PoPS: A Computational Tool for Modeling and Predicting Protease Specificity

Sarah E. Boyd and Maria Garcia de la Banda
School of Computer Science & Software Engineering and Victorian Bioinformatics Consortium
Monash University, Melbourne, Victoria 3800, Australia
{sboyd, mbanda}@csse.monash.edu.au

Robert N. Pike and James C. Whisstock
Department of Biochemistry & Molecular Biology and Victorian Bioinformatics Consortium
Monash University, Melbourne, Victoria 3800, Australia
{rob.pike, james.whisstock}@med.monash.edu.au

George B. Rudy
Genetics & Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research
1G Royal Parade, Parkville, Victoria 3052, Australia
georgerudy@stanfordalumni.org

Abstract

Proteases play a fundamental role in the control of intra- and extracellular processes by binding and cleaving specific amino acid sequences. Identifying these targets is extremely challenging. Current computational attempts to predict cleavage sites are limited, representing these amino acid sequences as patterns or frequency matrices. Here we present PoPS, a publicly accessible bioinformatics tool (http://pops.csse.monash.edu.au/) which provides a novel method for building computational models of protease specificity that, while still being based on these amino acid sequences, can be built from any experimental data or expert knowledge available to the user. PoPS specificity models can be used to predict and rank likely cleavages within a single substrate, and within entire proteomes. Other factors, such as the secondary or tertiary structure of the substrate, can be used to screen unlikely sites. Furthermore, the tool also provides facilities to infer, compare and test models, and to store them in a publicly accessible database.

1. Introduction

Proteases (also referred to as proteinases, peptidases or proteolytic enzymes) are a class of proteins which appear in all forms of life. Their function is to cleave other proteins, referred to as their *substrates*. This cleavage often activates, inactivates, or modifies the substrate, and thus controls a diverse range of biological processes.

Inappropriate proteolytic activity can have devastating consequences, and is the cause of numerous human diseases. Thus, much research focuses on identifying the target substrates and inhibitors of proteases in these disease states, with the ultimate goal of designing appropriate treatments. A primary step in identifying the target substrates and inhibitors of a protease is understanding its specificity, i.e. the specific amino acid preferences exhibited by the protease, since this is a major determinant of which substrates it will cleave and with what affinity.

One of the main factors affecting protease specificity is its *active site*, a cleft in the protease structure which binds to the substrate and cleaves it. This active site contains a number of contiguous pockets called *subsites* (see Figure 1). Each of these subsites binds a single amino acid within the substrate sequence, with consecutive subsites binding consecutive amino acids, and cleavage occurring at the scissile bond. We use the standard S/S' subsite and P/P' substrate notation for protease-substrate interaction as defined by Schechter and Berger [19], where P_1-P_1' represents the scissile bond, i.e. the bond that is cleaved in the substrate (see Figure 1). The particular number of subsites in the active site of a given protease, together with the specific size, charge and shape of each of these subsites, are the major components defining a protease's specificity. In order for an amino acid to be accommodated within a subsite, the size, charge and shape of the amino acid and the subsite must

Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)
0-7695-2194-0/04 $20.00 © 2004 IEEE

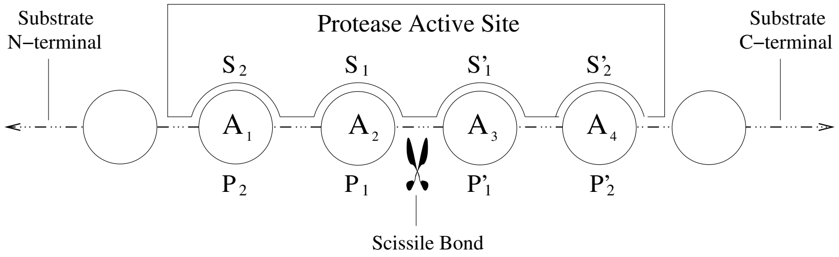


Figure 1. The interaction between the active site of a hypothetical protease with four subsites and the amino acids of a substrate, showing the notation of Schecter and Berger [19].

be compatible, with better compatibility resulting in a better binding and an increased likelihood of cleavage. In addition, the relative importance of the subsites in determining cleavage can vary, with one or more subsites clearly dominating the result. Other factors that can also influence specificity include the three-dimensional structure of the substrate, binding events between the substrate and the protease which occur outside the active site, and co-factors, i.e. molecules which can bind to the protease and modulate its specificity.

Unfortunately, while proteases represent around 2% of all gene products across all the genomes [18], the target substrates and inhibitors of many of these proteases remain uncharacterized. One reason is that generating specificity data, for example by testing in the laboratory the activity of a purified protease against a library of peptides, is costly and time consuming. Furthermore, even armed with specificity data, final identification of physiological targets often requires complex, time consuming *in vivo* experiments (experiments conducted in living cells and organisms) in order to fully understand the intricacies of a particular pathway. Another reason is the lack of accessibility to significant amounts of expert knowledge. There is, therefore, substantial demand for a publicly accessible computational approach to assist this process [18].

This paper presents PoPS (Prediction of Protease Specificity), an on-line computational tool (http://pops.csse.monash.edu.au/) for modeling protease specificity. PoPS supports the creation of computational specificity models of any protease from either measured experimental data or expert knowledge of primary sequence specificity. Thus, it allows models of *any protease* to be built from *any source of data* available to the user. These models are not only expressive enough to represent dependencies among subsites, but also to support quantitative analysis. Therefore, they can be used by PoPS to identify and rank likely cleavage sites within a target substrate sequence and within entire proteomes (containing all the known proteins for a particular organism). Specificity models can be stored in and retrieved from the publicly accessible database maintained by PoPS, based on the MEROPS protease database [18] which contains all proteases identified to date. In addition to primary sequence specificity, the tertiary or secondary structure of a target substrate can be used to predict inaccessible cleavage sites. Finally, PoPS provides tools to automatically infer, test and compare different models of the same protease, in order to determine the most suitable model.

2. Modeling and predicting protease specificity in PoPS

The PoPS computational model of protease specificity consists of three components. The first is the number of subsites within the active site of the protease. The second is the *specificity profile* of each subsite, which assigns a value to each of the 20 amino acids, with each value representing the relative contribution of the amino acid at that subsite to the overall substrate specificity of the protease. Values in the specificity profile are restricted to floating point numbers between -5.0 (most negative influence on binding) and +5.0 (most positive influence). This scale is large enough to accurately describe specificity, since floating point numbers allow a very high degree of precision, while still being meaningful for human users. In addition, the hash symbol '#' can be used to indicate amino acids that are known to prevent cleavage when appearing at a given subsite. The third and final component of the specificity model is the *weight* of the subsite, a positive floating point value which reflects the relative importance of each subsite in determining cleavage. The specificity model of a protease with J subsites can thus be represented by a $20 \times J$ position specific scoring matrix (PSSM) where each entry $r_{i,j}$ represents the relative contribution of amino acid i to subsite j, and a vector $w_1, ..., w_J$ representing the weights of each subsite.

The PSSM and weight vector are combined with a simple *sliding window* technique to obtain a score for each sequence of J consecutive amino acids in the substrate, as follows. Let $SS \equiv A_1, ..., A_J$ be the sequence of J consecutive amino acids in the substrate currently being aligned,

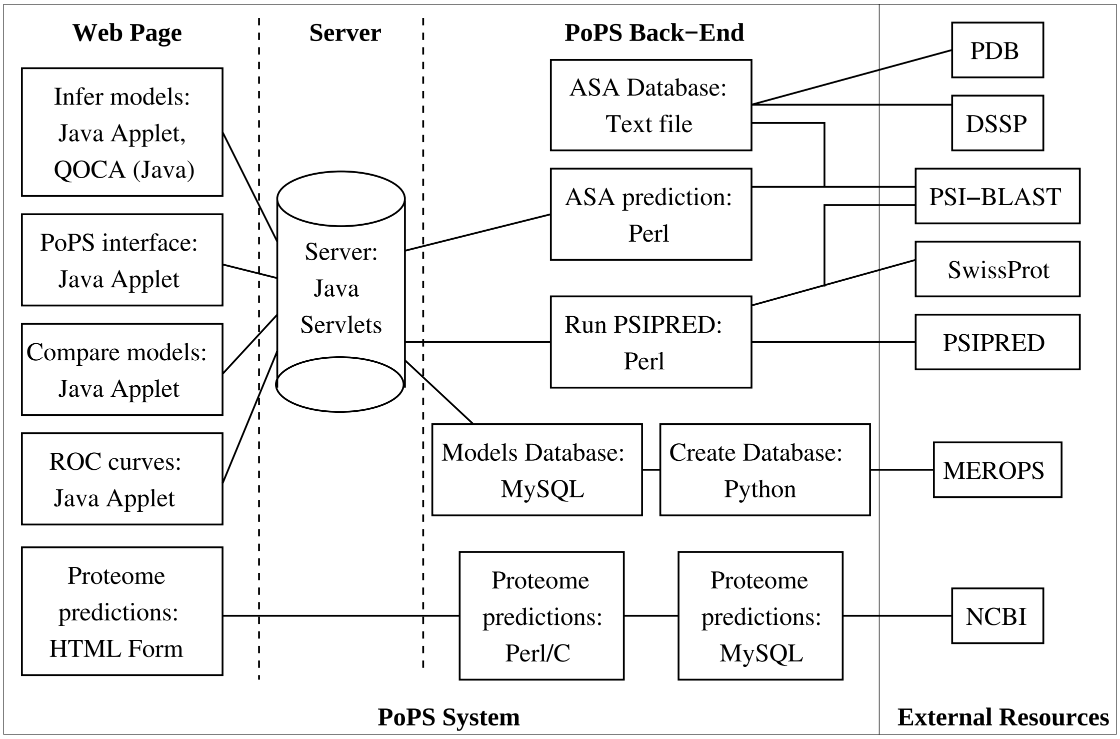


Figure 2. The PoPS system overview. Each rectangle indicates a distinct module in the system, together with its implementation language. The lines indicate how the modules are connected.

and let A_k, A_{k+1}, $1 \leq k \leq J-1$ represent the P_1, P_1' position of the scissile bond within the SS substrate sequence. The score at position A_k, A_{k+1} is computed as $\sum_i w_i * r_{i,A_i}$, and indicates the preference for a cleavage occurring at the position of the scissile bond. The higher the score, the more favourable the cleavage.

Note that the formula used to compute scores assumes independence among subsites. This is a common assumption in protease biology [5, 15, 16], made with the expectation that even if independence is not absolute, the results will still be useful. However, this assumption does not always hold [10] with cooperativity among subsites sometimes playing a significant role in determining substrate cleavage. For example, while the trypsin protease has been shown to be independent [16], it has two very specific exceptions: a proline in P_1' inhibits trypsin cleavage unless there is either a tryptophan in P_2 and a lysine in P_1, or a methionine in P_2 and an arginine in P_1 [13]. In order to support modeling of such proteases, PoPS allows users to enrich their specificity models with *dependency rules* of the form `(Mask,Kind,Value)`, where `Mask` is a sequence of amino acids in which `X` indicates any amino acid, `Value` is a signed decimal value, and `Kind` can be either `T` or `P`. These rules modify the usual matrix scoring method as follows. A rule with `Kind` set to `T` indicates that if the sequence SS of amino acids whose score is currently being computed “matches” that of `Mask`, then the score for SS is that given by `Value`, instead of the one computed using PSSM. For example, `(XAXB, T, 20)` replaces the sliding window score if A is found at position 2 and B is found at position 4 in the sliding window. A rule with `Kind` set to `P`, on the other hand, indicates a mixed approach: the final score for SS is that of `Value` plus the values of the matrix entries for the amino acids which matched an `X` in `Mask`. For example, `(XCXD, P, -5)` replaces the score for C and D with -5, but calculates the rest of the score using the PSSM. If more than one rule is applicable, the first one provided by the user will be used.

3. System Design

Figure 2 shows the general structure of the current PoPS system: a Web-based front-end which provides the user interface, a back-end which performs the predictions and manages the databases, and a server connecting the front-end to the back-end. Each module in the system is implemented using the programming language most appropriate for its needs (Figure 2). Upon access to the main program (Figure 2: *PoPS interface* module), a Java Applet containing the core PoPS program is downloaded to the user's

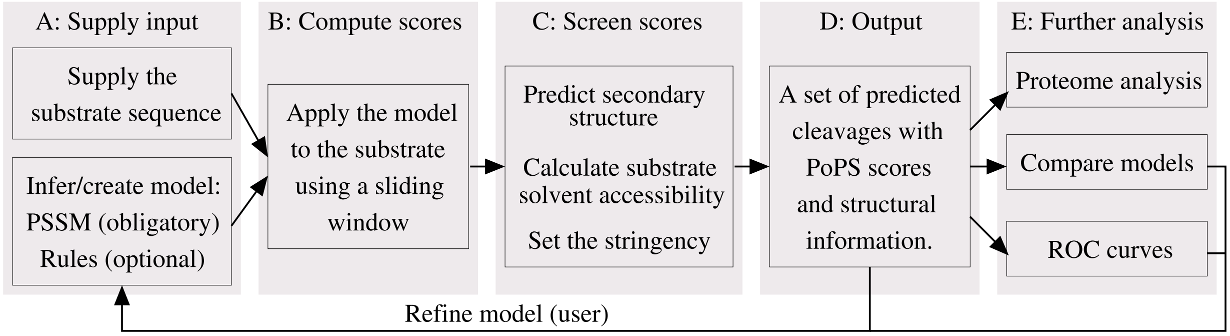

A: Supply input
B: Compute scores
C: Screen scores
D: Output
E: Further analysis
Supply the substrate sequence
Infer/create model: PSSM (obligatory) Rules (optional)
Apply the model to the substrate using a sliding window
Predict secondary structure
Calculate substrate solvent accessibility
Set the stringency
A set of predicted cleavages with PoPS scores and structural information.
Proteome analysis
Compare models
ROC curves
Refine model (user)

Figure 3. The processs of model development and cleavage prediction using PoPS.

computer. Most subsequent computations are executed locally, thus increasing the speed of the program's execution and removing the need for manual downloads, installations, or upgrades. However, the central databases are located on a server at Monash University, and any computations requiring the use of those databases are performed on the server. Figure 3 outlines the common sequence of steps that are used when creating and experimenting with protease specificity models. These steps are discussed in detail below.

3.1. Automatically building models from experimental data

The first step in using the PoPS system is to obtain a specificity model for the protease under investigation. Structured specificity studies, such as fluorescence quenched substrates [5], or position specific synthetic combinatorial libraries [23], build libraries of substrates whose sequence is carefully designed to represent the effect of single amino acids at individual subsites on the specificity of the protease. In theory, a model can be constructed from this data using linear regression but, in practice, the result of the regression analysis is mathematically equivalent to simply scaling the experimental measurements for all the amino acids at a given subsite to within the -5.0 to +5.0 range. This facility is provided to the user through the subsite profile window in the PoPS interface.

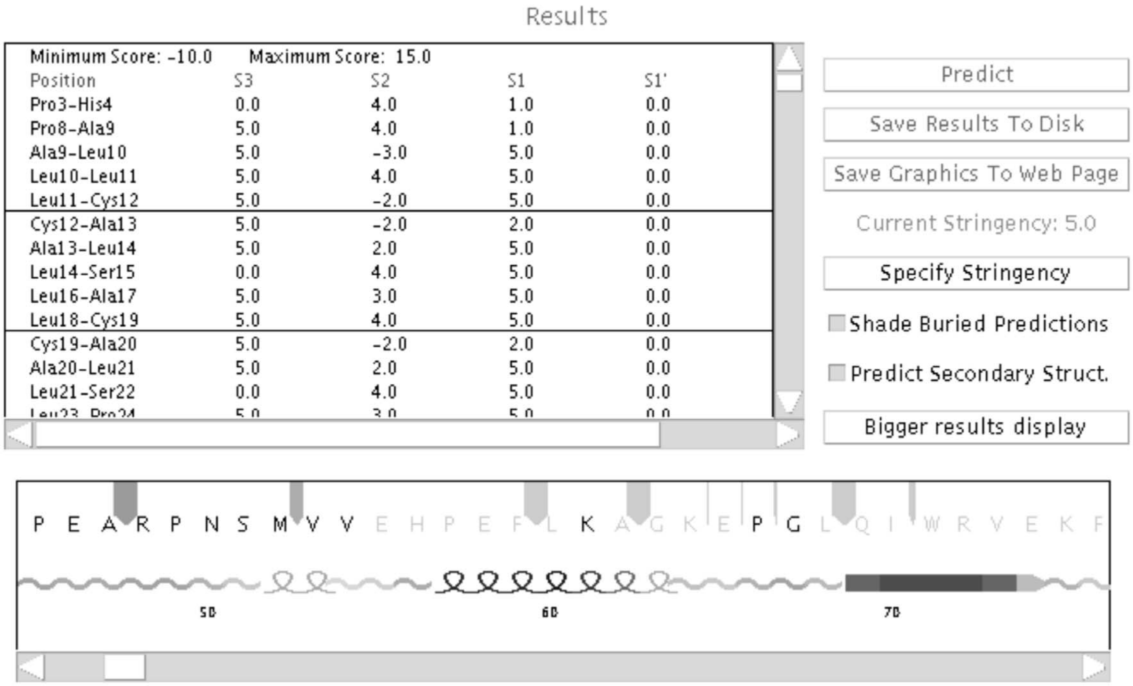
For unstructured data, PoPS provides a separate module (Figure 2: *Infer models* module) that applies regression analysis to produce a PoPS PSSM. The user must supply the amino acid sequences of the substrates and their associated kinetics data. If enough experimental data is available, the module will return the relative contributions of the amino acids together with the regression coefficient R^2.

Both of these methods produce a weight vector in which the weights of all subsites are set to 1, and an empty set of dependency rules. We expect the former to be always the case since the weights were always intended to be specified by expert users (see below). Regarding the latter, we are investigating different data mining based techniques capable of automatically inferring dependency rules from experimental data.

Incomplete specificity data will, of course, result in less accurate predictions. For example, if an amino acid's contribution is set to 0.0 because the real contribution is unknown, but in fact should have a negative score, PoPS will predict it as more favourable than it is, resulting in over-prediction of cleavages. Conversely, a favourable residue with missing specificity data (again set to 0.0) will not be selected by PoPS, resulting in an under-prediction of cleavage sites. Further, adding to a model subsites that do not influence cleavage may also affect the rate of over-/under-prediction. The PoPS interface allows easy investigation of how these subtle changes affect the predictive accuracy of a model, and therefore allows the user to gain a better understanding of the specificity of the protease.

3.2. Building models from expert knowledge

New specificity models for any protease can also be constructed through the PoPS model panel using expert knowledge. The model panel allows the user to determine the required number of subsites, assign weights to each subsite (as an intuitive and visually clear way to express relative importance among subsites), and directly provide subsite profiles by entering the values of each of the 20 amino acids for each subsite. Expert users may be able to specify the numbers directly, based on their experience with their protease, or any other resource. In addition, the subsite profile window also contains predefined common profiles, such as "Hydrophobic" or "Small", whose suggested values can be used and modified by the user. The model panel also allows the user to provide a set of dependency rules. Again, these rules can be obtained from expert knowledge, or any other available source. It is also possible to load a model of any protease previously stored in the Models database and modify it to build a new model for the protease of interest.



```
Minimum Score: -10.0    Maximum Score:  15.0
Position          S3        S2        S1        S1'
Pro3-His4         0.0       4.0       1.0       0.0
Pro8-Ala9         5.0       4.0       1.0       0.0
Ala9-Leu10        5.0       -3.0      5.0       0.0
Leu10-Leu11       5.0       4.0       5.0       0.0
Leu11-Cys12       5.0       -2.0      5.0       0.0
Cys12-Ala13       5.0       -2.0      2.0       0.0
Ala13-Leu14       5.0       2.0       5.0       0.0
Leu14-Ser15       0.0       4.0       5.0       0.0
Leu16-Ala17       5.0       3.0       5.0       0.0
Leu18-Cys19       5.0       4.0       5.0       0.0
Cys19-Ala20       5.0       -2.0      2.0       0.0
Ala20-Leu21       5.0       2.0       5.0       0.0
Leu21-Ser22       0.0       4.0       5.0       0.0
Leu23-Pro24       5.0       3.0       5.0       0.0
```

Figure 4. The results panel of the main PoPS program.

3.3. Predictions display

Individual substrates are supplied to PoPS through the substrate panel in the main program, using the single-letter amino acid coding. The predicted scores computed for a given specificity model and a substrate are displayed in both text and graphical format within the PoPS results panel (Figure 4). While our method obtains a score for every possible cleavage position in the substrate, not all scores might be of interest to the user. To avoid cluttering the screen, scores that involved a '#' symbol are recorded as -Infinity and never displayed, as they indicate cleavages that would not occur. Furthermore, a stringency value can be provided by the user to avoid displaying scores less than this value (Figure 4). The textual display then lists scores, with the contributing subtotals from each subsite, for each predicted cleavage that has a score that exceeds the stringency value, and the graphical display shows the substrate sequence with these predicted cleavages drawn as arrows (Figure 4). The more intense the green or red of an arrow the more positive or negative the score, respectively; the thicker the arrow, the greater the absolute value of the score.

The PoPS results panel can be used in two different contexts. During model development, the results panel is used to test the model by using substrates for which known cleavage data is available, and to refine the model according to the results. Once an accurate model has been defined, the results panel can then be used to predict the cleavage of target substrates.

3.4. Comparing different models of the same protease

PoPS allows the user to compare the accuracy of the predictions made by several models of a protease (Figure 2: *Compare models* module). To use this feature, the user must provide the name of the files containing each model, together with the name of the file containing the substrates against which the models will be compared. Typically, the substrates will be short sequences experimentally known to be cleaved or not cleaved (usually referred to as true positives and true negatives). If so, the expected (positive or negative) score of these known cleavage sites can be included in the file. PoPS will then apply each model to all substrates and compare the results of each model (and the expected score if any) using three different representation formats. The first is a simple bar chart graphing the predicted and expected scores. The second is an Excel table which includes for each model, substrate and cleavage site, such information as the expected score, the predicted score, the rank of the predicted score relative to every other score obtained for that substrate, and the maximum score for the entire substrate (Table 1 shows some of the information provided by this table). These two formats are aimed at providing easy means of comparison whenever low amounts of known cleavage data are available. Finally, PoPS can also create ROC curves that measure the performance of each model (Figure 2: *Compare models* module). These curves will only be meaningful if significant amounts of information regarding true positives and true negatives is provided.

3.5. Models database

As mentioned before, accessing protease specificity data is a difficult and time-consuming task. To assist in this, PoPS provides a publicly accessible database of specificity models that can be stored and retrieved by any user. The general classification data of each protease in the database is automatically derived from the MEROPS database [18], an on-line protease database which provides classification information for all known proteases, and has been used to index the PoPS models database. To preserve the integrity of the PoPS database, users are required to register their name, organization and email address, and choose a login name and password, before models can be saved. Only the name of the creator is made publicly available.

Each stored model has a unique identifier derived from the combination of the MEROPS protease identifier, the surname of the user, and the model number and version. In addition to storing the model, data such as the date, creator's name, specific organism, bibliographic details and extra comments are also included. Furthermore, if the newly stored model is a modification of an existing model, it is possible to indicate this (together with the reasons for the modification). The new model will then retain the original identifier, but will be assigned a different version number. All these data assist the users in finding the most appropriate model for their needs.

3.6. Accessible Surface Area (ASA) database

High scores might be calculated for positions that are inaccessible due to the 3-dimensional structure of the substrate. To help screen such predictions, PoPS maintains an Accessible Surface Area (ASA) database, used to determine the accessibility (surface or buried) of the substrate's amino acids (Figure 2: *ASA prediction* module). The database is created by automatically pruning the Protein Data Bank (PDB) database [4] to yield a single 3-dimensional model per protein. These models are processed by DSSP [12], which passes a 1.4 angstrom radius molecule over the surface of each 3-dimensional model to calculate the solvent accessibility of every residue in the protein. The results are stored with the corresponding protein sequence in the ASA database.

After cleavages have been predicted, the user can access the ASA data through the PoPS results panel (Figure 4). PoPS uses BLASTP [3] with an expect score of 0.001 to identify significant sequence similarity between the substrate and any sequence in the ASA database. These sequences (if any) are returned to the user as a list that includes the name and PDB accession number of the aligned protein, and the expect value from the BLASTP alignment. When the user selects one of these structures, the accessibility data already calculated by DSSP is mapped onto the substrate. Sections of the substrate that cannot be aligned by BLASTP are assumed to be accessible. The minimum percentage of an amino acid that must be solvent accessible before it is considered to be accessible to the protease (and therefore able to participate in a cleavage reaction) is by default 33%, but can be easily modified by the user if extra information about the size and shape of the active site suggests another value. Buried amino acids are shaded grey in the graphical display (Figure 4). Scores involving one or more buried amino acids are also shaded grey in both the graphical and textual displays. The grey shading is intended as an alert to potential inaccessibility. However, predictions should not be ignored without considering other factors, such as how many amino acids are buried across the active site, the significance of those amino acids in the cleavage process, and the accessibility of the regions surrounding the cleavage site.

3.7. Secondary structure prediction

If no 3-dimensional structure information is available, PoPS utilizes secondary structure prediction as a guide for further screening of cleavage sites (Figure 2: *Run PSIPRED* module). This can indicate where cleavages might be more likely to occur if, for example, a protease has a preference for coiled regions and loops. Secondary structure prediction is performed by aligning the substrate against the proteins in the SWISS-PROT database [7] using PSI-BLAST [3]. The PSI-BLAST output (after 2 iterations with an expect score of 0.001) is passed to PSIPRED [11], which uses a neural network to predict secondary structure with an average Q3 score of nearly 78%. The predicted secondary structure is drawn beneath the substrate in the graphical display (Figure 4). Helices are represented as blue coils, sheets as red arrows, and random coil as green waves. The intensity of the colouring of the secondary structure reflects PSIPRED's confidence of the prediction for the given amino acid: the more intense the color, the greater the confidence.

3.8. Analysis of proteomic data

Models can also be applied to the proteomic data of any organism currently available in PoPS (Figure 2: *Proteome predictions* module). PoPS currently supports analysis for *Homo sapiens*, *Saccharomyces cerevisiae*, *Escherichia coli*

Substrate	Cleavage Sequence	Cleavage Score	Maximum Substrate Score	Score Rank	Accessibility (33%)	Secondary Structure
Bcl-Xl	HLAD/S	9.99	12.80	2	Accessible	-
Pro-Interleukin 1 beta	FEAD/G	18.21	18.21	1	Unknown	Sheet/Coil
	YVHD/A	15.47	18.21	2	Unknown	Coil
Pro-Interleukin 18	LESD/Y	12.42	12.42	1	Unknown	Coil
Calpastatin	ALAD/S	9.75	12.20	9	Unknown	Helix/Coil
	LSSD/F	9.00	12.20	17	Unknown	Helix
	ALDD/L	7.05	12.20	47	Unknown	Coil

Table 1. Results for the Caspase 1 model F over known cleavage sites.

K12, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Rattus norvegicus*, *Mus musculus* and *Danio rerio*, all of which are obtained from the Reference Sequence database [17]. The results are returned by e-mail (usually within half an hour) and consist mainly of a file containing (for each putative substrate), the name, description, and a short reasoning table showing the top ten predicted cleavage sites. Results can be screened by selecting a cut-off value for the scores returned, and by choosing to receive only substrates containing less than a given number of cleavages. ASA information is included with the predictions. The proteomic analysis is intended to be used with models that are already known to be reasonably accurate.

4. Case Study: Caspase 1

This section illustrates the experimentation process supported by PoPS with a case study using caspase 1, an important protease involved in apoptosis (cell death) and cancer. The specificity of its S_4 to S_2 subsites has been profiled using positional scanning synthetic combinatorial libraries (PS-SCL) [23], and that of S_1' using fluorescence-quenched substrates [21]. Both studies yielded similar specificity profiles, indicating the two data sets could be successfully combined into one model. Therefore, a PSSM was created by scaling the S_4-S_2 PS-SCL data and the S_1' fluorescence quenched data to the range 0.0 to +5.0, and setting unprofiled amino acids to 0.0. Since caspase 1 can only accept the amino acid aspartate in the S_1 position [21], the entry for aspartate at S_1 was set to +5.0 and everything else in S_1 was set to '#'. All the weights were set to 1.0 and no dependency rules were added. The maximum score obtainable by the model is 25.0, and the minimum (other than -Infinity) is 0.0.

In addition, we developed another 5 different models of this protease using different combinations of the measured experimental data mentioned above, and general observations of behaviour, i.e. "expert knowledge" ([6, 9, 20, 22]). To choose the best model, we used the PoPS comparison module on a set of substrates that are known to be cleaved by caspase 1 [9]. Each substrate (first column, Table 1) has one or more known locations at which it is cleaved (second column; single-letter amino acid encoding). In addition, the ROC curves module was used to measure the performance of the 6 models, with the results shown in Figure 5. Only those (sub)-sequences in the substrates with an aspartate at the P_1 positions were entered as true positives (if they are known to be cleaved) or true negatives (all the rest). Including positions without an aspartate in P_1 would bias the ROC curves in favour of the models. Generally, the models incorporating measured data (models A, B, E and F) performed better than those using only expert knowledge (models C and D), although they all seem to perform reasonably well. The best model, F, is the one described in detail at the beginning of this section. This model obtained an area under the curve of 0.8, and is the model used for the remainder of the case study.

Having selected the model, we can use PoPS to further investigate its accuracy by using the PoPS main interface to apply the model to each substrate and carefully studying the results. Table 1 shows a summary of these results, including the predicted score for the known cleavage site, the maximum score in the same substrate and the associated rank. For example, the actual Bcl-Xl cleavage site is calculated to have the second-to-highest score of all possible cleavages within the Bcl-Xl sequence. Potential caspase 1 substrates have been investigated on the basis of the role of caspase 1 as a mediator of inflammation. Likely targets are selected and tested, usually *in vitro* ("in the test tube"), for cleavage by the caspase. The predictions for these substrates (Table 1) are quite accurate except for calpastatin. Calpastatin is an inhibitor of calpain, another protease prominent in apoptosis, and by cleaving calpastatin, caspase 1 could help promote apoptosis [24]. Thus, it has been tested *in vitro* and found to be cleaved by caspase 1, although the *in vivo* cleavage remains to be confirmed. It is worth noting, however, that if calpastatin is a true caspase 1 substrate, the PoPS model indicates that its primary amino acid sequence is not the main factor in determining cleavage. Other factors, such as a possible conformational change related to calpastatin's inhibitory mechanism, should be considered. It could also be the case that the *in vitro* environment allows

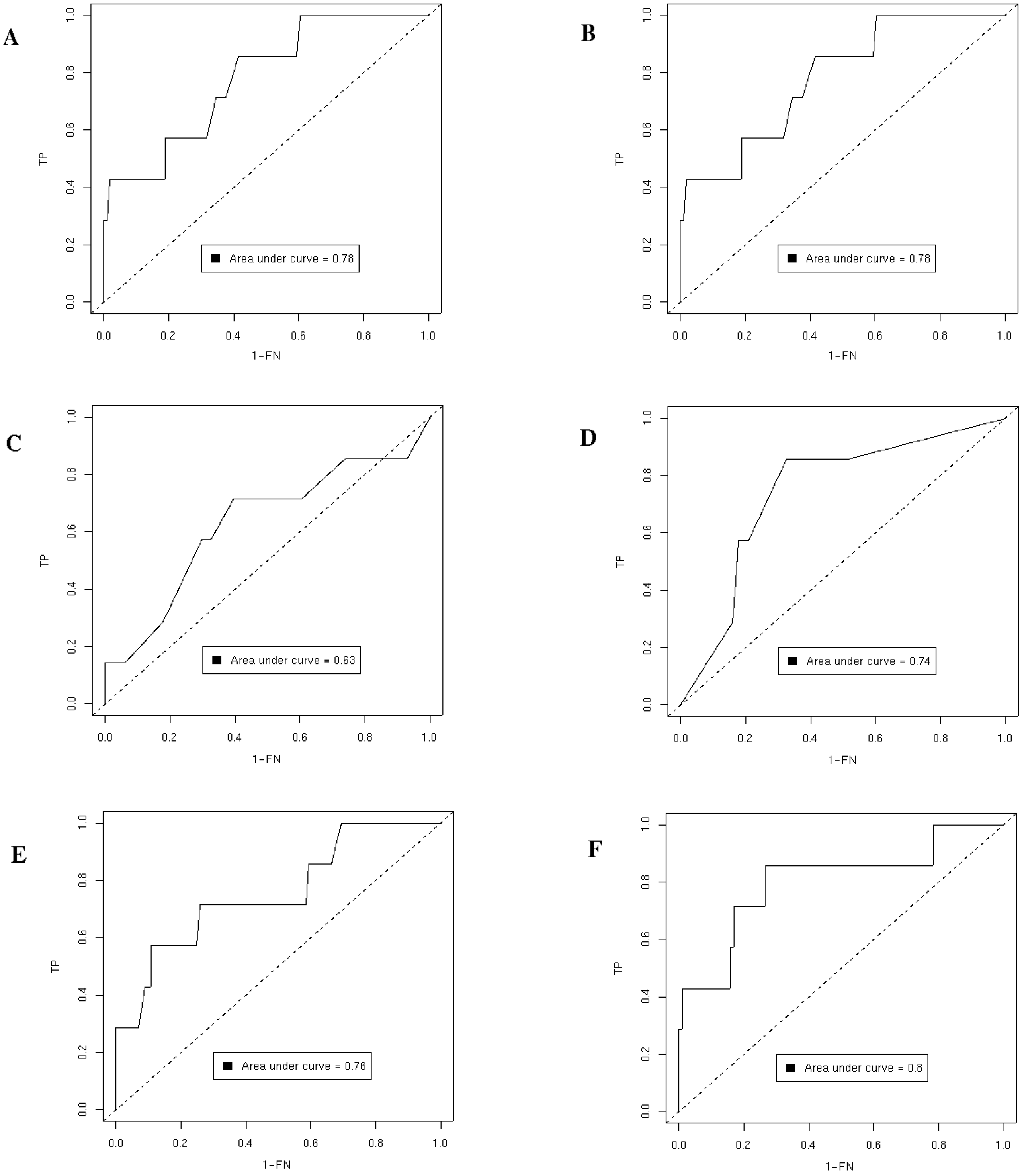


Figure 5. ROC curves measuring the different models constructed for caspase 1. Model A: [23]. Model B: [6, 20, 23]. Model C: [22]. Model D: [9]. Model E: [6, 20, 22, 23]. Model F: [21, 23].

NCBI Accession	Substrate	Possibly Buried?
NP_000383.1	ATP-binding cassette	Yes
NP_065816.1	Retinoblastoma-associated factor 600	No
NP_002717.2	Prolyl endopeptidase	Yes
NP_002850.1	Paxillin	Yes
NP_055559.1	TBC1 domain family, member 5	No
XP_291485.2	Similar to Myosin-binding protein H	No
XP_064873.2	Similar to carcinoembryonic antigen-related cell adhesion molecule 5	No
XP_058409.3	Similar to Potential carboxypeptidase-like protein X2 precursor	Yes

Table 2. The top scoring targets for caspase 1 from the human proteome analysis.

a cleavage to occur that would simply not occur *in vivo*.

In addition to primary sequence, we would also like to know if the real cleavage sites are calculated as accessible to the enzyme. As indicated by the sixth column of Table 1, the only substrate for which accessibility data is available is Bcl-Xl which, using the default minimum of 33% solvent accessibility as the threshold, is predicted as accessible to caspase 1. We may also choose to apply secondary structure prediction to substrates for which accessibility data was not available (column 7, Table 1). Random coil predicted across a cleavage site would be a positive indication for cleavage, but helices might be a negative indicator. All the known cleavages with rankings of 1 and 2 are predicted as having a sheet or random coil structure, which are generally considered to be cleavable structures. Again calpastatin was a notable exception, with two of the three cleavages being predicted to consist (at least partly) of a helical structure, again suggesting a possible conformational change during cleavage.

Finally, Model F was applied to the human proteome (currently 25,835 proteins) with a stringency of 22.00 (since the maximum possible score for the model is 25.00), and no limit on the number of predicted cleavages in the substrate. Table 2 shows the accession number and name of the potential targets returned by the model, the computation time being around 3 minutes. The third column indicates whether PoPS found accessibility data for the predicted cleavage that suggested that it was inaccessible.

Whilst these predictions remain to be tested for their biological relevance, it is interesting that in such a large database of potential hits, some very relevant substrates were returned. As mentioned before, caspase 1 is involved in cell death and cancer. “ATP-binding cassette” has been implicated in anti-cancer drug resistance, and “Retinoblastoma-associated factor 600”, “Paxillin” and “Similar to carcinoembryonic antigen-related cell adhesion molecule 5” are all implicated in various cancers. Such targets may be worthy of further investigation as potential substrates of caspase 1 or as therapeutic targets. The results may be improved even further by classifying (where possible) the proteins in a proteome using properties such as sub-cellular localization or functionality, to improve the screening of proteome predictions.

5. Related Work

The PoPS computational model of specificity was initially developed as an honours thesis by S.E. Boyd in [8] and was presented as a poster at the Gordon Research Conference on proteolytic enzymes and their inhibitors in July 2002. At this time there were only two publicly available, on-line computational tools capable of generalized prediction of protease cleavage of individual substrates: Cutter [1] and PeptideCutter [2]. Both operate on a fixed, limited set of proteases with predefined, unalterable models, where each model is simply a set of amino acid patterns derived from known cleavage sequences. Cleavage is predicted if an exact copy of a pattern in the model appears within the substrate (substring matching). A third tool, PEPS [14], is not available on-line but has recently and independently been proposed for predicting the specificity of cysteine proteases. Its specificity models are built from the frequency with which amino acids occur in known cleavages, and can be used to predict cleavage of individual substrates, and to search proteomes for new targets [14].

Regarding their computational models of specificity, neither Cutter nor PeptideCutter support quantitative analysis of likelihood and, therefore, favour simple models containing very few patterns (those associated with the most common cleavage sequences). While PEPS supports a quantitative analysis, its models can only be built from one source of data, and their accuracy relies on having substantial amounts of known cleavage sites, data which is rarely available. Furthermore, PEPS models cannot express dependencies among subsites. Models from all three tools can be easily represented in PoPS: those of Cutter and PeptideCutter as dependency rules, and those of PEPS as a particular kind of PSSM. Regarding their general capabilities, while both Cutter and PeptideCutter are on-line tools, they do not allow the user to provide their own specificity models, nor do they provide any support for further investigation of such models. PEPS is not an on-line tool and while it can be used to search entire proteomes, it does not seem to support any other kind of experimentation.

6. Conclusions

PoPS allows users to build expressive computational models of protease specificity for any protease from any data source available to the user. The models can be stored in a publicly accessible database, and used to predict likely cleavage sites within a given substrate, or to predict novel targets within an entire proteome. Importantly, PoPS provides a research environment that allows the user to automatically infer, compare and test computational models. Current work is focused on improvements to the specificity model through the investigation of data mining techniques for automatic inference of dependency rules, and docking techniques for modeling specificity factors other than primary sequence, such as conformational changes.

7. Acknowledgments

Kind thanks to Alan Barrett for giving permission to use the MEROPS database. Thanks to Michael Cameron, Di Wu, Suan Khai Chong, Stewart Hore, Stanley Gunawan, Dave Powell, Torsten Seemann, Sean Guo, and Laurent Tardiff for assistance with implementation and server administration. Thanks to Terry Speed and Frédéric Schütz for general comments and assistance with ROC curves. MGB and JCW are Monash University Logan Fellows, and JCW is an NHMRC (Australia) Senior Research Fellow. Thanks to the NHMRC, the Australian Research Council, and the Victorian Bioinformatics Consortium for their support.

References

[1] http://delphi.phys.univ-tours.fr/Prolysis/cutter.html.

[2] http://us.expasy.org/tools/peptidecutter/.

[3] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[4] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[5] E. Bianchini, V. Louvain, P.-E. Marque, M. Juliano, L. Juliano, and B. L. Bonniec. Mapping of the catalytic groove preferences of FXa reveals an inadequate selectivity for its macromolecule substrates. *J. of Biol. Chem.*, 277(23):20527–20534, June 2002.

[6] R. Black, S. Kronheim, and P. Sleath. Activation of interleukin-1beta by a co-induced protease. *FEBS Letters*, 247(2):386–390, April 1989.

[7] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31:365–370, 2003.

[8] S. Boyd. Cleave: A Tool to Model Enzyme Activity. *Honours Thesis, School of Computer Science, Monash University*, 2000.

[9] W. Earnshaw, L. Martins, and S. Kaufmann. Structure, activation, substrates, and functions during apoptosis. *Annu. Rev. Biochem.*, 68:383–424, 1999.

[10] H. Gron and K. Breddam. Interdependency of the binding subsites in subtilisin. *Biochemistry*, 31(37):8967–71, September 1992.

[11] D. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.

[12] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[13] B. Keil. *Specificity of proteolysis*. Springer-Verlag, First edition, 1992.

[14] T. Lohmüller, D. Wenzler, S. Hagemann, W. Kiess, C. Peters, T. Dandekar, and T. Reinheckl. Towards computer-based cleavage site prediction of cysteine endopeptidases. *Biol. Chem.*, 384:899–909, 2003.

[15] M. Pozsgay, G. Szabo, S. Bajusz, and R. Simonsson. Study of the specificity of Thrombin with Tripeptidyl-p-nitroanilide substrates. *Eur. J. Biochem.*, 115:491–495, 1981.

[16] M. Pozsgay, G. Szabo, S. Bajusz, R. Simonsson, R. Gaspar, and P. Elodi. Investigation of the substrate-binding site of Trypsin by the aid of Tripeptidyl-p-nitroanilide substrates. *Eur. J. Biochem.*, 115:497–502, 1981.

[17] K. Pruitt, T. Tatusova, and D. Maglott. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.*, 31(1):34–37, 2003.

[18] N. Rawlings, E. A. O'Brien, and A. Barrett. MEROPS: the protease database. *Nucleic Acids Res.*, 30:343–346, 2002.

[19] I. Schechter and A. Berger. On the size of the active site in proteases. I. papain. *Biochem. Biophys. Res. Comm.*, 18(2):77–82, 1967.

[20] P. Sleath, R. Hendrickson, S. Kronheim, C. March, and R. Black. Substrate specificity of the protease that processes human Interleukin-1beta. *J. Biol. Chem.*, 265(24):14526–14528, August 1990.

[21] H. Stennicke, M. Renatus, M. Meldal, and G. Salvesen. Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1,3,6,7 and 8. *Biochem J*, 350:563–568, 2000.

[22] H. Stennicke and G. Salvesen. Properties of the caspases. *Biochim. Biophys. Acta*, 1387:17–31, 1998.

[23] N. Thornberry, K. Chapman, and D. Nicholson. Determination of caspase specificities using a peptide combinatorial library. *Methods Enzymol.*, 322:100–110, 2000.

[24] K. Wang, R. Posmantur, R. Nadimpalli, R. Nath, P. Mohan, R. Nixon, R. Talanian, M. Keegan, L. Herzog, and H. Allen. Caspase-mediated fragmentation of calpain inhibitor protein calpastatin during apoptosis. *Archives of biochemistry and biophysics*, 356:187–196, 1998.