1 INTRODUCTION

To date 14 mammalian caspases have been identified; these molecules in function in signaling cascades that control critical processes such as apoptosis, necrosis, inflammation, migration and differentiation (Dix et al., 2008; Fischer et al., 2003; Lithé and Martin, 2007; Mahrous et al., 2008; Nicholson, 1999; Pop and Salvesen, 2009; Talanian et al., 1997; Timmer and Salvesen, 2007). Caspases cleave substrates that generally contain a conserved aspartate (D) at the P1 position (Nicholson, 1999; Pop and Salvesen, 2009; Talanian et al., 1997). However, in addition to the P1 Asp, the amino acids at the P3 and P4 positions contain important additional determinants of specificity that can dramatically affect cleavage efficiency. Accordingly, mammalian caspases can be divided into three groups (Nicholson, 1999). Group I caspases (caspase-1, -4, -5 and -13) prefer bulky hydrophobic amino acids at the P3 site and recognize peptide sequence (W/L)E(H/T)D. In contrast to the other caspases, caspase-14 is expressed and activated mainly in the epidermis and exhibits cleavage preference for the WEHD or IETD motif (Denecker et al., 2008).

The study of apoptotic pathways predominantly mediated by caspases has important implications for the development of therapies for cancer treatment. There is significant interest in gaining a better understanding of caspase substrate specificity (Dix et al., 2008; Lithé and Martin, 2007; Mahrous et al., 2008; Timmer et al., 2009). Even though almost 400 caspase substrates have been reported to date, there are likely to be hundreds of new caspase substrates that remain to be discovered (Fischer et al., 2003). Experimental identification and characterization of protease substrates is often difficult and time-consuming (Enoksson and Salvesen, 2008; Enoksson et al., 2007; Jon et al., 2007; Rawlings et al., 2008; Schilling and Overall, 2008). Hence, computational prediction of caspase substrate specificity may provide useful and experimentally testable information in regards to novel potential cleavage sites or candidate substrates.

Several computational approaches have been developed to predict caspase cleavage-site specificity. PeptideCutter utilized a limited experimental dataset and was initially used to predict the substrate cleavage sites for a variety of protease families including several
We extensively explored different sequence encoding schemes and sequences. In each validation step, one subset was singled out in turn as the In order to objectively evaluate the prediction performance, we employed 5-cleavage sites were verified experimentally. The current dataset contains 370 et al et al et al 2.1 Datasets caspases (Gasteiger et al., 2005). Lohmüller et al. (2003) developed a bioinformatic tool called prediction of endopeptidase substrates (PEPS) that utilized rule-based endopeptidase cleavage site scoring matrices (CSSM) and was deployed to predict caspase-3 substrates. Garay-Malpartida et al. built CasPredictor to predict caspase cleavage sites; this approach again utilizes sequence and specificity information but also incorporates an index of PEST-like sequences enriched in the vicinity of cleavage site regions (Garay-Malpartida et al., 2005; Singh, 2006). The latter program achieved an accuracy of 81% when evaluated on a dataset of 137 experimentally verified cleavage sites (Garay-Malpartida et al., 2005). Yang (2005) applied neural networks to build models for predicting caspase cleavage sites and investigated the impact of a sliding window size on the performance. This approach attained the highest reported prediction accuracy of 79%, but it is important to note that only a small dataset of 13 substrate sequences was considered. The GraB Cas software developed by Backes et al. (2005) provided position-specific scoring prediction of cleavage sites for caspases 1–9 and granzyme B. Wee et al. applied a support vector machine (SVM) approach and showed that CASVM predictors could achieve an accuracy ranging from 81.2 to 97.9% based on an extended dataset of 210 cleavage sites. More recently, they proposed a two-step model called multi-factor CASVM which exploits the structural factors to score and filter the false positives to improve the performance (Wee et al., 2009).

In this work, we extend the SVM approach proposed by Wee et al. (2006, 2007, 2009) to predict caspase cleavage sites from the flanking sequences of substrates. We started with the construction of a caspase substrate database with experimentally verified substrate sequences extracted from multiple resources. We then built support vector regression (SVR) models that not only provided conventional two-state (cleavage or non-cleavage) prediction, but also generated an estimated probability for each candidate cleavage site, thus providing a quantitative evaluation of caspase substrate specificity. We extensively explored different sequence encoding schemes and examined their effects on the prediction performance. Further, we took into consideration the characteristic sequence and structural features surrounding substrate cleavage and non-cleavage sites, such as the predicted secondary structure, solvent accessibility and natively disordered regions, based on a recently developed bi-profile Bayesian feature extraction method (Shao et al., 2009). Comparison with the published approaches reveals that our approach is generally useful for identification of caspase cleavage sites in large datasets.

2 METHODS
2.1 Datasets
We have constructed a caspase substrate sequence dataset from multiple resources, including manually curated CASBAH database (Fischer et al., 2003), the MEROPS database (Rawlings et al., 2008), the CASVM webserver (Wee et al., 2006, 2007), the Uniprot database (Bairoch and Apweiler, 2000), as well as a literature search. All the annotated substrate cleavage sites were verified experimentally. The current dataset contains 370 caspase substrate sequences and 563 cleavage sites (Supplementary Table 1).

In order to objectively evaluate the prediction performance, we employed 5-fold cross-validation and leave-one-out cross-validation (LOOCV) methods. In the case of 5-fold cross-validation, substrate sequences in this dataset were randomly divided into five subsets with roughly equal numbers of substrate sequences. In each validation step, one subset was singled out in turn as the testing dataset, while the rest were used as the training dataset. In the case of LOOCV, each substrate sequence in the dataset was singled out in turn as the testing set and this procedure was repeated for all the substrate sequences. Moreover, in order to make a more stringent comparison with CASVM, we also used an independent testing set extracted from a recent experimental study (Dix et al., 2008), which contains 64 caspase substrate sequences and 69 cleavage sites. The complete list of these substrate sequences used in this study is available in the Supplementary Data and can be downloaded from the Cascleave website (http://sunflower.kaiu.kyoto-u.ac.jp/~sjin/Cascleave/).

Cleavage and non-cleavage sites were collected as positive and negative datasets, respectively. Peptide sequences in positive and negative datasets were extracted using a sliding window approach surrounding the experimentally verified cleavage sites and other residues that were found not to be cleaved by caspases, respectively. The predictive models were then built based on the extracted positive and negative peptide sequences. As the obtained datasets this way might contain sequence redundancy which will lead to the overestimation of the prediction performance of the models, we need to reduce sequence homology for the extracted positive and negative peptide sequences. Local window-based homology reduction (Shao et al., 2009) was adopted for this purpose. Sequence homology reduction within the training and testing datasets was performed in such a way that sequence identity between any two peptide sequences should not be larger than 70%.

2.2 Binary encoding amino acid sequence profiles (BEA)
In the first instance, the SVR models were trained and tested using binary encoding amino acid sequence profiles (BEA) where substrate sequences were transformed into n-dimensional vectors using an orthonormal encoding scheme, in which each amino acid is represented by the 20-D binary vector composed of either zero or one element (Shao et al., 2006). For example, Ala (00000000000000000000), Cys (01000000000000000000), Asp (00100000000000000000), Thr (00000000000000000001), etc. For the sake of simplicity, we termed this binary encoding amino acid sequence profile as the encoding scheme ‘BEA’. Since increasing sequence window size is supposed to provide more local sequence information, we used a sliding window approach to derive the local sequence profiles based on the ‘BEA’ scheme and examined the corresponding prediction accuracy. The window size w is defined as the residue numbers involved in the local sequence windows surrounding the cleavage sites from P0 to Pw positions, either in a symmetrical or non-symmetrical manner (Supplementary Fig. 1): i.e. w = 3 (Pw-P0), 4 (Pw-1-P0 or P0-Pw-1), 5 (Pw-2-P0 or P0-Pw-2), 6 (Pw-3-P0), …, 14 (Pw-7-P0), 16 (Pw-8-P0), etc.

2.3 Predicted structural information
Based on the observation of structural determinants of caspase substrate specificity, we also incorporated into Cascleave the structural information predicted by state-of-art algorithms, specifically, secondary structures, solvent accessibility and natively unstructured regions. Secondary structure was predicted using the PSIPRED program, which provides one of the most accurate predictions for protein secondary structures and generates the probability profiles of three secondary structure assignments (helix, strand and coil) for each residue in a protein (Jones, 1999). For a given residue, we extracted the w × 3 matrix from the output file of PSIPRED by selecting the sliding window size w. Solvent accessibility was predicted using the SSpro program implemented in the SCRATCH package (Cheng et al., 2005). SSpro predicts the solvent accessibility status for each residue in a protein sequence, producing a binary output—either as ‘exposed’ or ‘buried’ (Cheng et al., 2005). Solvent accessibility was encoded as binary units into the SSR model. Natively unstructured region was predicted using the DISOPRED2 server (Ward et al., 2004), which is one of the leading servers for predicting natively disordered regions in proteins. The probability of each residue being disordered generated by DISOPRED2 is used as the input to the SVR models.
2.4 Bi-profile Bayesian signatures

Shao et al. (2009) recently proposed a novel approach called bi-profile Bayes and applied it to predict methylation sites in proteins. This approach has been demonstrated to provide a significant improvement of prediction performance for predicting methylation sites from protein sequences (Shao et al., 2009). We will describe the application of bi-profile Bayesian signature extraction approach to predict caspase cleavage sites in this study. The rationale behind this approach is that peptide sequences that can be cleaved by caspases should exhibit distinct amino acid or characteristics relative to those that cannot be cleaved. Therefore, integrating the bi-profile Bayesian signatures by representing each sample in a bi-feature manner would be more informative than the single BEAA mentioned above. This approach would be particularly useful when dealing with an unbalanced dataset comprising a smaller amount of positive samples and greater number of negative samples. Given a substrate peptide sequence \( S = (a_1, a_2, a_3, ..., a_n) \), where \( a_i \) denotes an amino acid at position \( i \), \( n \) denotes the sequence length of substrate peptide (i.e. the local sliding window size), \( S \) can be classified as one of the two classes \( C_1 \) (representing cleavage sites of caspases, i.e. positive samples) or \( C_0 \) (representing non-cleavage sites of caspases, i.e. negative samples), according to whether or not it can be cleaved by caspases. The posterior probability of both positive and negative samples can be calculated as the occurrence of each amino acid at each position in the training dataset. We integrated the bi-profile Bayesian signatures to predict cleavage sites of caspase from primary sequence of substrates (for more details about the bi-profile Bayesian feature extraction method, refer to the original work of Shao et al.).

2.5 Sequence encoding schemes

We extracted four different types of sequence-profiles based on the bi-profile Bayesian feature extraction approach: (i) bi-profile Bayesian amino acid profile (BPBAA); (ii) bi-profile Bayesian secondary-structure profile (BPBSS); (iii) bi-profile Bayesian solvent accessibility profile (BPBISA); and (iv) bi-profile Bayesian disordered profile (BPBDISO), through calculating the frequency of each amino acid or the corresponding structural types at each position of the non-canonical 'XXXD' motif (Dix et al., 2008; Fischer et al., 2003; Lüthi and Martin, 2007; Mahrous et al., 2008; Nicholson, 1999; Pop and Salvesen, 2009; Talanian et al., 1997; Timmer and Salvesen, 2007; Timmer et al., 2009).

Our analysis reveals that 96.99% of cleavage sites have a P₁ aspartate, while only a small percentage of cleavages sites exhibit the non-canonical 'XXXD' motif (Dix et al., 2008) and half sphere exposure (Song et al., 2008). We used the SVM_light package, an implementation of Vapnik’s SVM for SVC, SVR and pattern recognition (Joachims, 1999). We selected radial basis kernel function (RBF) at \( \epsilon = 0.01 \), \( \gamma = 0.01 \) and \( C = 100.0 \) to build the prediction models. This parameter set was optimized based on 5-fold cross-validation.

The number of negative samples is much larger than that of the positive samples, which will incur the imbalance problem and result in biased prediction in favor of the negative data (Song et al., 2006; Wee et al., 2006). We employed the under-sampling approach to overcome this imbalance problem by reducing the size of the over-represented negative samples. We set the ratio of the positive to negative data at 1:3. Random sampling still retained the original distribution of negative samples, but avoided their over-representation.

2.7 Performance evaluation

The predictive performance of our approach was evaluated using the accuracy, sensitivity, specificity, F-score and Matthews correlation coefficient (MCC) measures (Matthews, 1975) (see Supplementary Data for more details). The predictive performances were evaluated using these five measures based on 5-fold cross-validation and LOOCV tests.

3 RESULTS

3.1 Statistical distribution of substrate cleavage sites

Using the compiled dataset of experimentally determined caspase substrates, we analyzed the statistical distributions of substrate cleavage sites for P₈–P₉ positions. As shown in Figures 1 and 2, caspase-3 was used as an example to generate a heat map and sequence logo diagram, respectively, for the P₈–P₉ specificities of caspase-3. As expected, one of the prominent characteristics of caspase-3 cleavage site specificity is that this enzyme preferentially cleaves after the aspartate (t) at P₃ and P₄ position, which forms the well-known canonical ‘XXXD’ motif (Dix et al., 2008; Fischer et al., 2003; Lüthi and Martin, 2007; Mahrous et al., 2008; Nicholson, 1999; Pop and Salvesen, 2009; Talanian et al., 1997; Timmer and Salvesen, 2007; Timmer et al., 2009).

![Fig. 1. Caspase-3 substrate cleavage sites for P₈–P₉ position. The average amino acid occurrences in P₈–P₉ were calculated and displayed in the form of a two-dimensional heat map. The aspartic peptide bond between sites P₁ and P² is indicated by a vertical white line.](https://academic.oup.com/bioinformatics/article-abstract/26/6/752/245474/fig1)

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/26/6/752/245474 by guest on 04 March 2019
As indicated in previous studies, the three-dimensional (3D) and proline at P8 for caspase-10. of cleavage sites for P4–P′
positions. We also generated heat maps for the other caspases such as caspase-1, -6, -7, -8, -9 and -10, as well as sequence 
logos for P8–P′ positions (Supplementary Figs 1–3). Comparison of different caspase groups reveals distinct patterns of 
subsite specificity through the P8–P′ sites, for example, serine at P′ for caspase-7 and proline at P′ for caspase-10.

3.2 Analysis of structural determinants that characterize caspase substrate specificity

As indicated in previous studies, the three-dimensional (3D) context and the appropriate presentation of solvent accessible surface are key factors, which determine whether the presence of a particular substrate motif can be accessed and cleaved by a caspase (Nicholson, 1999; Timmer et al., 2009). The expanded dataset of caspase substrates used in this study allows us to perform a comprehensive analysis of the structural determinants that characterize the caspase substrate specificity. As such, we performed the prediction of three secondary-structures (helix, strand and coil), two-state solvent accessibility (exposed, buried), native disorder (disordered, ordered), as well as functional domains, using the PSIPRED (Jones, 1999), SCRATCH (Cheng et al., 2005), DISOPRED (Ward et al., 2004) and HMMER (Finn et al., 2008) programs, respectively. All of these programs have been widely accepted as the state-of-the-art in their respective functions.

The frequency of secondary-structure types occurring at each position from P3 to P′ reveals that caspase most frequently cleave substrates that contain coils or loops, which is consistent with a recent study of large-scale proteomics-based profiling of caspase-mediated proteolytic events (Malhus et al., 2008). Depending on positions P3 through P′, the majority of these cleavage sites (71–80%) are observed to be located within the predicted coils, 2–9% in beta sheets and 17–19% are located within alpha-helices (71–80%) are observed to be located within the predicted coils, 2–9% in beta sheets and 17–19% are located within alpha-helices (71–80%) are observed to be located within the predicted coils, 2–9% in beta sheets and 17–19% are located within alpha-helices. To better characterize caspase substrate cleavage sites relative to functional domain boundaries annotated in Pfam (Finn et al., 2008) indicates that 35.7% of these cleavage sites are located within functional domains, 32.8% are located between functional domains, 14.0% are located before the first domain, and 17.5% are located after the last domain (Fig. 3D).

3.3 Prediction of caspase cleavage sites using the BEA profile

In this section, we focused on predicting caspase cleavage sites from the flanking amino acid sequences of substrates. This is different from previous studies where statistical rules based on amino acid preference in the vicinity of cleavage sites (Backes et al., 2005; Garay-Malpartida et al., 2005; Lohmüller et al., 2003) or classification-oriented algorithms like SVM (West et al., 2006, 2007) or neural networks (Yang, 2005) are typically used to generate predictive models. Here, we formulated the prediction task of cleavage sites as a regression task for which SVR was utilized to build the models which predict the cleavage probabilities given primary sequences. First, we wanted to examine the influence of different local window sizes of single sequence inputs on the predictive performances of the SVR models. We thus used different local window sizes, both symmetric and asymmetric (Supplementary Fig. 4), to build the SVR models in order to find out which size could lead to the best performance. The accuracy of the 5-fold cross-validation is shown in Supplementary Table 2.

The predictive performance based on asymmetric windows appears to be better than the corresponding symmetric windows (Supplementary Table 2). For example, the prediction accuracy based on P3–P′ (80.8%) is better than that based on P3–P′ (79.6%). Another interesting finding is that the non-prime-side cleavage sites (P1–P′) appear to have more diverse effects on the prediction performance than the prime-side sites (P′–P′) in that smaller window sizes are required to improve the predictive performance for the latter. Based on a local window of P3–P′, the SVR method can achieve the best prediction accuracy of 82.2% and an MCC of 0.667. Therefore, in the following analysis, we fixed the local window at P3–P′ when evaluating the effects of incorporating other informative features.

3.4 Improving predictive performance by incorporating relevant structural information of substrates

The BEA profile used above is based on the binary encoding of single sequences only; however, we need to take into account additional informative features to further improve the predictive performance. Recently Shao et al. (2009) proposed a novel approach called bi-profile Bayesian feature extraction to extract the key features in the positive and negative data. This method outperformed other methods due to its advantage over binary encoding and single feature encoding schemes. On the other hand, caspase cleavage sites exhibit different features in terms of predicted secondary-structure, solvent accessibility and natively unstructured regions, which might be suitably captured by bi-profile Bayesian feature extraction approach. It is conceivable that incorporating this information might be useful for improving the prediction accuracy of caspase

![Sequence logo diagram representation of the occurrences of amino acid residues in the caspase substrate cleavage site P8–P′ positions. The occurrence rate of each amino acid type, the sequence logo ordinates have been scaled in bits (Schneider and Stephens, 1990).](https://academic.oup.com/bioinformatics/article-abstract/26/6/752/245474/sequence-logo-diagram-representation-of-the-occurrences-of-amino-acid-residues-in-the-caspase-substrate-cleavage-site-p8-p-positions-the-occurrence-rate-of-each-amino-acid-type-the-sequence-logo-ordinates-have)
Fig. 3. Structural determinants of caspase substrate specificity based on the amino acid occurrences in P₈–P'₈ positions for cleavage sites. (A) Three secondary-structure assignments at each position for P₈–P'₈ positions of caspase substrate cleavage sites. H: alpha-helix, E: beta-stand, C: coil. (B) Two-state solvent accessibility distribution at each position for P₈–P'₈ positions of caspase substrate cleavage sites. Two states are defined: ‘e’ indicates ‘exposed’, while ‘b’ denotes ‘buried’. (C) Natively unstructured region or disordered region distribution at each position for P₈–P'₈ positions of caspase substrate cleavage sites. ‘Asterisk’ indicates the natively unstructured/disordered region while ‘dot’ denotes the structured/ordered region. (D) Localization of caspase substrate cleavage sites relative to functional domain boundaries annotated in Pfam. Four domain boundaries are defined: inter-, intra-domain, before first domain and after last domain.

cleavage sites. From this, we extracted different types of sequence profiles based on the bi-profile Bayesian feature extraction: BPBAA, BPBSS, BPBBSA and BPBBDISO, as described in the ‘Materials and methods’ section.

We evaluated the predictive performances of different combinations of these profiles with the gradual increase in the complexity of features, i.e. the encoding schemes BEAA, BEAA_BPBAA, BEAA_BBPBSA, BEAA_BPBBSA, BEAA_BPBBSA_BPBBDISO, BEAA_BPBAA_BBPBSA_BBPBDISO and BEAA_BPBAA_BBPBSA_BBPBDISO (for brevity, these encoding schemes are represented by 1, 2, ..., 8, respectively, in Table 1). This step-wise procedure can reveal the contribution of individual features to the predictive performances.

When selecting the sequence encoding scheme BEAA_BPBAA which uses single binary amino acid profile coupled with bi-profile Bayesian signature, Cascleave achieved an accuracy of 84.1%, F-score of 74.2%, and MCC of 0.686. Moreover, the incorporation of the bi-profile Bayesian features based on binary encoding amino acid profile and solvent accessibility (BEAA_BPBAA_BBPBSA) leads to a further improvement of 1.9% accuracy. The F-score and MCC also increased by 4.4% and 0.031, respectively.

It is observed that the predictive performance of Cascleave gradually increases with the addition of input features, and attains the best performance when using the sequence encoding scheme BEAA_BPBAA_BBPBSA_BBPBDISO—achieving a prediction accuracy of 87.6%, F-score of 80.4% and MCC of 0.747 and was the best performing encoding scheme observed in this study. However, the addition of BPBSS did not improve the predictive performance, but rather decreases the accuracy, presumably because that there might exist redundant information between different feature sets.

To further evaluate the predictive performance, we plotted the receiver operating characteristic (ROC) curves for the assessment of these eight sequence encoding schemes in Figure 4. ROC curves indicate the performance of all the SVR models based on the same training and testing datasets. The uppermost curve with the largest area under the curve indicates the best prediction model. The ROC

756
Table 1. Predictive performances of Cascleave using different sequence encoding schemes

<table>
<thead>
<tr>
<th>Sequence encoding schemes</th>
<th>Prediction accuracy (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC SE SP F-score MCC Dimensionality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>82.2 65.2 92.0 72.8 0.667 120</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>84.1 67.6 92.5 74.2 0.686 132</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>84.6 71.4 91.3 75.7 0.699 132</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>86.5 73.7 93.1 78.6 0.730 144</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>86.3 76.5 91.2 79.0 0.731 144</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>87.1 74.6 93.4 79.5 0.739 168</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>87.3 72.6 94.8 79.5 0.741 156</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>87.6 76.2 93.3 80.4 0.747 156</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results were obtained by 5-fold cross-validation.

Fig. 4. The ROC curves to assess the predictive performance of the eight Cascleave models based on different sequence encoding schemes. See the main text for more details about these encoding schemes. The curve based on the scheme BEAA_BPBAABPBSABPBDISO dominates the curves representing other encoding schemes. This reinforces the notion that this is the best amongst all eight encoding schemes. Taken together, all the results obtained from Table 1, Figure 4 and Supplementary Table 5 suggest that the integration of relevant structural features in terms of predicted secondary structure, solvent accessibility and natively unstructured regions based on bi-profile Bayesian feature extraction approach can significantly enhance the predictive performance of Cascleave.

3.5 Comparison with other methods

Next, we compared the predictive performance of our Cascleave predictor with other methods. Due to the unavailability of the two previously developed web servers CasPredictor (Garay-Malpartida et al., 2005) and GraBCas (Backes et al., 2005), we compared the predictive performance between Cascleave, CASVM (Wee et al., 2007) and Multi-factor CASVM (Wee et al., 2009).

Multi-factor CASVM is a two-step model: the first step is based on CASVM and the second step filters out the predicted false positives using the structural factors such as disorder and solvent exposure in the vicinity of cleavage sites (Wee et al., 2009). The methodological differences among these three methods are detailed in Supplementary Table 5.

Since cross-validation performance comparison is only logical when the training and testing datasets being employed are identical to each other, these two methods were first tested on the same training and testing datasets using 5-fold cross-validation. For each method, there were a total of six models indexed to a particular local window size. As previously observed by Wee et al. (2005), because of the large percentage of cleavage sites with the canonical ‘XXXD’ motif compared to the much smaller percentage of cleavage sites with the non-canonical ‘XXXE’ and ‘XXXG’ motifs, there might exist the possibility for the built models to be over-trained, meaning the built model tends to overlook the other non-canonical cleavage sites and simply predict them as being negatives. Cascleave generally results in an improvement of ~4–5% and 2–5% with respect to CASVM and Multi-factor CASVM, respectively (Supplementary Table 4). Cascleave obtained the best performance with an accuracy of 87.4%, F-score of 80.3% and MCC of 0.747.

Further, we compared the predictive performance of Cascleave, CASVM and multi-factor CASVM using the LOOCV test, which is a more stringent test compared with 5-fold cross-validation (Supplementary Table 5). For the sake of a comprehensible comparison, the predictive performances of different Cascleave models based on different encoding schemes were also presented. As can be seen, Cascleave model using the encoding scheme BEAA_BPBAABPBSABPBDISO achieved the best prediction performance in the LOOCV test, with the accuracy of 89.0 ± 2.8% and the MCC of 0.742 ± 0.062, respectively, while multi-factor CASVM achieved the accuracy of 88.3 ± 2.8% and the MCC of 0.715 ± 0.072, respectively. Moreover, we tested the predictive power of the online Cascleave server to recognize novel caspase substrates and compared with the online CASVM server using an independent testing set. It was extracted from a recent study of Dix et al. (2008) and contains newly identified caspase substrate cleavage sites that were not previously reported. The percentage of true positives, i.e. the percentage of the cleavage sites that were correctly predicted by Cascleave is 81.2%, while CASVM was only able to identify 66.7% of the novel substrate cleavage sites, again demonstrating the predictive power of Cascleave (Supplementary Table 6).

In summary, Cascleave outperforms CASVM and multi-factor CASVM by explicitly integrating primary sequence features with the predicted solvent accessibility/secondary structures/native disorder features, which serve as an important supplement to the primary sequence. By integrating these features, Cascleave is capable of distinguishing more difficult cleavage sites that cannot be readily detected by methods based only on primary sequence information.

3.6 Case study

We further investigated the predictive performance of Cascleave by studying four different caspase substrates for which the cleavage sites have been experimentally validated. Substrate sequence scanning results with Cascleave are shown in Figure 5 and...
J. Song et al.

Fig. 5. The predicted cleavage probability for caspase cleavage sites using Cascleave based on the best sequence encoding scheme. (A) Caspase-activated DNase inhibitor (ICAD, Uniprot ID: O54786); (B) Apoptotic protease Mch-2 (Caspase-6, Uniprot ID: P55212). The two substrates have two and three cleavage sites, respectively. The predicted coiled, solvent exposed and natively disordered regions on the top of each panel are highlighted by magenta, green and red, respectively. A threshold value of 0.5 for making the positive cleavage site prediction is denoted by a red dashed line. The predicted cleavage sites in P_{-1}^{p+4} positions by Cascleave are also labeled.

Supplementary Figure 5. The first example is the caspase-activated DNase inhibitor (ICAD) which inhibits the activity of Caspase-activated DNase (CAD) (Enari et al., 1998). Importantly, cleavage of ICAD by caspase-3 activates the CAD nuclease and deactivates its CAD-inhibitory effect. Site-directed mutagenesis indicated that ICAD carries two specific cleavage sites by caspase-3: DEPD|SRAG at positions 117–118 and DAVD|TGVG at positions 224–225 (| denotes the cleavage site). Cleavage of ICAD can liberate the active CAD nuclease that can otherwise be inhibited by ICAD and mediate the apoptotic DNA degradation (Sakahira et al., 1998).

The second example is the apoptotic protease Mch-2 (caspase-6), which is further cleaved by caspase-3, -8 or -10 to produce the two active subunits (Srinivasula et al., 1996), resulting in the activation cascade of caspases responsible for apoptosis execution. Mch-2 has three cleavage sites of caspase-3: TETD|AFYK at positions 23–24, DVVD|NQTE at position 179–180 and TEVD|AASV at positions 193–194 (Srinivasula et al., 1996). All the experimentally identified cleavage sites of both ICAD and Mch-2 were successfully predicted with the Cascleave predictor based on the best-performing sequence encoding scheme (Fig. 5). These cleavage sites were among the top ranking results according to the predicted probabilities and could be correctly identified as putative cleavage sites using a cutoff threshold of 0.5.

The third and fourth examples are the heterogeneous nuclear ribonucleoprotein inhibitor (hnRNP, Uniprot ID: O43390) (Brockstedt et al., 1998) and the Ras GTPase-activating protein (RasGAP, Uniprot ID: P20936) (Yang and Widmann, 2001), respectively. They represent difficult caspase substrates for which the cleavage sites cannot be readily deduced from the use of sensitive machine learning models like Cascleave. hnRNP has four caspase cleavage sites, however, Cascleave only successfully predicted two of them: DYYD|DYYG and DYHD|YRGG, failing to identify the other two cleavage sites: RAID|ALRE and KESD|LSHV (Supplementary Fig. 5). RasGAP is a special substrate as it is cleaved through sequential caspase cleavage during apoptosis (Yang and Widmann, 2001). Assigning a higher cutoff threshold of 0.6, Cascleave was only able to predict one of the two cleavage sites DEGD|SLDG while missing the primary cleavage site DTVD|GKEI. Nevertheless, if a lower cutoff of 0.5 is used, DTVD|GKEI site will be included in the predicted sites, with the ranking of third place, among the five predicted sites. All the above results suggest that silico computational sequence scanning using...
The predictive performance of Cascleave was further showcased by the development of therapeutics that target specific caspase-mediated substrates; however, this information cannot at present be readily derived from experimental approaches. Solving this problem is fundamental for both understanding caspase biology and the derived from experimental approaches. Solving this problem is fundamental for both understanding caspase biology and the development of therapeutics that target specific caspase-mediated apoptotic pathways.

The recent proliferation of large-scale in vitro peptide cleavage libraries and proteome-wide global profiling platforms offers promising prospects of identifying specific caspase substrates in combination with computer-based screening of genome sequences (Timmer et al., 2009). The advantage of machine learning techniques such as SVR make them particularly appealing in solving the difficult problem of caspase substrate specificity prediction, as they can be effectively applied to better describe the complex non-linear relationships underlying the caspase-substrate interactivity by using the kernel functions to build the predictive models (Shao et al., 2009). To address this problem, we developed a novel bioinformatic approach based on SVR to make testable predictions on the substrate specificity of caspases. Considering that caspase cleavage sites show preferences for predicted secondary structure, solvent accessibility and native unstructured regions, we used a bi-profile Bayesian feature extraction approach to derive these profiles and train the SVR models. The efficiency of the resulting Cascleave predictors for predicting caspase cleavage sites has been demonstrated by comparing to one of the most accurate existing algorithms CASVM (Woo et al., 2006, 2007) and multi-factor CASVM (Woo et al., 2009). The predictive performance of Cascleave was further showcased by predicting four caspase substrates for which the cleavage sites have been experimentally validated. The results indicate that sequence scanning using Cascleave should be very useful for identifying the putative caspase cleavage sites.

In this study, our goal was to predict all the potential cleavage sites, irrespective of the spatiotemporal environments or conformation changes that a substrate may be subject to. The features used as input to the Cascleave predictor are derived from primary sequences and the input data does not contain specific information about the order of sequential cleavage events in proteolytic cascades. In these exceptional cases, there is a chance that the predictor will fail as a purely statistical model. The predicted solvent exposure and native disorder features are a supplement to the primary sequence, and in cases where the primary sequence alone contains a strong consensus indicating a cleavage site, we hope these amino acid sequence encoding features are overriding. For example, in RasGAP (Uniprot ID: P20936) there is a correctly predicted cleavage site (DEGD/SLDG) by Cascleave which shows low probability of being solvent exposure and natively disordered (Supplementary Fig. 5). In the specific cases where a sequential proteolytic cascade occurs, features based on primary sequence alone are unlikely to be sufficient. Training based on features from the 3D structure of the substrate protein may help in this regards, although this is not amenable to a general predictor since the 3D structures of many substrates, in their intact or cleaved-form mid-cascades, are generally not known.

There are a number of additional measures with potential to further increase the accuracy of prediction in the future. The first approach is to use either the accurate 3D structure of the substrate or the structure of a protease in complex with a substrate, instead of the predicted structural information to build the predictive models. This will allow us to ascertain whether the high-resolution crystal structure of a protease/substrate complex can be used to derive specificity information. The second is to incorporate other informative and complementary features, such as sequence-order-dependent context that can better describe the sequential neighborhood surrounding the substrate cleavage sites (Mahrous et al., 2008; Nicholson, 1999). The third is to investigate how to effectively represent the negatives (non-cleavage sites) that are true negatives and are non-cleavable under any cellular or physiological conditions, and how to better discriminate these absolute negatives from those that are cleavable under a given physiological condition. Thus the precise detection of the true negatives is more likely to contribute to an improvement of the prediction accuracy, based on which characteristic feature sets regarding positives and negative samples can be more accurately represented and established. Further improvement can also be achieved by using refined training and testing datasets with high-quality coverage of the cleavage sites, identified by high-throughput proteome-wide techniques.

5 CONCLUSION

In summary, we have proposed a novel approach to predict caspase cleavage sites from the flanking amino acid sequences of caspase substrates. We analyzed the structural determinants of caspase cleavage sites from the flanking amino acid sequences and the input data does not contain specific information about the order of sequential cleavage events in proteolytic cascades. In these exceptional cases, there is a chance that the predictor will fail as a purely statistical model. The predicted solvent exposure and native disorder features are a supplement to the primary sequence, and in cases where the primary sequence alone contains a strong consensus indicating a cleavage site, we hope these amino acid sequence encoding features are overriding. For example, in RasGAP (Uniprot ID: P20936) there is a correctly predicted cleavage site (DEGD/SLDG) by Cascleave which shows low probability of being solvent exposure and natively disordered (Supplementary Fig. 5). In the specific cases where a sequential proteolytic cascade occurs, features based on primary sequence alone are unlikely to be sufficient. Training based on features from the 3D structure of the substrate protein may help in this regards, although this is not amenable to a general predictor since the 3D structures of many substrates, in their intact or cleaved-form mid-cascades, are generally not known.

There are a number of additional measures with potential to further increase the accuracy of prediction in the future. The first approach is to use either the accurate 3D structure of the substrate or the structure of a protease in complex with a substrate, instead of the predicted structural information to build the predictive models. This will allow us to ascertain whether the high-resolution crystal structure of a protease/substrate complex can be used to derive specificity information. The second is to incorporate other informative and complementary features, such as sequence-order-dependent context that can better describe the sequential neighborhood surrounding the substrate cleavage sites (Mahrous et al., 2008; Nicholson, 1999). The third is to investigate how to effectively represent the negatives (non-cleavage sites) that are true negatives and are non-cleavable under any cellular or physiological conditions, and how to better discriminate these absolute negatives from those that are cleavable under a given physiological condition. Thus the precise detection of the true negatives is more likely to contribute to an improvement of the prediction accuracy, based on which characteristic feature sets regarding positives and negative samples can be more accurately represented and established. Further improvement can also be achieved by using refined training and testing datasets with high-quality coverage of the cleavage sites, identified by high-throughput proteome-wide techniques.
Natural Science Foundation of China (60704047), Science and Technology Commission of Shanghai Municipality (08ZR1410600, 08JC1410600), and sponsored by Shanghai Pujiang Program and Innovation Program of Shanghai Municipal Education Commission (10ZZ17). J.W. is supported by the Australian Research Council (DP0772238). J.C.W. is a NHMRC Principal Research Fellow and an ARC Federation Fellow.

Conflict of Interest: none declared.

REFERENCES


Liu,W. et al. (2006) Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. BMC Bioinformatics, 7, 182.


