BIOINFORMATIC APPROACHES FOR PREDICTING SUBSTRATES OF PROTEASES

JIANGNING SONG∗,+†‡§,‡‡, HAO TAN†, SARAH E. BOYD§, HONGBIN SHEN§, KHALID MAHMOOD†∗∗, GEOFFREY I. WEBB††, TATSUYA AKUTSU‡, JAMES C. WHISSTOCK∗,+,†,§§ and ROBERT N. PIKE∗,+,¶¶

†Department of Biochemistry and Molecular Biology
Monash University, Victoria 3800, Australia
‡Bioinformatics Center, Institute for Chemical Research
Kyoto University, Kyoto 611-0011, Japan
§Tianjin Institute of Industrial Biotechnology
Chinese Academy of Sciences
Tianjin 300308, China
∗AgriBio, La Trobe University
Victoria 3086, Australia
¶¶Institute of Image Processing and Pattern Recognition
Shanghai Jiaotong University, Shanghai 200240, China
**ARC Centre of Excellence for Structural and Functional Microbial Genomics, Monash University
Victoria 3800, Australia
††Faculty of Information Technology, Monash University
Victoria 3800, Australia
‡‡Jiangning.Song@monash.edu
§§James.Whisstock@monash.edu
¶¶Rob.Pike@monash.edu

Received 15 July 2010
Revised 8 October 2010
Accepted 9 October 2010

Proteases have central roles in “life and death” processes due to their important ability to catalytically hydrolyze protein substrates, usually altering the function and/or activity of the target in the process. Knowledge of the substrate specificity of a protease should, in theory, dramatically improve the ability to predict target protein substrates. However, experimental identification and characterization of protease substrates is often difficult and time-consuming. Thus solving the “substrate identification” problem is fundamental to both understanding protease biology and the development of therapeutics that target specific protease-regulated pathways. In this context, bioinformatic prediction of protease substrates may provide useful and experimentally testable information about novel potential cleavage sites in candidate substrates. In this article, we provide an overview of recent advances in developing bioinformatic approaches for predicting

∗Corresponding authors.
protease substrate cleavage sites and identifying novel putative substrates. We discuss the advantages and drawbacks of the current methods and detail how more accurate models can be built by deriving multiple sequence and structural features of substrates. We also provide some suggestions about how future studies might further improve the accuracy of protease substrate specificity prediction.

**Keywords**: Proteases; substrate specificity; substrate cleavage site; bioinformatics; sequence analysis; machine learning; support vector machine; feature selection; structural information.

### 1. Introduction

Proteases, also known as proteolytic enzymes, peptidases or proteinases, are enzymes that catalyze the breakdown of proteins by the hydrolysis of peptide bonds in their substrates.\(^1\)–\(^5\) This process is referred to as proteolysis or substrate cleavage and is used as a biological switch to activate/deactivate protein function in numerous biological processes.\(^4\) Thus the ability to catalytically hydrolyze protein substrates, along with the removal of damaged or undesirable products generated after protein synthesis, is fundamental to all forms of life.\(^2\)–\(^5\) Indeed, controlled proteolysis is one of the major pathways by which the estimated 1–1.5 million peptides and proteins are produced from the \(\sim 26,000\) human genes, in order to fulfill the complexity of human life.\(^1\) Proteases thus have central roles in “life and death” processes, such as neural, endocrine and cardiovascular signalling, digestion, degradation of misfolded or unwanted proteins, immunity, cell division and apoptosis. The importance of protease-controlled protein regulation is apparent in numerous human pathological conditions related to alterations in protease activity levels, including cancer, arthritis, inflammation, neurodegenerative and cardiovascular diseases, development, infection and immunity.\(^2\)–\(^4\) Their fundamental roles in multiple biological processes and their implication in numerous pathological processes resulting from common biochemical functions have led to the concept of the protease degradome,\(^5\) which is defined as the complete set of proteases in an organism.

The key to understanding the physiological role of a protease is to identify the repertoire of its natural substrate(s), known as the substrate degradome.\(^5\) Proteases act as processing enzymes that carry out either highly or moderately selective cleavage of the scissile bond within the cleavage site of their substrates. The specificity and thereby role of the enzymes vary, primarily depending on their active sites, which display selectivity ranging from preferences for a number of specific amino acids at defined positions (e.g. thrombin and the caspases) to more generic sites with limited discrimination at one position (e.g. chymotrypsin).\(^6\)\(^7\) In addition to the primary amino acid sequence of the substrate, specificity is also influenced by the three-dimensional conformation of the substrate (secondary and tertiary structures). In particular, proteases preferentially cleave substrates within extended loop regions,\(^8\)–\(^11\) while residues that are buried within the interior of the protein substrate are clearly inaccessible to the protease active site. Finally, cleavage is regulated by the temporal and physical co-location of the protease and substrate.
In particular, some proteases are sequestered within specific compartments, with limited access to proteins, while others are able to cleave multiple substrates in different physiological compartments.

Knowledge of the substrate specificity of a protease can dramatically improve our ability to predict target protein substrates; however, this information can at present only be derived from experimental approaches. In recent years, many specificity-profiling experimental techniques have been developed to globally profile protease–substrate relationships and characterize the substrate repertoire of proteases. These include phage display,\textsuperscript{12,13} peptide libraries,\textsuperscript{14,15} PROTOMAP (SDS-PAGE integrated protein topography and Migration Analysis platform),\textsuperscript{16} and, more recently, proteomic technologies that use chemical labelling strategies to distinguish N-termini of peptides that are newly formed from proteolytic events such as iTRAQ,\textsuperscript{17,18} or iTRAQ-TAILS (iTRAQ with terminal amine isotopic labelling of substrates),\textsuperscript{19–21} as well as other modified N-terminal labelling or positional proteomics approaches.\textsuperscript{22–26} Despite the fact that these global profiling techniques have greatly increased the coverage of uncharacterized protease substrates, they are complex, costly and time-consuming. However, in the absence of such data, the targets of protease function cannot be deduced \textit{a priori} from the structure or sequence of the protease. In-depth understanding of the substrate specificity of proteases is essential for obtaining deep insight into their function, and is a prerequisite for the development of specific inhibitors.\textsuperscript{11,27–29} Thus, solving the “substrate identification” problem is fundamental both to understanding protease biology and to the development of therapeutics that target specific protease-regulated pathways.\textsuperscript{2}

In this context, bioinformatic prediction of the substrates of proteases can provide important and experimentally testable information with regard to the discovery of novel substrates and the putative cleavage sites for a protease.

In the past few decades, a number of computational approaches have been developed to predict proteases and their substrates based on different algorithms, including prediction of protease family (serine, cysteine, aspartic, glutamic, metallo or threonine),\textsuperscript{30,31} development of general predictors to model substrate specificity,\textsuperscript{32–36} and prediction of the substrate cleavage sites for various proteases, for example, caspases,\textsuperscript{37–43} granzyme-B,\textsuperscript{44,45} thrombin,\textsuperscript{46–50} Factor Xa,\textsuperscript{51,52} trypsin,\textsuperscript{53–55} HIV-protease,\textsuperscript{56–61} SARS-CoV\textsuperscript{62} and C virus proteases.\textsuperscript{63,64} These methods generally take known substrate sequences as input, and the resulting models can predict substrate cleavage sites with accuracy ranging from 70% to over 90%. This indicates that these sequence-based substrate specificity tools can be helpful in identifying novel physiological substrates and their cleavage sites \textit{in vivo}.

In this review, we summarize recent progress in the development of bioinformatic approaches to predict protease substrate specificity including benchmark datasets, feature extraction, prediction algorithms, web servers and software implementations. We focus on the methods that are publicly available and highlight some of the major results and challenges in developing more accurate approaches for predicting substrate specificity. We will discuss the shortcomings and drawbacks of current methods and raise issues that remain to be addressed in future studies. We will
also illustrate the utility of the methods by presenting several case studies for the identification of putative substrate cleavage sites and discuss possible future directions for the current methods and their utility for identifying novel putative substrates.

2. Biochemistry and Nomenclature of the Substrate Specificity of Protease

Proteases specifically cleave protein substrates near the N- or C-terminus (termed *exopeptidases*), or in the middle of the substrate (termed *endopeptidases*), through the binding of the protease active site to the substrate residues flanking the cleavage site (Fig. 1). As defined by Schechter and Berger, the active site residues in the protease are composed of contiguous pockets termed *subsites*. Each subsite pocket binds to a corresponding residue in the substrate sequence, referred to here as the sequence *position*. According to this definition, amino acid residues in the substrate sequence are consecutively numbered outward from the cleavage site as $\cdots$-P$_{1}$-P$_{3}$-P$_{2}$-P$_{1}$-P$_{1}'$-P$_{3}'$-P$_{2}'$-P$_{1}''$$\cdots$ (the scissile bond is located between the P$_{1}$ and P$_{1}'$ positions), while the subsites in the active site are correspondingly labelled as $\cdots$-S$_{4}$-S$_{3}$-S$_{2}$-S$_{1}$-S$_{1}'$-S$_{2}'$-S$_{3}'$-S$_{4}'$$\cdots$.

Proteases bind to their substrates through interactions between the subsites in the active site cleft, and the amino acids within the cleavage site. However, the subsites exhibit varying binding affinities for the amino acids in the substrate, ranging from a restricted one or few specific amino acids, to generic binding with little or no discrimination between different amino acids. For instance, thrombin is

![Diagram of protease substrate specificity](image-url)
a serine protease with a classical role in the blood coagulation system, acting as a pro-coagulant enzyme by converting fibrinogen into insoluble strands of fibrin in the final stages of blood coagulation, as well as catalyzing many other coagulation and complement-related reactions. Thrombin possesses a trypsin-like specificity, as it preferentially cleaves after arginine (R) residues and, under some circumstances, after lysine (K) residues, although with a lower efficiency. However, in addition to the requirement of P₃ R or K, other amino acids from the P₈₋P₈′ positions confer important additional determinants of specificity that can affect the substrate cleavage efficiency, particularly across the P₄₋P₄′ positions (Fig. 2).

In order for a substrate to be cleaved, the subsites usually need to bind to the substrate residues in a “cooperative” or “concerted” manner, which means that the binding process of each amino acid residue in the substrate to its corresponding subsite is not entirely independent, a phenomenon referred to as subsite cooperativity. This has been observed in several proteases, such as thrombin, Factor Xa, trypsin, HIV-1 and other viral proteases, where researchers have identified a number of cooperative subsites whose interactions can have either positive or negative influences on cleavage efficiency. For example, in double mutants of thrombin, the concerted interactions between subsites were abolished, resulting in a great loss of specificity and the inability to cleave substrates. This synergistic effect illustrates

Fig. 2. Thrombin substrate cleavage sites for P₈₋P₈′ positions. The results are calculated from the experimentally determined substrate cleavage sites of thrombin published in the MEROPS database. The amino acid occurrences from P₈₋P₈′ are calculated and displayed in the form of a two-dimensional heat map. The scissile bond between P₄ and P₄′ is indicated by a vertical white line.
the important role of cooperativity between two or more subsites, a subject that has recently been reviewed.6

Protease specificity is not only influenced by substrate sequence, but also substrate conformation.69,70 This includes the compatibility between the shape of the active site cleft and the structure of the amino acid side chains within the cleavage site,35 and the presentation of the cleavage site at the surface of the three-dimensional structure of the substrate.42,45 Furthermore, the specificity of many proteases can be dramatically affected by the existence and occupancy of one or more exosites on the protease, i.e. binding sites that are remote from the active site, but which regulate the function of the protease.50,71 The presence of these multiple influences on substrate cleavage, and the complexity of enzyme-substrate recognition, combined with the fundamental importance of characterizing the substrate degradome, make it highly desirable to develop powerful bioinformatic approaches to accurately predict the substrates of proteases. Publicly available approaches to this problem are reviewed here.

3. Methodology Development

3.1. Benchmark dataset construction

An essential component of substrate prediction is the construction of high-quality benchmark datasets, which ideally contain experimentally verified protease substrates. These benchmark sets are used for the model optimization and performance evaluation of the developed approaches. The most comprehensive information resource is the MEROPS database,1,72 which contains all identified proteases, together with a curated collection of their known substrates and inhibitors. In addition to MEROPS, there are other curated databases providing detailed annotation of proteases and their substrates, such as CutDB,73 a database of proteolytic events for physiologically relevant proteins in vivo or in vitro; PMAP,74 an integrated database for analyzing proteolytic events and pathways; and the Degradome database,75 which assembles all of the relevant mammalian proteases and the relationship between protease alterations and disease. Specific collections for individual proteases or protease families, such as the CASBAH database for caspases, also exist.76 Details of these databases are provided in Table 1.

Using resources such as these, the first step is to collect high-quality data and build a benchmark dataset containing experimentally verified substrate sequences with detailed annotation of the cleavage sites. This dataset is used to build the predictive models and evaluate performance. However, as the starting dataset might contain some homologous sequences that share higher sequence similarity between two or more substrate sequences, programs such as CD-HIT77 need to be run to reduce the overall sequence redundancy in the dataset. The remaining substrate sequences will then be partitioned into training and testing datasets (see Sec. 3.4, “Performance evaluation”).
Table 1. A list of publicly available databases of proteases and their substrates/inhibitors, in alphabetical order. For each database, the area of specialization, current data statistics, and the URL are provided. The statistics were collected as of July, 2010.

| Database | Specialization | Statistics | URL
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CASBAH76</td>
<td>Caspase Proteolytic events</td>
<td>776 substrates</td>
<td><a href="http://bioinf.gen.tcd.ie/casbah/">http://bioinf.gen.tcd.ie/casbah/</a></td>
</tr>
<tr>
<td>CutDB73</td>
<td>Proteolytic events</td>
<td>470 proteases and 3,070 proteolytic events</td>
<td><a href="http://cutdb.burnham.org/">http://cutdb.burnham.org/</a></td>
</tr>
<tr>
<td>Degradome75</td>
<td>Mammalian proteases and proteolytic diseases</td>
<td>570 human, 568 chimpazee, 651 mouse and 641 rat proteases</td>
<td><a href="http://degradome.uniovi.es/">http://degradome.uniovi.es/</a></td>
</tr>
<tr>
<td>MEROPS1,72</td>
<td>Proteases, substrates, inhibitors</td>
<td>160,176 sequences, 2,995 identifiers</td>
<td><a href="http://merops.sanger.ac.uk/">http://merops.sanger.ac.uk/</a></td>
</tr>
<tr>
<td>PMAP74</td>
<td>Proteolytic events and pathways</td>
<td>&gt;45,000 proteases, &gt;5,000 proteolytic events</td>
<td><a href="http://www.proteolysis.org/">http://www.proteolysis.org/</a></td>
</tr>
</tbody>
</table>

3.2. **Sequence feature extraction and representation**

Cleaved and uncleaved peptide sequences are collected as positive and negative training datasets, respectively. A sliding window strategy is commonly employed to extract local sequence features from both the positive and negative data, in which the P$_1$ cleavage site is either symmetrically or non-symmetrically flanked by the upstream and downstream residues (Fig. 3). At the substrate sequence level, sequences information is encoded using a binary encoding scheme. In addition to the local amino acid sequences surrounding the cleavage sites, the predicted structural information in the form of secondary structure, solvent accessibility, and natively unstructured regions can be further incorporated to take into consideration the local structural determinants (Fig. 3). These derived features are respectively summarized below.

![Fig. 3. A sliding window approach to extract the local sequence features from cleaved and uncleaved sequences. A caspase substrate, the apoptotic protease Mch-2 (Uniprot ID: P55212), is used as an example here. Two typical local window sizes are presented: P$_8$-P$_8$ (a symmetrical sliding window), and P$_4$-P$_2$ (a non-symmetrical sliding window). The arrow indicates the substrate cleavage site after the P$_1$ position, while the predicted structural features that are taken into account are highlighted in different colors (C, coil; B, buried; E, exposed; *, disordered; -, ordered). Examples shown here include the predicted secondary structure, solvent accessibility and native disorder.](image-url)
3.2.1. Binary encoding scheme of amino acid sequence

A binary encoding scheme is widely adopted in many applications, where each amino acid residue is represented by the 20-dimensional binary vector, composed of either 0 or 1 elements, whose positional order depends on the alphabetical order of the amino acid type. For example, Ala is represented as \[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]\]. Cys is represented as \[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]\]. Asp is represented as \[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]\]. Tyr is represented as \[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]\]. Therefore, for a peptide sequence with \(L\) amino acid residues, it will be represented by a \((L \times 20)\)-dimensional vector.

3.2.2. PEST sequences

PEST sequences were originally proposed by Rogers et al. to function as proteolytic signals. These sequences contain regions enriched in proline (P), glutamate (E), serine (S) and threonine (T). PEST sequences are hydrophilic stretches of at least 12 amino acids in length, with the entire region flanked by lysine (K), arginine (R) or histidine (H), but not interrupted by positively charged residues. Approximately 10% of the mammalian proteins contain PEST sequences. The presence of PEST sequences in the vicinity of protease cleavage sites might be responsible for target recognition of protein substrates for the rapid proteolytic degradation by proteases such as the ubiquitin-26S proteasome and calpain. Due to the implication of PEST sequences as targets for potential substrates, the PEST-like index has been used in combination with other descriptors to improve the prediction accuracy of cleavage sites by the CaSPredictor, PoPS and SitePrediction programs.

3.2.3. Physicochemical properties

Physicochemical properties of amino acid residues provide useful features for discriminating cleaved sequences from uncleaved sequences. Each amino acid type can be encoded by physicochemical properties extracted from the AAindex database, including hydrophobicity, hydrophilicity, volumes of side chains, polarity, isoelectric point and accessible surface area. The peptide sequence within the sliding window can then be translated into a vector, with each of the amino acids numerically represented by one of the set of descriptors.

3.2.4. Secondary structure

The secondary structure of protein substrates can be predicted using PSIPRED, which is a neural network-based predictor for estimating the probabilities of three secondary structure classes: helix, strand and coil. PSIPRED has been considered as one of the most accurate sequence-based predictors of protein secondary structures with an overall accuracy of up to 80%. In our previous work, predicted secondary structure information has been shown to significantly improve the accuracy of prediction of various properties of proteins, such as cis/trans...
isomerization, residue-wise contact order, disulfide connectivity pattern, half-sphere exposure and residue depth. Recent studies revealed that cleavage sites of some proteases show clear preferences towards specific secondary structure motifs. For example, all-helix and all-loop motifs are commonly cleaved by caspases. Predicted secondary structure has been taken into account by specificity prediction programs, such as PoPS, Cascleave and SitePrediction.

3.2.5. Solvent accessibility

Appropriate presentation of cleavage sites in a solvent exposed region can be important for efficient proteolysis. Two-state (exposed or buried) solvent accessibility is an additional descriptor that can be predicted using the SSpro program implemented in the SCRATCH suite. This program outputs the estimated probability of a residue being solvent-exposed or buried within the substrate structure, which in turn can be used as input to machine learning predictors. Thus, predicted structural information can be exploited to identify substrate sites that are relatively exposed, allowing protease binding and cleavage. However, the fact that some proteolytic cleavages occur at solvent inaccessible regions implies that being predicted as solvent buried does not always rule out the possibility of a site being cleaved. Several methods including PoPS, Cascleave and SitePrediction have exploited predicted solvent accessibility information, either as a filter or direct input to the developed predictors. In cases where 3D structural information is available, PoPS can also use solvent accessibility derived from known structural information in PDB.

3.2.6. Native disorder

The prevalence and significance of natively disordered or intrinsically unstructured proteins has been recognized, as a consequence of large-scale genome sequencing projects and the development of experimental techniques to analyze the structural properties of proteins in solution. Natively disordered proteins have been found to have important roles in many biological processes including transcription, translation, signal transduction, protein recognition, targeted degradation and cell cycle regulation. The natively disordered regions of these proteins have been shown to be essential for their functions. Native disorder of a substrate sequence can be accurately predicted from amino acid sequences using machine learning predictors such as DISOPRED2 and PrDOS. Several studies have shown that the predicted native disorder information can be used to improve the predictive performance.

3.3. Computational methods for predicting the substrate specificity

Predicting protease substrate cleavage sites can be viewed as a binary classification problem. The underlying assumption is that the sequence features encoded in the local sequence surrounding the cleavage sites should contain important information
for determining the substrate specificity. In particular, the combination of multiple sequence and structural features should successfully improve performance in terms of predicting substrate cleavage sites. Here, the discussion will mostly focus on methods for predicting protease substrate cleavage that have been implemented as publicly available web servers or stand-alone software.

Current methods for predicting protease substrates are generally categorized into two main classes: machine learning or empirical scoring function. The former is mainly based on the selection and subsequent representation of useful sequence features and further mapping of input features to the property of being cleaved (positive) or non-cleaved (negative), e.g., BBFNN,\textsuperscript{37} CASVM,\textsuperscript{39,40} Multi-factor CASVM,\textsuperscript{41} Cascleave,\textsuperscript{42} Pripper,\textsuperscript{43} fXaWeb,\textsuperscript{51} and ProtIdent.\textsuperscript{31} The latter relies on learning the underlying rules using the distribution of positive and negative sequences, and building empirical scoring functions to discriminate between the two classes, e.g., PeptideCutter,\textsuperscript{33} CasPredictor,\textsuperscript{38} GraBCas,\textsuperscript{44} PoPS,\textsuperscript{34} HIVcleave,\textsuperscript{61} and SitePrediction.\textsuperscript{36} Table 2 summarizes the features of these methods, and a brief review of their predictive performances is provided in the Results and Discussion section.

3.3.1. Empirical scoring approaches

Empirical scoring approaches rely on learning the underlying rules based on the observed distribution of samples and building empirical scoring functions to discriminate between different classes. In terms of predicting substrate cleavage, most of these prediction methods use experimentally verified substrate data to establish the scoring rules that will determine the prediction quality. Generally, the scoring methods assign a score to each amino acid residue in the cleavage sites and the final score is summarized or averaged within a given local window.

PeptideCutter is an empirical scoring function-based tool used to identify the cleavage sites of protease substrates.\textsuperscript{33} However, many of the cleavage sites used by PeptideCutter were derived from a relatively limited experimental dataset. In addition to proteases, PeptideCutter also provides prediction for specific chemicals that are also used to break down proteins. The success of this tool partly lies in the fact that the proteases and chemicals modeled by this tool have well-defined actions in the breakdown of proteins. Another scoring tool, CasPredictor, predicts caspase cleavage sites by taking into account the amino acid substitution index, and the presence of PEST-like sequences in the vicinity of cleavage sites.\textsuperscript{38} This program achieved a sensitivity value of 81\% (111/137) when predicting cleavage sites for a dataset of 137 experimentally determined cleavages.

Another prediction algorithm, GraBCas, provides a position-specific scoring prediction of potential substrate cleavage sites for caspases and granzyme B.\textsuperscript{44} This algorithm was trained on the relative frequencies of amino acids measured in the study of Thornberry \textit{et al.} who used positional scanning combinatorial libraries to determine the cleavage specificities of caspases and granzyme B.\textsuperscript{95}
Table 2. Publicly available web servers and tools for predicting the substrate specificity of proteases.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Algorithm</th>
<th>Dataset</th>
<th>Feature</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CasPredictor</td>
<td>Empirical scoring</td>
<td>223 substrates, 283 cleavage sites</td>
<td>Fast-like index and position-dependent amino acid matrices</td>
<td><a href="http://icb.usp.br/~farmaco/Jose/">http://icb.usp.br/~farmaco/Jose/</a></td>
</tr>
<tr>
<td>GraBCas</td>
<td>Empirical scoring</td>
<td>47 substrates, 59 cleavage sites</td>
<td>Position-specific scoring matrices</td>
<td>CaS predictor.files</td>
</tr>
<tr>
<td>PoPS</td>
<td>Empirical scoring</td>
<td>From the MEROPS database or user-derived 13 substrates</td>
<td>Position-specific scoring matrices, solvent accessibility, secondary structure</td>
<td><a href="http://pops.csse.monash.edu.au/">http://pops.csse.monash.edu.au/</a></td>
</tr>
<tr>
<td>BBFNN*</td>
<td>Bayesian bio-basis neural network</td>
<td>13 substrates</td>
<td>Local sub-sequences using a sliding window with a fixed size</td>
<td>Not available</td>
</tr>
<tr>
<td>CASVM</td>
<td>Support vector machine</td>
<td>219 cleavage sites</td>
<td>Binary encoding amino acid sequence</td>
<td><a href="http://www.casbase.org/casvm/">http://www.casbase.org/casvm/</a></td>
</tr>
<tr>
<td>Multi-factor CASVM</td>
<td>Support vector machine</td>
<td>74 cleavage sites</td>
<td>Binary encoding amino acid sequence, disordered and solvent exposed propensity</td>
<td><a href="http://www.casbase.org/casvm/">http://www.casbase.org/casvm/</a></td>
</tr>
<tr>
<td>SitePrediction</td>
<td>Empirical scoring</td>
<td>From the MEROPS database</td>
<td>PEST sequences, solvent accessibility and secondary structure</td>
<td><a href="http://www.dmbr.ugent.be/prx/">http://www.dmbr.ugent.be/prx/</a> bioit2-public/SitePrediction/</td>
</tr>
<tr>
<td>Cascleave</td>
<td>Support vector regression</td>
<td>370 substrates, 562 cleavage sites</td>
<td>Predicted secondary structure, solvent accessibility and native disorder based on bi-36 profile Bayesian feature extraction</td>
<td><a href="http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/Cascleave/">http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/Cascleave/</a></td>
</tr>
<tr>
<td>Pripper</td>
<td>Supervised machine learning</td>
<td>358 substrates, 443 cleavage sites</td>
<td>Binary encoding amino acid sequence</td>
<td><a href="http://users.utu.fi/mijopi/Pripper/">http://users.utu.fi/mijopi/Pripper/</a></td>
</tr>
<tr>
<td>Peptide Specifier</td>
<td>Support vector machine</td>
<td>359 granzyme B and 602 caspase substrates</td>
<td>Secondary structure, solvent accessibility and native disorder</td>
<td><a href="http://modbase.compbio.ucsf.edu/peptide/">http://modbase.compbio.ucsf.edu/peptide/</a></td>
</tr>
<tr>
<td>ProtIdent</td>
<td>Two-layer ensemble classifier</td>
<td>3,051 protease sequences</td>
<td>Functional domain and position-specific scoring matrix</td>
<td><a href="http://www.csbio.sjtu.edu.cn/bioinf/Protease/">http://www.csbio.sjtu.edu.cn/bioinf/Protease/</a></td>
</tr>
<tr>
<td>fXaWeb</td>
<td>Bootstrap aggregation algorithm</td>
<td>132 substrate sequences</td>
<td>Six-residue sequence fragments</td>
<td><a href="http://asqa.iis.sinica.edu.tw/fXaWeb/">http://asqa.iis.sinica.edu.tw/fXaWeb/</a></td>
</tr>
</tbody>
</table>

Notes: *BBFNN: Bayesian Bio-basis Functional Neural Network. 

b) fXaWeb: Multi-level bootstrap aggregation predictor for predicting the substrate specificity of factor Xa protease.
PoPS is a comprehensive bioinformatic tool that allows users to build computational models of protease substrate specificity that can be used to predict and rank potential cleavage sites for any protease. In addition, users can also augment PoPS models with dependency rules to account for subsite cooperativity effects that have been experimentally observed to be significant for some proteases such as thrombin, trypsin and HIV protease. Another advantage of PoPS is that it also supports expert users to build new and accurate specificity models from their expert knowledge gleaned from experimental work. Other important complementary information about the solvent accessibility, secondary structure and the presence or absence of PEST sequences can be used within the PoPS prediction engine to further screen the false positive sites.

SitePrediction is a state-of-the-art empirical scoring tool for predicting substrate specificity of any protease. It combines the frequency score, based on the occurrence of each amino acid type at each position in a substrate site with an amino acid substitution matrix score that indicates the similarity of the potential cleavage sites to the known cleavage sites. The final score is calculated as the product of these two scores. Similar to PoPS, SitePrediction also provides extra analysis such as sequence logo representation of the predicted site, identification of PEST regions, and details on solvent accessibility and secondary structure, giving the users a better idea of the likelihood of cleavage of a site of interest. Finally, both PoPS and SitePrediction can be employed on a large scale to search entire proteomes for putative protease substrates and their potential cleavage sites.

3.3.2. Machine learning approaches

Machine learning approaches can automatically learn and recognize complex patterns and use these to construct predictive models to discriminate different classes of samples, especially for those cases in which it is difficult to extract explicit rules. Appropriate representation of sequence or structural features in combination with feature selection, is required to reduce data dimensionality and improve predictive power. An issue for machine learning approaches, however, is how to efficiently avoid or overcome the over-fitting problem. Representative machine learning methods include artificial neural networks, naïve Bayes, decision trees, nearest neighbor, random forest, support vector classification and support vector regression.

The first machine learning algorithm applied to predict caspase substrate cleavage sites is BBFNN (Bayesian Bio-basis functional neural network) which was originally proposed by Yang. Based on 13 substrate sequences with experimentally determined caspase cleavage sites, each caspase sub-sequence was obtained by scanning the protein sequence with a fixed-sized sliding window and input into the BBFNN algorithm to build the classifiers which achieved the highest accuracy of 97.15 ± 1.13% when evaluated on this smaller dataset.

The SVM approach to machine learning has been designed to avoid the over-fitting problem and can be trained efficiently to maximize classification accuracy.
with higher generalization ability.\textsuperscript{96,97} With the increasing availability of experimental data for caspase substrate specificity, Wee \emph{et al}. have recently employed the SVM algorithm with the radial basis function (RBF) kernel for training classifiers.\textsuperscript{39,40} Their SVM-based models (termed as CASVM) extracted local amino acid sequence profiles using a sliding window approach and transformed them into an \(n\)-dimensional vector using an orthonormal encoding scheme. This approach achieved an accuracy ranging from 81.25\% to 97.92\%, when evaluated on a dataset containing 390 caspase substrate sequences.\textsuperscript{39,40} Moreover, Wee \emph{et al}. have further developed the multi-factor CASVM with the first prediction step based on the output from CASVM and the second filtering step based on structural factors such as native disorder and solvent accessibility which filters out false positives to improve predictive performance.\textsuperscript{41}

Song \emph{et al}. developed a new approach to predicting the cleavage sites of caspases from the substrate primary sequences.\textsuperscript{42} This approach, named Cascleave, uses support vector regression (SVR) to build prediction models based on multiple sequence and structural features derived from: (i) the local amino acid sequence profile; (ii) the predicted secondary structure; (iii) the predicted solvent accessibility; and (iv) the predicted native disorder.\textsuperscript{42} These features were extracted and used as input to train the models based on a recently developed bi-profile Bayesian feature extraction method.\textsuperscript{98} The idea of this approach is that the intrinsic difference between two classes of samples can be better reflected and learned in a bi-profile manner by calculating the posterior probabilities of each amino acid at each position from \(P_5\) to \(P_5'\) sites in the training dataset. Thus integrating the bi-profile Bayesian features might be more informative than the general encoding scheme. The researchers investigated the effects of different window sizes on the predictive performance and found that the Cascleave model based on the local window of \(P_4-P_2'\) achieves the best prediction accuracy. Experiments on a dataset of 562 cleavage sites show that Cascleave correctly predicts 82.2\% of the known caspase cleavage sites with a Matthews correlation coefficient (MCC) of 0.667,\textsuperscript{42} performing favorably compared with the CASVM\textsuperscript{39} and multi-factor CASVM.\textsuperscript{41}

Pripper is a recently developed tool for predicting caspase substrate cleavage sites.\textsuperscript{43} It uses three different machine learning-based classifiers (SVM, random forest and J48 algorithm) to make the prediction. Pripper can be used to identify caspase substrates from tandem mass spectrometry based proteomics experiments. More recently, Barkan \emph{et al}. presented another SVM-based tool to computationally recognize the substrate peptides of granzyme B and caspases.\textsuperscript{45} They used sequence and structural properties including the frequencies of amino acid residue types occurring at each position through \(P_4\) to \(P_4'\) sites, physiochemical properties, secondary structure, solvent accessibility and native disorder. This approach is able to predict caspase substrates at an 87\% true positive rate and 13\% false positive rate, and granzyme B substrates at a 79\% true positive rate and a 21\% false positive rate, respectively.\textsuperscript{45} The developed tool has been further applied to scan the human proteome to generate a list of high-confidence candidate substrates.
that were subjected to experimental validation. Two high-confidence predictions, Apoptosis Inducing Factor (AIF-1) and Survival Motor Neuron 1 (SMN1), were experimentally validated as substrates of granzyme B.\(^{45}\)

In addition, there are other predictors specifically developed for predicting the substrate specificity of other proteases based on machine learning techniques. For example, the neural network method for predicting HIV protease cleavage sites,\(^{55,57-59}\) the decision tree method for the SARS-CoV protease,\(^{62}\) as well as the SVM-based fXaWeb predictor for Factor Xa substrate sites.\(^{51}\) However, in the case of modeling the substrate specificity of the HIV-1 protease, Rögnvaldsson and You pointed out that it is inappropriate to apply complex non-linear algorithms, such as neural networks based on a linearly separable database, especially when there is no evidence that the researched problem is non-linear and no serious attempts have been made to rule out the possibility of using simple linear classifiers.\(^{99}\)

Together, the consensus of these studies is that the sequence profiles of protease substrates are very informative for determining the cleavage sites, while structural features have also been useful when real 3D structure is available or even when predicted or modeled structural information is used.\(^{42,45}\)

### 3.4. Performance evaluation

In order to objectively evaluate the predictive performance of any software program, three rigorous tests are regularly utilized: \(n\)-fold cross-validation, leave-one-out cross-validation (or jack-knife test) and independent tests. In the case of \(n\)-fold cross-validation, substrate sequences in the dataset are randomly divided into \(n\) equally sized subsets. In each validation step, one subset is singled out in turn as the test data, while the remaining are used as the training data. This procedure is repeated \(n\) times until each subset has been used both as training and testing data. In the case of leave-one-out cross-validation, each substrate sequence is singled out in turn as the test case, and the remaining as the training data. This needs to be repeated such that each substrate sequence is used once as the testing data. In the case of independent testing, the set of cleavage sites that are used to derive the model are independent from the set of cleavage sites that are used to test the model, with no overlap between the two datasets.

The predictive performances of various specificity predictors are evaluated using the following measures:

1. Sensitivity (percentage of correctly predicted caspase cleavage sites):
   \[
   \text{Sensitivity} = \frac{TP}{TP + FN}
   \]

2. Specificity (percentage of correctly predicted non-cleavage sites):
   \[
   \text{Specificity} = \frac{TN}{TN + FP}
   \]
(3) Accuracy (percentage of correct predictions for both cleavage and non-cleavage sites):

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

(4) Mathew’s Correlation Coefficient (MCC) is a measure of the quality of binary classifications. \(\text{MCC} = 1\) signifies a perfect classification, while \(\text{MCC} = 0\) indicates completely random classification. It is defined as:

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

In each of these measures, \(TP\), \(TN\), \(FP\) and \(FN\) denote the number of true positives, true negatives, false positives and false negatives, respectively.

In addition, the \(F\)-score which is a harmonic mean of precision and recall, is given as:

\[
F\text{-score} = \frac{2 \times TP}{2 \times TP + FP + TN}
\]

4. Results and Discussion

4.1. Sequence and structural determinants of the substrate specificity

As previously discussed, the substrate specificity of proteases can vary considerably, ranging from restricted specificity of one or only a few specific amino acids, to generic specificity without any discrimination between amino acids. Here we focus on two representatives, caspase-3 and thrombin, which are respectively cysteine and serine proteases, to illustrate the sequence and structural determinants of substrate specificity.

The caspases are a family of intracellular cysteine proteases that play essential roles in the initiation and execution of programmed cell death and apoptosis, and are often referred to as “executioner” or “killer” proteases. A hallmark of the substrate specificity of caspases is that they specifically cleave after aspartate (D) residues, although additional requirements must also be met in order for a caspase to efficiently cleave its substrate.

As one of the critical members of the caspase family, caspase-3 is characterized by its recognition for the canonical cleavage site motif “DXXD↓X”, where an aspartate residue (D) occurs in both the \(P_4\) and \(P_1\) positions, “X” denotes any of the 20 amino acids, and “↓” denotes the cleavage site. More specific known cleavage motifs of caspase-3 include “DEVD↓G”, “DGPD↓G”, “DEVD↓N”, “DMQD↓N”, “DEPD↓S”, “DEAD↓G”, “DETD↓S” and “DAVD↓T”.

Using the WebLogo program to analyze caspase-3 substrate specificity data from the MEROPS database, the sequence logo representation of amino acids from \(P_8\) to \(P_8\)' shows that the major determinants for the \(P_1\)' position are small amino acids such as glycine (G), serine (S) and alanine (A) (Fig. 4).
Thrombin is a serine protease from the chymotrypsin family of proteases, and plays a critical role in the blood coagulation system. Thrombin preferentially cleaves substrates that contain an arginine (R) residue at the P₁ position, but can also cleave after a lysine (K) residue, although with far less efficiency. In addition to the P₁ R/K preference, thrombin has also been shown to have a strong preference for proline (P) residues at the P₂ position, while serine (S) and glycine (G) residues at the P₁′ position are also suggested to improve substrate cleavage by thrombin (Fig. 4).

Recent studies have found different preferences for secondary structure types occurring at protease cleavage sites. As indicated by the distribution of secondary structure (SS) motifs for P₄-P₄′ caspase and thrombin cleavage sites (Fig. 5), SS motifs consisting of only loop structures are the most common secondary structure motifs, and are frequently observed in both caspase and thrombin substrates (54% and 22%, respectively). Surprisingly, the helical SS motifs are not uncommon compared to other SS motifs (8% and 2% for caspase and thrombin substrates, respectively), suggesting potential roles for structural dynamics or conformational switching in these regions upon substrate binding and catalytic cleavage by these two proteases. In fact, substrate recognition and catalytic proteolysis by proteases is a complicated multi-step process, that sometimes
Fig. 5. Distribution of secondary structure (SS) motifs of substrate cleavage sites from P$_i$-P$_{i+1}$ for (a) caspase-3 and (b) thrombin. Datasets for both caspase-3 and thrombin substrates were extracted from the MEROPS database. Secondary structure is represented as H: alpha-helix, E: beta-strand, C: coil.

requires dramatic conformational changes of the substrate cleavage site to form the correct conformation for cleavage. The insulin-degrading enzyme (IDE), a Zn$^{2+}$-metalloprotease, serves as a suitable example. On the one hand, IDE needs to transit from a substrate-free open state to a closed state in order to entrap substrate peptides. On the other hand, in order to fit into the catalytic cleft for cleavage, IDE substrates must undergo substantial conformational changes on binding to IDE, in order to fit into the enclosed active site. In particular, sequences derived from an N-terminal loop and an $\alpha$-helix both switch to $\beta$-strand structures prior to degradation. This mechanistic insight might explain why certain regular SS motifs, such as the all-helix SS motif, can be effectively cleaved by proteases.

At the level of solvent accessibility and native disorder, local sequences surrounding cleaved and uncleaved sites also exhibit differences. Enrichment analyses performed by different groups based on solved 3D structures and comparative models, or sequence-based structure prediction, both led to the same conclusion that cleaved sequences are more likely to be exposed to the solvent, flexible and natively disordered compared to uncleaved sequences. Together, it might be useful to use these distinct sequence and structural features that flank the cleaved and uncleaved sites to build more accurate models for predicting the localization of cleavage sites.

4.2. Case studies

In this section, we illustrate how putative protease cleavage sites can be predicted by performing in silico scanning of caspase substrate sequences. The caspase family is a focus of extensive research, and as a result, a number of bioinformatic tools
have been developed for predicting caspase substrates.\textsuperscript{37–45} Prediction of caspase cleavage sites is facilitated by the requirement of Asp (D) at the P\textsubscript{1} position, but is confounded by the varying degree of specificity at other cleavage site positions and across the proteases. As some web servers are not available, we compare predictions from CASVM,\textsuperscript{40} Cascleave,\textsuperscript{42} PoPS,\textsuperscript{34} and SitePrediction\textsuperscript{36} for two known caspase substrates. The first two tools are machine learning-based, while the latter two are empirical rule-based. Except for CASVM, which does not provide a ranking (prediction results given in Supplementary Table 1), the top five ranked predictions from each of the other three servers are shown in Supplementary Tables 2, 3 and 4, respectively.

The two substrates investigated here represent cases for which it is difficult to predict cleavage specificity as both contain sequential cleavage sites during apoptosis. The first example is the heterogeneous nuclear ribonucleoprotein inhibitor (hnRNP) (UniProt\textsuperscript{109} ID: O43390). This protein has four caspase-3 cleavage sites: DYYD|DYYG (P\textsubscript{1} position located at D472 within the protein sequence), KESD|LSHV (D87), DYHD|YRGG (D481) and RAID|ALRE (D66). CASVM predicts seven cleavage sites without ranking and all four physiological cleavage sites are included in the predictions (Supplementary Table 1). The other three methods, Cascleave, PoPS and SitePrediction, give high ranking to the two cleavage sites DYYD|DYYG and DYHD|YRGG, but failed to identify the other two cleavage sites RAID|ALRE and KESD|LSHV. Cascleave gave DYYD|DYYG and DYHD|YRGG a ranking of 1 and 2, respectively (Supplementary Table 2), PoPS gave them a ranking of 5 and 2, respectively (Supplementary Table 3), while SitePrediction gave them a ranking of 2 and 1, respectively (Supplementary Table 4). Interestingly, these two correctly identified cleavage sites are also found to be located in predicted solvent accessible and natively disordered regions [Fig. 6(a)].

The second caspase substrate is the Ras GTPase-activating protein (RasGAP) (UniProt\textsuperscript{109} ID: P20936). It also contains sequential caspase-3 cleavage sites: DTVD|GKEI (D455) and DEGD|SLDG (D157). The former is the primary cleavage site, while the latter is a secondary cleavage site whose cleavage will only occur after the cleavage of the primary site. In the case of RasGAP, CASVM outputs 11 predicted sites where the two cleavage sites are included in the predictions (Supplementary Table 1). PoPS provided the best prediction results compared with Cascleave and SitePrediction, giving a ranking of 1 and 2, respectively (Supplementary Table 3). Cascleave gave a ranking of 1 and 3, respectively (Supplementary Table 2), whereas SitePrediction gave a ranking of 2 and 3, respectively (Supplementary Table 4).

The performance difference between these three methods likely suggests the different characteristics of the information used by different algorithmic frameworks. Combining the predictions of all three methods might be helpful to predict high-confidence caspase substrates that can be subjected to further experimental validation in the laboratory,\textsuperscript{45} all the while being mindful of balancing the improvement of the algorithms with overfitting of the models.
5. Discussion and Conclusions

Due to their important function of catalytically hydrolyzing protein substrates, proteases have central roles in “life and death” processes. Protease-controlled events have been implicated in numerous human diseases.2–5 Solving the “substrate identification” problem is fundamental for both understanding protease biology and the development of therapeutics that target specific protease-regulated pathways. Determining the substrate specificity of proteases is an important step towards deeper understanding of the structure and function of proteases and the underlying mechanisms that govern site-specific proteolysis. Due to the gap between functionally characterized proteins and the proteins for which there is no experimental
functional annotation, there is great demand for powerful bioinformatic methods that can accurately predict protease substrates based on the current available specificity data.

In this review, we summarize recent advances in bioinformatic approaches for predicting the substrate specificity of proteases and their application to identify putative substrates. Existing approaches can be generally classified into two groups: empirical scoring or machine learning based. A recent trend in this field is to recruit various sources of sequence and structural information to further improve the predictive performance of the models, including positional-specific scoring matrices, predicted secondary structure, solvent accessibility and native disorder. Studies have indicated that use of these different but complementary features can lead to improved predictive performance.

Future improvements in the performance of predictive methods may be derived in several ways. The first is to integrate relevant sequence and structural features of cleavage sites based on feature selection. Integrating structural information from the active site cleft of a protease might also be very helpful for developing better predictive models. The second option is to exploit the ensemble learning approach, which provides a better framework for different data integration. This approach uses multiple diverse and independent basic classifiers, and combines their original predictions to make a final, more reliable consensus prediction. The third method is to improve the representation of true “negatives” (sites that cannot be cleaved under any experimental conditions), which is important in order to resolve the open question of how to efficiently reduce false positives, which are often high where there is a highly unbalanced dataset (fewer sparse positive samples against many negative samples). On the other hand, the presence of “false negatives” could further confound the analysis, as in some cases those false negatives can actually be cleaved when the physiological conditions change, meaning that they are actually unidentified or unannotated positives. Finally, with greater availability of specificity data generated by proteome-wide profiling techniques, it is becoming feasible to build high-quality datasets that will allow further development and evaluation of computational methods in order to improve the prediction performance of protease substrate specificity.

Acknowledgments

JS would like to thank the National Health and Medical Research Council of Australia (NHMRC) and the Japan Society for the Promotion of Science (JSPS) for financially supporting this research via the NHMRC Peter Doherty and JSPS Postdoctoral Fellowships. JS was also supported by the Hundreds of Talents Fellowship Program of the Chinese Academy of Sciences (CAS). HS was supported by the National Natural Science Foundation of China (60774047), Science and Technology Commission of Shanghai Municipality (08ZR1410600, 08JC1410600), and sponsored by Shanghai Pujiang Program and Innovation Program of Shanghai
Municipal Education Commission (10ZZ17). GIW is supported by the Australian Research Council (ARC) (DP0772238). JCW is an ARC Federation Fellow and an honorary NHMRC Principal Research Fellow.

References


Jiangning Song received both his B.Eng. degree in Biotechnology and his Ph.D. degree in Bioinformatics from Jiangnan University, China in 2000 and 2005, respectively. From 2005 to 2007, he was a postdoctoral research fellow at the Advanced Computational Modelling Centre (ACMC) of the University of Queensland, Australia. From 2007 to 2009, he was a JSPS Research Fellow at the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. He is currently an NHMRC Peter Doherty Fellow in the Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Australia. His research interests include structural bioinformatics, integrative systems biology, machine learning, data mining and complex networks. Dr. Song is a member of the International Society for Computational Biology (ISCB), the Japan Society of Bioinformatics (JSBi) and the International Proteolysis Society (IPS).

Hao Tan received his B.Eng. degree in Communication Engineering from Southeast University China, and his M.Sc. degree in Applied Information Technology from Monash University, Australia, in 2004 and 2008, respectively. He was employed by China Unicom Co. Ltd. from 2004 to 2008 as a software engineer. He is currently a research assistant in the Faculty of Medicine, Monash University Australia. His research interests are computer software design and bioinformatics applications.

Sarah E. Boyd received a B.Sc. (Hons) degree in Computer Science and Biochemistry (2000) and a Ph.D. degree in Computer Science (2005) from Monash University, Australia. Since 2000, Dr. Boyd has researched computational prediction of protease specificity. In 2006, she worked as a research fellow for the Centre On Proteolytic Pathways, an NIH program grant housed at the Burnham Institute, San Diego, USA. From 2007 to 2009, Dr. Boyd was awarded an Australian Postdoctoral
Fellowship to research cooperativity in protease specificity. In 2010, Dr. Boyd was appointed as a research fellow at La Trobe University, Australia, to research the systems biology of cell signalling. Dr. Boyd is interested in the acceleration of biomedical research and discovery using the systems biology paradigm, which combines experimental and computational research to integrate, model, visualize and exchange large, complex, multi-dimensional datasets.

Hongbin Shen received his Ph.D. degree from Shanghai Jiao-tong University, China in 2007. He was a postdoctoral research fellow of Harvard Medical School from 2007 to 2008. Currently, he is a professor in the Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, China. His research interests include data mining and bioinformatics. He is particularly interested in research aimed at predicting protein structure and function, as well as intelligent modeling for complex biological networks. Dr. Shen has published more than 60 papers and constructed 20 bioinformatics servers in these areas and he serves on the editorial boards of several international journals.

Khalid Mahmood is a Ph.D. student at Monash University, Australia. He is supervised by Prof. James Whisstock and Prof. Geoff Webb. His research interests include computational biology with focus on comparative genomics and protein sequence/structure analysis. In particular, he is interested in developing methods for identifying gene orthologs and conserved gene segments in various genomes. Further, he is interested in developing data mining methods to study how protein families evolve.

Geoffrey I. Webb received his Ph.D. from La Trobe University, Australia, in 1987. He held appointments at Griffith University and then Deakin University, where he received a personal chair. He is a professor in the Faculty of Information Technology at Monash University, where he heads the Centre for Research in Intelligent Systems. His primary research areas are machine learning, data mining, and user modelling. He is known for the development of numerous methods, algorithms and techniques for machine learning, data mining and user modelling. His commercial data mining software, Magnum Opus, incorporates many techniques from his association discovery research. Many of his learning algorithms are included in the widely-used Weka machine learning workbench. He is editor-in-chief of *Data Mining and Knowledge Discovery*.
**Discovery**, co-editor of the *Encyclopedia of Machine Learning* (to be published by Springer), a member of the advisory board of *Statistical Analysis and Data Mining* and a member of the editorial boards of *Machine Learning* and *ACM Transactions on Knowledge Discovery in Data*.

**Tatsuya Akutsu** received both an M.Eng. degree in Aeronautics (1986) and a Dr. Eng. degree in Information Engineering (1989) from the University of Tokyo, Japan.

From 1989 to 1994, he was in the Mechanical Engineering Laboratory, Japan. He was an associate professor in Gunma University from 1994 to 1996 and in the Human Genome Center, University of Tokyo, from 1996 to 2001, respectively. He joined the Bioinformatics Center of the Institute for Chemical Research, Kyoto University, Japan as a professor in October 2001. His research interests include bioinformatics and discrete algorithms. He is a member of ACM, IEICE, JSAI and JSBi.

**James C. Whisstock** received his M.Phil. and D.Phil. degrees in Biochemistry from the University of Cambridge, UK in 1992 and 1996, respectively.

In 1997 he moved to Monash University to take up a two-year Monash University Faculty of Medicine Fellowship. In 1999 he was awarded an NHMRC Peter Doherty Fellowship, and in 2002 an NHMRC Senior Research Fellowship and Monash University Logan Fellowship. He is currently an ARC Federation Fellow and an honorary NHMRC Principal Research Fellow.

Professor Whisstock’s research group is based at Monash University and investigates three broad themes: (1) the structural biology of proteases, serpins and lipid phosphatases, (2) the structural biology of membrane-attack perforin (MACPF) and perforin-like proteins, and (3) a wide range of bioinformatic projects aimed at understanding the structure, function and evolution of large protein families.

**Robert N. Pike** received a B.Sc. degree in Agriculture and a Ph.D. degree in Biochemistry from the University of Natal, Pietermaritzburg, South Africa in 1987 and 1991, respectively.

From 1992 to 1994, he was a postdoctoral fellow in the Department of Biochemistry and Molecular Biology, University of Georgia, USA, following which he was a Foundation for Research and Development Research Fellow at the University of Natal, Pietermaritzburg, South Africa from 1994 to 1995.
From 1996 to 1997, he was a postdoctoral fellow in the Department of Haematology, University of Cambridge, UK. In 1997 he was appointed to a lectureship in the Department of Biochemistry and Molecular Biology at Monash University, Melbourne, Australia, where he is now a professor. Professor Pike’s research work has consistently involved investigations into the interplay between proteolytic enzymes and their inhibitors and receptors.