEST-like sequence index (Garay-Malpartida *et al*., 2005). Wee *et al*. developed CASVM based on support vector machine (SVM) to achieve a prediction accuracy ranging from 81.2 to 97.9% (Wee *et al*., 2007). PoPS is a systematic computational tool that supports the creation of specificity models of any protease given the knowledge of its substrate specificity data (Boyd *et al*., 2005). Verspurten *et al*. developed a scoring tool SitePrediction, which incorporated predicted secondary structure, solvent accessibility and PEST sequence occurrence (Verspurten *et al*., 2009). Song *et al*. proposed the first version of Cascleave predictor based on amino acid sequence, predicted secondary structure, solvent accessibility and natively disordered regions as the input features to the support vector regression (SVR) models (Song *et al*., 2010). Piippo *et al*. built Pripper using four different classifiers including SVM, J48, Random Forest and Vote (Piippo *et al*., 2010). CAT3 is a tool based on scoring matrices, however, it can predict substrate cleavage sites of caspase-3 only, which has limited its practical application (Ayyash *et al*., 2012). Barkan *et al*. developed an SVM-based approach, termed PCSS, which used characteristic sequence and structure features important for substrate recognition to predict novel substrates of caspases and GrB (Barkan *et al*., 2010). These tools can be generally divided into two categories: scoring function- or machine learning-based. PeptideCutter, PoPS and SitePrediction belong to the first category, whereas Cascleave 1.0, PCSS, CASVM and Pripper belong to the second category. In terms of machine learning algorithms used for model training, Cascleave 1.0, PCSS, CASVM and Pripper are developed using SVM, whereas Pripper combines random forest, J48 and vote predictions.

In this study, we present a novel approach termed Cascleave 2.0 to predict caspase- and GrB-specific cleavage sites by integrating heterogeneous information from various sources. We first constructed non-redundant substrate datasets from MEROPS (Rawlings *et al*., 2008) and then selected useful sequence and protein functional features as the input our SVR models. To this end, we performed extensive feature selection by using a two-stage feature selection procedure to comprehensively investigate and characterize features arising from different levels that are important for determining protease-specific substrate. Independent tests indicate that Cascleave 2.0 outperforms other existing methods and can be used as a powerful tool for high-throughput *in silico* screening of novel putative substrates of human caspase-1, 3, 6, 7, 8 and GrB and their respective cleavage sites.

2 METHODS

2.1 Benchmark and independent test datasets

We curated substrate datasets for human caspase-1, 3, 6, 7, 8 and GrB from the MEROPS database (version 9.6) (Rawlings *et al*., 2008). All the annotated substrate cleavage sites in the datasets were experimentally determined. The number of experimentally verified substrates and cleavage sites for different enzymes, after removing redundancies, is shown in Supplementary Table S1. Moreover, a background dataset containing all human proteins was retrieved from UniProt (version released in August 2012) (Bairoch *et al*., 2005). Given a protease, its positive (cleavage site peptide) and negative (non-cleavage site peptide) samples were generated and used as training/testing data. In particular, it would be difficult to prove definitely that a particular peptide bond is not cleaved under any conditions. Accordingly, it is hard to collect a set of protein sequences that can be safely regarded as non-cleavable. In this work, the negative samples were selected from the background protein sets that excluded the experimentally verified substrates. A sliding window approach commonly used to extract sequence features was adopted here for both positive and negative samples. For positive samples, the flanking sequences around each cleavage site were considered (Song *et al*., 2010, 2012; Zhao *et al*., 2005). For negative samples, the same number of sites not known as cleavage sites was randomly selected from the background proteins (Zhao *et al*., 2008). The schematic flowchart of Cascleave 2.0 can be found in Figure 1.

Sequence identity for the initial datasets was reduced using CD-HIT (Huang *et al*., 2010) in such a way that the sequence identity between any two protein sequences should be no larger than 30%. This step is essential for eliminating sequence redundancy and avoiding overestimation of the performance of machine learning predictors. To objectively evaluate the performance, the final substrate datasets after removing redundancies were randomly divided into two subsets: a benchmark dataset and an independent dataset (~20% of the size of the benchmark dataset). The performance of the models was comprehensively evaluated using 5-fold cross-validation on the benchmark dataset and further validated on the independent test set. In 5-fold cross-validation, the substrate sequences were randomly divided into five subsets of roughly equal size, four of which were used as the training set, whereas the rest one was left as the test set. This procedure was repeated five times such that each subset was used as the test set once. Using the independent dataset, we performed the independent test by training the model based on the benchmark dataset so that the performance of the model can be objectively evaluated.

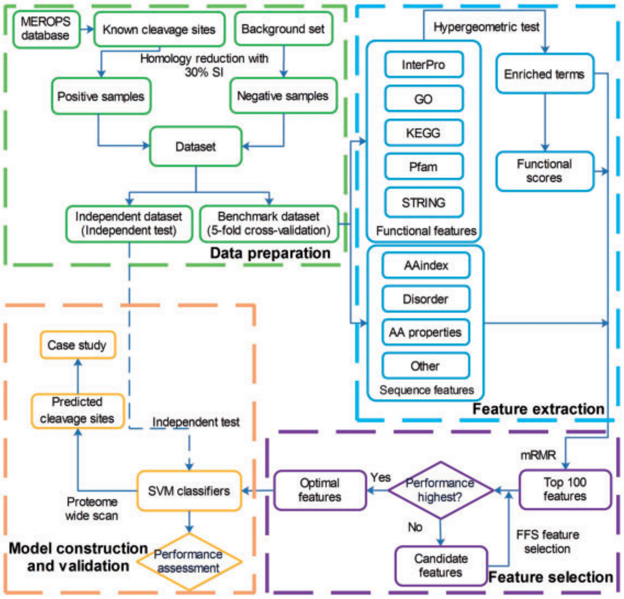

Fig. 1. Flowchart of the Cascleave 2.0 methodology

2.2 Feature extraction

To improve prediction accuracy, we considered both sequence and functional features to identify potential cleavage sites for each protease. These features include (i) secondary structures predicted by SABLE (Wagner *et al.*, 2005); (ii) solvent accessibility predicted by SABLE (Wagner *et al.*, 2005); (iii) disordered region predicted by DISOPRED2 (Ward *et al.*, 2004); (iv) amino acid index (AAindex) representing various physicochemical and biochemical properties of amino acids (Kawashima *et al.*, 1999); (v) position-specific scoring matrix generated by PSI-BLAST (Altschul *et al.*, 1997); (vi) isoelectric point (pI) calculated by BioPerl (Stajich *et al.*, 2002); (vi) residue conservation score; (vii) amino acid mass weight; (ix) biological process (BP) feature from Gene Ontology (Ashburner *et al.*, 2000); (x) cellular component (CC) feature from Gene Ontology; (xi) molecular function (MF) feature from Gene Ontology; (xii) functional domain information from InterPro (Hunter *et al.*, 2012); (xiii) pathway information from KEGG (Kanehisa and Goto, 2000); (xiv) functional domains from Pfam (Punta *et al.*, 2012); and (xv) protein–protein interaction (PPI) from STRING (Jensen *et al.*, 2009). Details of how to extract the features are provided in the Supplementary File (Supplementary Table S2).

2.3 Over- and underrepresented feature analysis

For each protein substrate, the set of various heterogeneous features generated above are highly dimensional, heterogeneous, noisy and redundant. This will lead to a time-consuming practice to train classifiers using these noisy features, thereby resulting in possible biased model training and prediction (Zhao *et al.*, 2010). To address this, a two-sided hypergeometric test was used to identify over- or underrepresented terms for the protein substrates of each protease as opposed to the background protein set. Given a certain feature, the hypergeometric test was performed using the R Stats package (Team, 2011):

$$p = F_{hypergeom}(q, m, n, k) \qquad (1)$$

where q is the number of samples annotated with the feature term in the 'set' of interest (here, we refer to the benchmark dataset excluding the independent test dataset), m is the number of samples annotated with the feature in the background set, n is the number of samples without the feature, whereas k is the number of samples in the set of interest. P-values derived from hypergeometric test were corrected by Bonferroni correction for testing on multiple terms. Features with $P < 0.01$ were taken as significant. We then scored potential substrate proteins based on the overrepresented features, as described in previous work (Li *et al.*, 2010).

2.4 SVR implementation

We used SVR to build the models to estimate the cleavage probability of substrates for each protease (Song *et al.*, 2007). SVR is the regression mode of SVM (Cortes and Vapnik, 1995), which is a supervised machine learning approach (Burges, 1998). Therefore, SVR allows us to build models that provide quantitative evaluation of the cleavage probability. We used the LIBSVM package (Chang and Lin, 2011) for SVR implementation, where the radial basis kernel function with default parameters was selected. The radial basis kernel function function is defined as follows:

$$y(x) = \sum_{i=1}^{N} w_i \phi(x - x_i) \qquad (2)$$

where the approximating function $y(x)$ is represented as a sum of N radial basis functions [$\phi(x - x_i)$ means the distance from x to x_i], each associated with a center x_i and weighted by an appropriate coefficient w_i.

All the input feature values were normalized as follows:

$$X_{norm} = \frac{1}{1 + e^{-x}} \qquad (3)$$

where x is the real value of each feature, and X_{norm} is the value after normalization.

2.5 Two-step feature selection

Because it is likely that many of the derived features contain redundant information, we performed feature selection to filter those features that do not contribute to the predictive performance. In this work, a two-step feature selection approach, composed of the maximum relevance minimum redundancy (mRMR) (Peng *et al.*, 2005) and forward feature selection (FFS) (Wang *et al.*, 2012) was used to select the most informative features for predicting substrate cleavage sites.

The first step is to evaluate the relative importance of each feature using mRMR that is able to rank the features according to both their relevance to the response variables and the redundancy between the features themselves. Features assigned with higher rankings are thought to have better trade-off between the maximum relevance and minimum redundancy. We selected the top 100 features as our optimal feature candidates (OFCs).

The second step involves use of the FFS method to select a condensed subset of optimal features from the above 100 OFCs. FFS is a feature selection method that examines each feature from the candidate feature set sequentially and results in a feature set that achieves the highest accuracy (using AUC as described later in the text). Briefly, FFS starts with the feature that is ranked as the best feature with mRMR, and then more features are added to the preselected feature set that can lead to higher prediction accuracy. This procedure continues until no more features can be added to the selected feature set to improve prediction accuracy. The resulted feature set will be used as the final feature set for further analysis.

2.6 Performance evaluation

The prediction performance was evaluated using the following measures:

(1) Sensitivity (SEN) or true-positive rate (percentage of correctly predicted cleavage sites):

$$Sensitivity = TP/(TP + FN) \qquad (4)$$

(2) Specificity (SPE, percentage of correctly predicted non-cleavage sites):

$$Specificity = TN/(TN + FP) \tag{5}$$

(3) False-positive rate (percentage of not correctly predicted non-cleavage sites):

$$FPR = 1 - Specificity = FP/(TN + FP) \tag{6}$$

(4) Precision (PRE) is defined as:

$$Precision = TP/(TP + FP) \tag{7}$$

(5) Accuracy (ACC, percentage of correct predictions of both cleavage and non-cleavage sites):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

(6) Mathew's correlation coefficient (MCC), a measure of the quality of binary classifications (Matthews, 1975), is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

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

In addition, we also used the AUC measure, the area under the receiver operating characteristic curve (ROC), to evaluate the performance of classifiers. Moreover, the performance was comprehensively evaluated using the six measures based on 5-fold cross-validation and independent tests.

3 RESULTS AND DISCUSSION

3.1 Analysis of sequence-level determinants of the protease substrate specificity

Supplementary Table S1 lists the number of cleavage sites and sequences clustered at the two sequence identity levels. Because we are more interested in analyzing the substrate specificity of human proteases, only human substrates were retained in our final datasets. We then calculated the frequencies of amino acid residue types appearing at each position of P16-P16′ (upstream and downstream 16 sites) sites surrounding the cleavage sites.

Using the human proteome as the reference set, sequence logo representations (Supplementary Fig. S1) were generated by using IceLogo (Colaert *et al.*, 2009), allowing us to identify conserved sequence motifs or distinct patterns that differentiate between different proteases, as shown in Supplementary Figure S1. It can be seen that a primary feature of the substrate specificity of caspases is the requirement of Asp at P1 position, whereas a lesser selectivity of Asp residue was also observed on the P4 position, constituting the canonical DXXD motif (Nicholson, 1999). GrB is known to have similar substrate specificity as the caspases, e.g. it primarily cleaves after Asp at P1 position. Nonetheless, we noted that there exist subtle, yet important differences between the substrate specificities of different proteases. For example, caspase-3 and 7 strongly disfavor Leu at P4 position, whereas other proteases do not. On the other hand, caspase-6 and GrB prefer to have 'V' at P4 position, which is not the case for other proteases. In addition, more subtle differences between different proteases were also observed at the prime-side positions, especially P1′ and P2′ positions. For example, caspase-1, 3, 7 and 8 prefer to have 'G/S' at P1′ position, whereas caspase-6 and GrB favor 'D/E' and 'S/T', respectively. Moreover, caspase-1 and 8 prefer to have 'P' at P2′ position, whereas others do not have such preference. All together, these results highlight the need and importance to address this problem by developing protease-specific predictors that might allow more precise recognition of its substrate cleavage sites.

3.2 Characterization of functional features

In this section, we investigated functional features that are significantly enriched for substrate cleavage. First, we identified overrepresented functional features by performing two-sided hypergeometric tests based on the benchmark dataset. The significantly enriched KEGG, GO and STRING features for each protease are shown in Supplementary Tables S3–S9 (only top ten features are listed). Taking caspase-3 as an example, in terms of the PPI features, there are 836 enriched protein interaction partners of caspase-3. In terms of the KEGG pathway features, caspase-3 substrates are enriched in signaling pathway terms, such as 'Fc epsilon RI signaling pathway', 'ErbB signaling pathway', 'toll-like receptor signaling pathway', 'transcriptional misregulation in cancer' and 'GnRH signaling pathway'. In addition, caspase-3 substrates are also found to be enriched in certain cancer-related pathways, including 'Small cell lung cancer', 'Transcriptional misregulation in cancer', 'Endometrial cancer' and 'Prostate cancer' (Supplementary Table S4). These results suggest that cleavage of caspase-3 substrates is implicated in cellular processes related to signaling or cancer pathways.

In terms of BP terms, again, we found that caspase substrates are enriched in 'platelet activation', 'cell proliferation', 'mRNA splicing', 'phosphorylation' and 'signaling pathways' (Supplementary Table S4). In terms of MF, caspase-3 substrates are enriched in functions related to binding, such as 'protein C-terminus binding', 'p53 binding', 'enzyme binding' and 'protein phosphatase binding'. It is also noteworthy that some enriched terms are related to protein post-translational modifications, such as 'protein autophosphorylation', 'protein phosphatase binding' and 'ubiquitin-protein ligase activity'. This finding recapitulates our refreshed knowledge about the presence and extent of functional cross-talks between caspase cleavage, phosphorylation and other types of post-translational modifications (Dix *et al.*, 2012; Kurokawa and Kornbluth, 2009). On the other hand, although there are certain functional terms shared in common by different proteases, such as 'Alzheimer's disease', most of the enriched terms are different and indicate protease-specific preferences of the target substrates. These results again highlight the needs to develop specific models for each protease to better understand their substrate specificity and facilitate the discovery of their native substrates.

Next, we wanted to find out whether substrate proteins can be discriminated from the background protein set by considering the cohort of the significantly enriched functional features. Using a simple log-odd approach (Li *et al.*, 2010), we calculated the function scores for each protein (See the Supplementary File for its definition). The distributions of the functional scores are shown in Figure 2. The majority of the proteins in the background set have scores <0, which is in sharp contrast to those

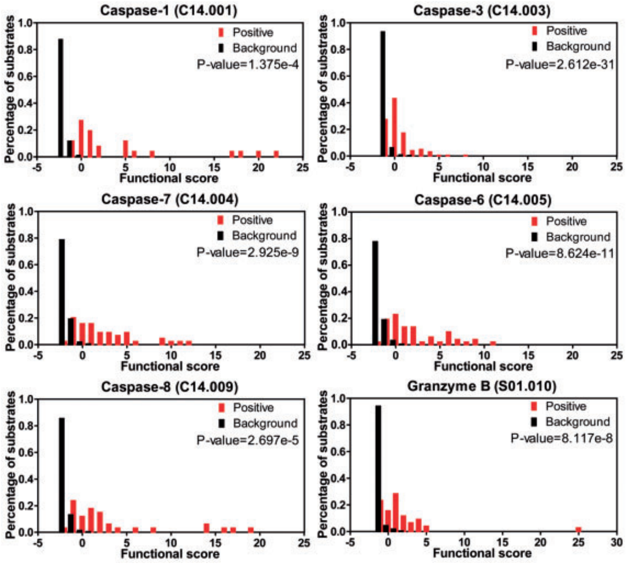

Fig. 2. Functional score distributions for the background protein set (black) and known substrates (red) of caspase-1, 3, 6, 7, 8 and granzyme B

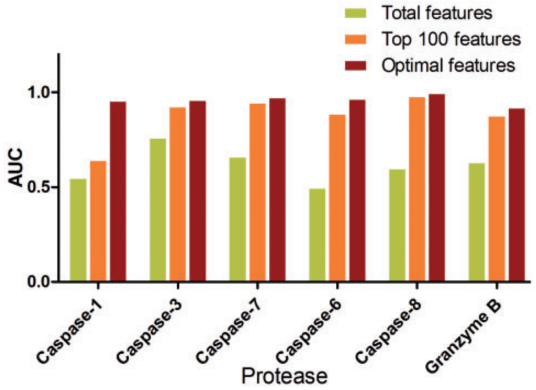

Fig. 3. Predictive performance of models for six proteases using the total features, the top 100 OFCs and the final optimal features, respectively, as evaluated by the AUC scores

of the substrate proteins most of which have positive score values. The statistical *t*-tests also suggest that the difference is significant across all the six proteases, with a *P*-value of 1.5e-4 or less (Fig. 2). These results indicate that real substrate proteins can be discriminated from the background protein set and our selected characteristic functional features might be informative for improving the substrate prediction (results discussed in Section 3.3).

3.3 Performance improvement based on a two-step feature selection using mRMR and FFS

Owing to the heterogeneous nature of the complex features extracted from various sources in this study, we are motivated to apply feature selection techniques (Saeys *et al.*, 2007; Wang *et al.*, 2012), to select more relevant features that are critical for cleavage site prediction. As identifying the most informative features is necessary for minimizing the classification error (Peng *et al.*, 2005) and shedding light on the important determinants of the protease substrate specificity, we used both mRMR and FFS methods to select the most informative feature subset in a two-step fashion. In the first step, we used mRMR to rank all the initial features and generate the top 100 features list, respectively, for the six proteases. The mRMR can output two different feature lists: the MaxRel list that provides the rank of features according to their relevance to the target and the mRMR list that provides the final mRMR rank after considering both max-relevance and min-redundancy. We used the mRMR list as the final selected features.

After the first step feature selection using mRMR, we selected the top 100 features for each protease, and then compared the performance of the three types of models based on three different groups of features when using the AUC scores as the metric for predictive power, as shown in Figure 3. The final optimal features set is provided in Supplementary Table S11. This figure indicates expected results, i.e. that by training models using the selected top 100 OFCs, we can achieve a significantly better performance. Taking caspase-1 as an example, we observe a 17% improvement of the overall AUC score obtained by the model based on the top 100 features (AUC = 0.645), when compared with the AUC score obtained by the model based on all the total features (AUC = 0.552). We observe a performance improvement across all the six proteases under investigation, ranging from 17 to 78%. These findings suggest that the mRMR feature selection can substantially improve the prediction performance with the selection of a compact subset of more informative and relevant features.

Next, we wanted to ascertain if it is possible to further select a subset of optimal features that could lead to the overall best performance. To answer this question, we focused on the selected 100 top OFCs and performed a second-step feature selection by iteratively adding/removing each of the features and examining the subsequent performance change. The performance changes during the course of feature selection were reflected by the FFS curves in Supplementary Figure S2, from which we could identify the corresponding subset of optimal features for each protease that achieved the best AUC score (i.e. the highest peaks in Supplementary Fig. S2). It can be seen from Figure 3 that the performance was consistently improved after the second-step feature selection. Compared to the models based on the top 100 features, the models based on the final optimal have further improved the predictions, with an improvement between 2 and 48%, depending on the protease type. In summary, these results suggest that this two-step feature selection is critical for improving the performance of the models. To our knowledge, this study represent the first systematic effort to scrutinize and select more relevant features that make a significant contribution to the prediction of caspase and GrB cleavage sites based on feature selection, in contrast to previous studies few of which have used feature selection methods to address this important task.

Supplementary Table S10 shows the number of the final selected features for the six proteases and their categorization. It can be seen that the total numbers of final features for each protease vary, ranging from 6 to 73. In addition, the number of final features is not correlated with the model's performance, i.e. a model with less number of final features may achieve a better performance, such as caspase-1, 7 and 6 (all with ≤15 final features), whereas the reverse is also true, namely, a model may require more final features to achieve a better performance, such as caspase-3 and GrB (both with >50 final features). On the other hand, we also observe some common features included in all the final feature sets across the six proteases, such as the STRING, GO and KEGG functional features. Inclusion of these common functional features may indicate the commonality of the substrate functions of these proteases. Models of caspase-3 and granzyme B also include Pfam and Interpro functional features, suggesting that that they are important for cleavage site prediction of these two enzymes.

We evaluated the relative importance of each class of functional features to the predictive power of the models, based on the mRMR scores, as shown in Supplementary Table S11. The analysis shows that functional features are highly abundant, accounting for almost 50% of the final selected features, and are generally assigned with higher rankings. For example, STRING and GO CC features are ranked as the top five features for most proteases. To further quantitatively evaluate their contribution to the performance, we measured the performance loss using the AUC score by removing a feature group from the model (Supplementary Table S12). The results show that STRING functional features are the most important feature group for all the six proteases except caspase-1 because removal of this group resulted in the largest performance loss.

Finally, using the mRMR scores, we characterized the relative importance of each individual optimal feature. Taking caspase-6 as an example, the most important feature type is the STRING feature, followed by GO CC, AA_index_subst.V1872 and GO BP (see the Supplementary File for explanation of the features) (Supplementary Fig. S3). More than half of the features (7 out of 13) belong to functional features, indicating that they can interact cooperatively with other complementary features to achieve a better performance. A complete list of the final selected features for all the six proteases and the respective mRMR scores is provided in Supplementary Table S11.

3.4 Prediction performance of Cascleave 2.0 based on the benchmark datasets

In this section, we assessed the prediction performance of the SVR models of Cascleave 2.0 using the final optimal features by the 5-fold cross-validation tests based on the benchmark datasets. The performance was assessed using six different measures, including MCC, ACC, SEN, SPE, PRE and AUC. The results are shown in Table 1. One can see that the models of all the six proteases achieved relatively high performance, with MCC and AUC ranging from 0.744 to 0.978 and 0.922 to 0.997, respectively. In contrast, the performance of Granzyme B is the worst predicted with an MCC of only 0.744, whereas the performance of caspase-8 is the best with MCC and AUC scores close to 1.

Table 1. Performance of Cascleave 2.0 based on the benchmark datasets by 5-fold cross-validation tests for caspase-1, 3, 7, 6, 8 and granzyme B

Protease	MCC	ACC	SEN	SPE	PRE	AUC
Caspase-1	0.890	0.945	0.958	0.935	0.920	0.958
Caspase-3	0.819	0.909	0.926	0.893	0.894	0.962
Caspase-7	0.900	0.949	0.918	0.980	0.978	0.976
Caspase-6	0.851	0.925	0.913	0.937	0.935	0.969
Caspase-8	0.978	0.989	1.000	0.978	0.979	0.997
Granzyme B	0.744	0.871	0.835	0.907	0.899	0.922

Our work distinguished itself from previous studies in that previous studies mostly used sequence or sequence-derived structural features to build the models, failing to consider and include informative functional features. As previously discussed, in this study, we found that there were a number of functional features that significantly contribute to the substrate cleavage site prediction, such as GO BP, MF, CC, KEGG and PPI features (Supplementary Table S11). These findings actually make sense, because protease and its substrates have to be co-localized in the same cellular compartments or function in the related pathways or interaction networks to satisfy its physiological role to cleave target substrates *in vivo*(Gromiha *et al*., 2010). As a result, global functional features help predict potential target substrates.

On the other hand, sequence features also prove useful for predicting substrate cleavage sites. The selected optimal sequence features include AA index, amino acid properties, predicted secondary structure, predicted disorder region, pI, etc (Supplementary Table S11). Because these sequence features has been well established in previous studies (Ayyash *et al*., 2012; Barkan *et al*., 2010; Garay-Malpartida *et al*., 2005; Huang *et al*., 2006; Li *et al*., 2012; Mizianty *et al*., 2010; Song *et al*., 2010; Verspurten *et al*., 2009), we did not focus on the analysis of these features in this section. One important issue is that with the application of powerful feature selection techniques, like the ones used in this study, enables us to quantify the relative importance and contribution of each of the selected feature types to substrate cleavage site prediction, providing important insights into various determinants of the substrate specificities of the proteases arising from different levels.

We further investigated the contribution of the selected functional features to the classifiers' performance by comparing the AUC measures between models trained with and without functional features (Fig. 4). The results indicate that the selected functional features are critical for the performance improvement, especially for caspase-1, 6 and 7, for which the AUC scores increased from 0.575 to 0.958, 0.602 to 0.969 and 0.737 to 0.976, respectively. Taken together, these results again highlight the complementarity and importance of the selected functional features to the predictive power of the SVR models.

3.5 Comparison with other existing tools based on the independent test datasets

In this section, we compared the performance of Cascleave 2.0 with other state-of-the-art tools including PoPS (Boyd *et al*., 2005), SitePrediction (Verspurten *et al*., 2009), PCSS (Barkan *et al*., 2010), Pripper (Piippo *et al*., 2010), PeptideCutter (Wilkins *et al*., 1999), CASVM (Wee *et al*., 2007), Cascleave 1.0 (Song *et al*., 2010) and CAT3 (Ayyash *et al*., 2012), based on the independent test datasets. In addition, we also compared with BLAST prediction, which is one of the most readily

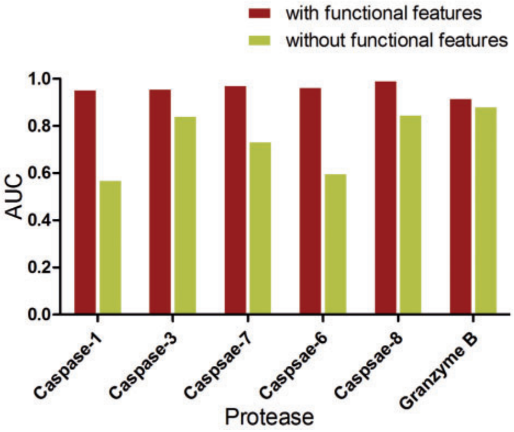

Fig. 4. Performance comparison using the AUC scores between models that incorporated selected functional features and that did not

available sequence analysis tools using sequence similarity. Although CaSPredictor and GraBCas are previously developed as useful tools, they are currently unavailable. Therefore, we did not include them in the comparison. It should be noted that while Cascleave 2.0 is able to predict protease-specific substrate cleavage sites for all the six proteases under investigation, some tools could only be used to predict cleavage sites for one or more, but not for all the proteases. For example, CAT3 can be applied to predict cleavage sites of caspase-3 only. Therefore, we compared with the tools that could provide valid predictions for a protease of interest. In Figure 5, some ROC curves are not smooth because the sizes of corresponding independent datasets are small.

For most of the proteases, we found that SVM was better than RF, J48 and Vote for Pripper except caspase-3. The performance of CASVM showed that the local window size of P16-P10′ better than P4-P2′ and P4-P1, except for caspase-6. The performance comparison was made using the ROC curves (Fig. 5) and measures such as MCC and ACC (Supplementary Tables S13 and S14). These results clearly show that Cascleave 2.0 outperformed other tools for five proteases, i.e. caspase-1, 3, 6, 7 and granzyme B, but slightly worse than SitePrediction in the case of caspase-8, for which Cascleave 2.0 achieved an AUC of 0.938, slightly lower than that of SitePrediction (AUC = 0.979). To examine the statistical significance, we further performed pairwise *t*-test and showed that the performance differences between Cascleave 2 and other methods were, in most cases, statistically significant (Supplementary Table S16). The performance improvement of Cascleave 2.0 compared with other tools might be attributed to the incorporation of informative functional features and the deployment of an efficient two-step feature selection in an integrated framework.

SitePrediction is an empirical scoring tool for protease cleavage site prediction (Verspurten *et al.*, 2009) by combining amino acid frequency score with amino acid substitution matrix. It is especially effective for predicting potential cleavage sites that are similar to known cleavage sites. Our studies confirm the predictive power of SitePrediction. Similarly, PoPS, CAT3 and PeptideCutter are also frequency-based scoring tools and performed worse than SitePrediction. On the other hand, machine learning-based tools such as PCSS, CASVM and Cascleave 1.0, also achieved varying prediction performance, depending on the protease type. PCSS is an SVM-based tool that is developed using sequence and structural features important for recognition of substrate peptides. It also achieved a prediction performance comparable or better than other tools, particularly for caspase-3, 7, 8 and granzyme B. Finally, we would like to point out that it is practically difficult to perform an unbiased comparison, as we are faced with a situation where the knowledge of the training substrate data used by other tools is not available and re-training of the models of other tools based on the same independent datasets is often not possible.

In addition, we observed the prediction performance of granzyme B substrate cleavage sites is relatively worse, across all the compared tools, as opposed to the caspases. The underlying reason behind this result is unclear. One possibility is that extended substrate specificity (including for example the contribution of exosites) may be less well captured by sequence-derived models.

In summary, our analysis indicates that Cascleave 2.0 has outperformed other tools for most proteases, through the integration of informative sequence and functional features. Cascleave 2.0 can be used as a powerful tool for predicting difficult cleavage sites that cannot be readily identified by tools based on sequence information only.

3.6 Application of Cascleave 2.0 to human proteome

After showing the predictive ability of Cascleave 2.0 to predict substrate cleavage sites using both the benchmark and independent tests, we further applied it to screen the human proteome (with a total of 86 728 proteins) for novel substrates and their cleavage sites. Models were trained using the final optimal features based on the complete training dataset. Substrates containing predicted cleavage sites at the 99% specificity level were considered as high-confidence putative targets of the protease. A complete list of the predicted substrates and cleavage sites is given at the Web site (http://www.structbioinfor.org/cascleave2), along with the implemented Java Applet and user instructions. An attractive advantage of the Java implementation is that it provides a user-friendly interface and allows for a proteome-wide screening analysis (see Supplementary Fig. S5 for screenshots). The statistics of predicted cleavage sites of caspase-1, 3, 7, 6, 8 and granzyme B are shown in Supplementary Table S15. In contrast, caspase-1 has the lowest number of predicted substrates and cleavage sites (509 substrates/9591 sites). On the other hand, granzyme B substrates have on average the smallest number of predicted cleavage sites (~2.2 sites per substrate). In addition, scalability of our algorithm is also important, as it will hinder a wider adoption if it takes hours to process some of the results. We performed proteome-wide substrate screening of the mouse proteome and this process took about 24 h to finish. Owing to limited availability of experimentally verified substrate data for the murine caspases, we only trained the model for the murine caspase-1 and applied the trained model to identify a number of high-confidence substrates. All proteome-wide prediction results can be downloaded from our Web site. In addition, as more experimentally verified substrate data become available, we will

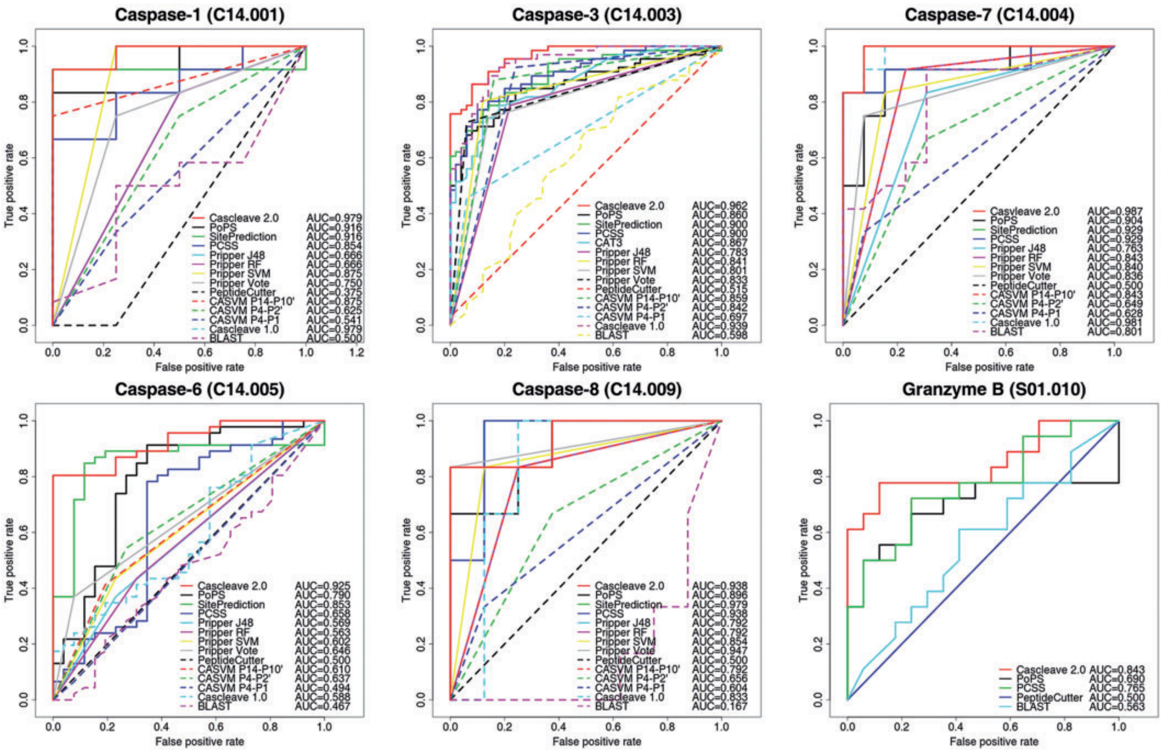

Fig. 5. ROC curves of Cascleave 2.0 and other tools for cleavage site prediction for different proteases based on the independent test datasets

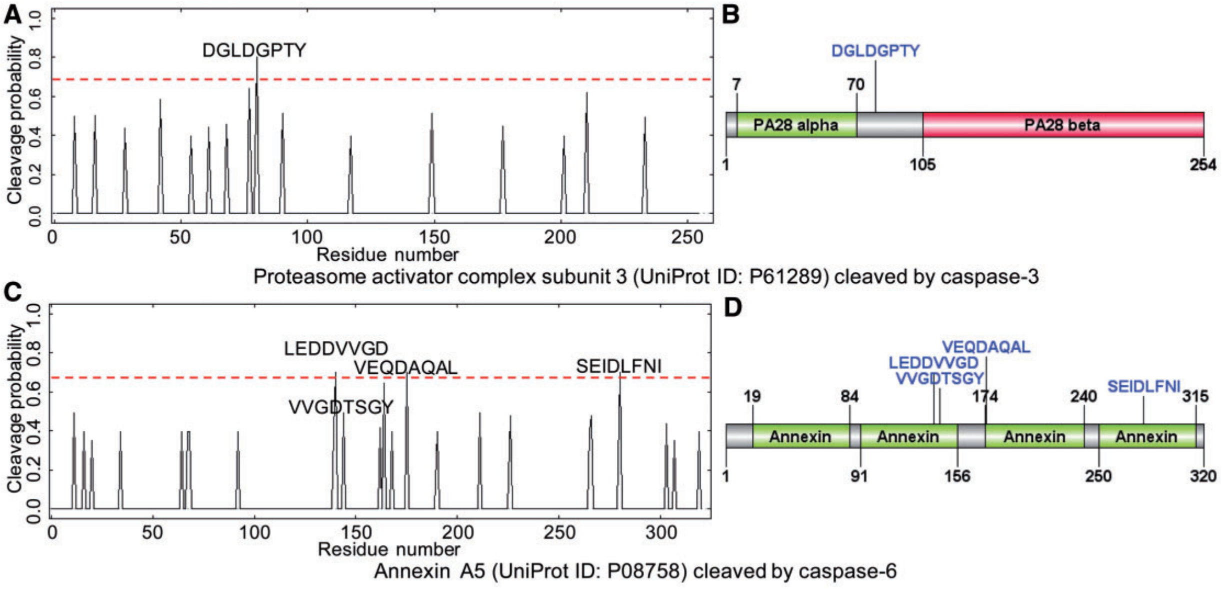

Fig. 6. The predicted cleavage probability for caspase cleavage sites using Cascleave 2.0. (A) Sequence-scanning results and (B) predicted cleavage site locations with respect to the domain of proteasome activator complex subunit 3; (C) sequence-scanning results and (D) predicted cleavage sites with respect to the protein domain of Annexin A5 (Uniprot ID: P08758)

regularly incorporate these data to update prediction models and expand the utility of our tool. These proteome-wide predictions provide a valuable resource for experimental validation of novel human substrates and the proposition of useful hypotheses.

3.7 Case study

To further demonstrate the predictive power of Cascleave 2.0, we performed a case study of two different substrates for which the cleavage sites have been experimentally validated.

Sequence-scanning results and the predicted cleavage sites with respect to the protein domain are displayed in Figure 6.

The first example is the proteasome activator complex subunit 3 (REG-gamma UniProt ID: P61289), which activates the trypsin-like catalytic subunit of the proteasome but inhibits the chymotrypsin-like and postglutamyl-preferring subunits (Realini *et al*., 1997; Wilk *et al*., 2000). Site-directed mutagenesis indicates that REG-gamma has one cleavage site of caspase-3: DGLD|GPTY ('|' denotes the cleavage site) (Araya *et al*., 2002). Cascleave 2.0 correctly predicted this cleavage site, with a cleavage probability score of 0.804. The second example is the Annexin A5 (UniProt ID: P08758) (Bogdanova *et al*., 2007), an anticoagulant protein that acts as an indirect inhibitor of the thromboplastin-specific complex (Grundmann *et al*., 1988). Annexin A5 has four experimentally determined cleavage sites of caspase-6 (Klaiman *et al*., 2008). Cascleave 2.0 successfully predicted three of them: LEDD|VVGD, VEQD|AQAL and SEID|LFNI, failing to identify the fourth cleavage site: VVGD|TSGY, when assigning a higher cutoff threshold of 0.65. Nevertheless, if a lower cutoff of 0.49 was used, RDPD|AGID site would have been included as a cleavage site, with a ranking of fourth place, among the five predicted sites. These results suggest that Cascleave 2.0 can be used as a useful tool for *in silico* cleavage site prediction.

4 CONCLUSION

In summary, we provided a novel approach, termed Cascleave 2.0, which has significantly improved the prediction of protease-specific substrate cleavage sites for caspase-1, 3, 6, 7, 8 and granzyme B, by integrating heterogeneous sequence and functional features from multiple sources. It uses a two-step integrative framework to characterize protein sequence and functional features that are important for determining the substrate specificity of different proteases. Benchmarking experiments indicate that Cascleave 2.0 is able to provide a performance that is better than or competitive with other existing tools. Prioritization and dissection of novel protease-substrate interactions will likely benefit from the development of caspase-specific bioinformatic tools such as Cascleave 2.0. To the best of our knowledge, this study represents the first systematic effort to scrutinize and select more relevant features that make a significant contribution to the prediction of caspase and GrB cleavage sites based on feature selection, in contrast to previous studies few of which have used feature selection methods to address this important task. Moreover, the high-confidence predicted substrates and their cleavage sites produced in this study also provide a rich knowledge base for follow-up hypothesis-driven experiments in the field of protease biology.

Funding: Major Inter-disciplinary Research Project Grant awarded by Monash University, Hundred Talents Program of the Chinese Academy of Sciences (CAS), National Natural Science Foundation of China (91130032, 61103075, 61202167, 61303169), Innovation Program of Shanghai Municipal Education Commission (13ZZ072), National Health and Medical Research Council of Australia (NHMRC) (490989) and Australian Research Council (ARC) (LP110200333). XMZ was partly supported by the National Natural Science Foundation of China (91130032, 61103075), Innovation Program of Shanghai Municipal Education Commission (13ZZ072) and Shanghai Pujiang Program (13PJD032). J.S. is a NHMRC Peter Doherty Fellow and Recipient of the Hundred Talents Program of CAS. Chinese Academy of Sciences Visiting Professorship for Senior International Scientists (2011T2S34) (to T.A.). J.C.W. is an ARC Federation Fellow and NHMRC Senior Principal Research Fellow.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Araya,R. *et al.* (2002) Yeast two-hybrid screening using constitutive-active caspase-7 as bait in the identification of PA28gamma as an effector caspase substrate. *Cell Death Differ*., **9**, 322–328.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Ayyash,M. *et al.* (2012) Developing a powerful in silico tool for the discovery of novel caspase-3 substrates: a preliminary screening of the human proteome. *BMC Bioinformatics*, **13**, 14.

Backes,C. *et al.* (2005) GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Res*., **33**, W208–213.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.

Barkan,D.T. *et al.* (2010) Prediction of protease substrates using sequence and structure features. *Bioinformatics*, **26**, 1714–1722.

Bogdanova,N. *et al.* (2007) A common haplotype of the annexin A5 (ANXA5) gene promoter is associated with recurrent pregnancy loss. *Hum. Mol. Genet*., **16**, 573–578.

Boyd,S.E. *et al.* (2005) PoPS: a computational tool for modeling and predicting protease specificity. *J. Bioinform. Comput. Biol*., **3**, 551–585.

Bredemeyer,A.J. *et al.* (2005) Use of protease proteomics to discover granzyme B substrates. *Immunol. Res*., **32**, 143–153.

Burges,C.J.C. (1998) A tutorial on Support Vector Machines for pattern recognition. *Data Min. Knowl. Discov*., **2**, 121–167.

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM:a library for support vector machines. *ACM Trans. Intell. Syst. Technol*., **2**, 1–26.

Chen,C.T. *et al.* (2008) Protease substrate site predictors derived from machine learning on multilevel substrate phage display data. *Bioinformatics*, **24**, 2691–2697.

Chowdhury,I. *et al.* (2008) Caspases — an update. *Comp. Biochem. Physiol. B Biochem. Mol. Biol*., **151**, 10–27.

Colaert,N. *et al.* (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods*, **6**, 786–787.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn*., **20**, 273–297.

Dix,M.M. *et al.* (2008) Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell*, **134**, 679–691.

Dix,M.M. *et al.* (2012) Functional interplay between caspase cleavage and phosphorylation sculpts the apoptotic proteome. *Cell*, **150**, 426–440.

duVerle,D.A. and Mamitsuka,H. (2012) A review of statistical methods for prediction of proteolytic cleavage. *Briefings Bioinformatics*, **13**, 337–349.

Enoksson,M. and Salvesen,G.S. (2008) Proteolytic needles in the cellular haystack. *Nat. Chem. Biol*., **4**, 651–652.

Enoksson,M. *et al.* (2007) Identification of proteolytic cleavage sites by quantitative proteomics. *J. Proteome Res*., **6**, 2850–2858.

Fischer,U. *et al.* (2003) Many cuts to ruin: a comprehensive update of caspase substrates. *Cell Death Differ*., **10**, 76–100.

Garay-Malpartida,H.M. *et al.* (2005) CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics*, **21 (Suppl. 1), i169–i176.**

Gasteiger,E. *et al.* (2005) Protein identification and analysis tools on the ExPASy server. In: Walker,J.M. (ed.) *The Proteomics Protocols Handbook*. Humana Press, Totowa, New Jersey, pp. 571–607.

Gromiha,M.M. *et al.* (2010) Sequence and structural analysis of binding site residues in protein-protein complexes. *Int. J. Biol. Macromol*., **46**, 187–192.

Grundmann,U. *et al*. (1988) Characterization of cDNA encoding human placental anticoagulant protein (PP4): homology with the lipocortin family. *Proc. Natl Acad. Sci. USA*, **85**, 3708–3712.

Huang,D.S. *et al*. (2006) Classifying protein sequences using hydropathy blocks. *Pattern Recogn*., **39**, 2293–2300.

Huang,Y. *et al*. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Hunter,S. *et al*. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*., **40**, D306–D312.

Jensen,L.J. *et al*. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*., **37**, D412–D416.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.

Kawashima,S. *et al*. (1999) AAindex: amino acid index database. *Nucleic Acids Res*., **27**, 368–369.

Klaiman,G. *et al*. (2008) Targets of caspase-6 activity in human neurons and Alzheimer disease. *Mol. Cell Proteomics*, **7**, 1541–1555.

Kurokawa,M. and Kornbluth,S. (2009) Caspases and kinases in a death grip. *Cell*, **138**, 838–854.

Li,B.Q. *et al*. (2012) Prediction of protein cleavage site with feature selection by random forest. *PLoS One*, **7**, e45854.

Li,T. *et al*. (2010) Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. *PLoS One*, **5**, e15411.

Los,M. *et al*. (2001) Caspases: more than just killers? *Trends Immunol*., **22**, 31–34.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Mizianty,M.J. *et al*. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.

Nicholson,D.W. (1999) Caspase structure, proteolytic substrates, and function during apoptotic cell death. *Cell Death Differ*, **6**, 1028–1042.

Pardo,J. *et al*. (2009) The biology of cytotoxic cell granule exocytosis pathway: granzymes have evolved to induce cell death and inflammation. *Microbes Infect*., **11**, 452–459.

Peng,H. *et al*. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell*., **27**, 1226–1238.

Piippo,M. *et al*. (2010) Pripper: prediction of caspase cleavage sites from whole proteomes. *BMC Bioinformatics*, **11**, 320.

Punta,M. *et al*. (2012) The Pfam protein families database. *Nucleic Acids Res*., **40**, D290–D301.

Rawlings,N.D. *et al*. (2008) MEROPS: the peptidase database. *Nucleic Acids Res*., **36**, D320–D325.

Realini,C. *et al*. (1997) Characterization of recombinant REGalpha, REGbeta, and REGgamma proteasome activators. *J. Biol. Chem*., **272**, 25483–25492.

Russell,J.H. and Ley,T.J. (2002) Lymphocyte-mediated cytotoxicity. *Annu. Rev. Immunol*., **20**, 323–370.

Saeys,Y. *et al*. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Schilling,O. and Overall,C.M. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat. Biotechnol*., **26**, 685–694.

Song,J. *et al*. (2007) Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics*, **23**, 3147–3154.

Song,J. *et al*. (2010) Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics*, **26**, 752–760.

Song,J. *et al*. (2011) Bioinformatic approaches for predicting substrates of proteases. *J. Bioinform. Comput. Biol*., **9**, 149–178.

Song,J. *et al*. (2012) PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One*, **7**, e50300.

Stajich,J.E. *et al*. (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.

Team,R.D.C. (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Austria.

Turk,B. (2006) Targeting proteases: successes, failures and future prospects. *Nat. Rev. Drug Discov*., **5**, 785–799.

Turk,B.E. *et al*. (2001) Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat. Biotechnol*., **19**, 661–667.

Verspurten,J. *et al*. (2009) SitePredicting the cleavage of proteinase substrates. *Trends Biochem. Sci*., **34**, 319–323.

Wagner,M. *et al*. (2005) Linear regression models for solvent accessibility prediction in proteins. *J. Comput. Biol*., **12**, 355–369.

Wang,M. *et al*. (2012) FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS One*, **7**, e43847.

Ward,J.J. *et al*. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol*., **337**, 635–645.

Wee,L.J. *et al*. (2007) CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics*, **23**, 3241–3243.

Wilk,S. *et al*. (2000) Properties of the nuclear proteasome activator PA28gamma (REGgamma). *Arch. Biochem. Biophys*., **383**, 265–271.

Wilkins,M.R. *et al*. (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol*., **112**, 531–552.

Zhao,X.M. *et al*. (2010) A discriminative approach for identifying domain-domain interactions from protein-protein interactions. *Proteins*, **78**, 1243–1253.

Zhao,X.M. *et al*. (2005) A novel approach to extracting features from motif content and protein composition for protein sequence classification. *Neural Netw*., **18**, 1019–1028.

Zhao,X.M. *et al*. (2008) Protein classification with imbalanced data. *Proteins*, **70**, 1125–1132.