

SitePredicting the cleavage of proteinase substrates

Jelle Verspurten^{1,2}, Kris Gevaert^{3,4}, Wim Declercq^{1,2*} and Peter Vandenabeele^{1,2*}

¹ Department for Molecular Biomedical Research, VIB (the Flanders Institute for Biotechnology), B-9052 Ghent, Belgium

² Department of Biomedical Molecular Biology, Ghent University, B-9052 Ghent, Belgium

³ Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium

⁴ Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium

Proteinases are enzymes that play important roles in vital cellular and extracellular processes by hydrolytically cleaving peptide bonds in their protein substrates. This cleavage can be non-specific as part of degradation during protein catabolism or highly specific as part of proteolytic cascades and signal transduction events. Several web tools are available for predicting possible cleavage sites in candidate substrates. Here, we compare existing prediction tools with SitePrediction, a novel and user-friendly tool for identifying potential cleavage sites. This prediction is based on known datasets found in the literature, stored in web-accessible repositories or generated by our own experiments. Comparison of the different programs shows that SitePrediction makes it possible to derive more reliable predictions. In addition, this tool allows the use of a wide range of proteinases.

Predicting proteinase cleavage sites using web tools

Proteinases are enzymes that hydrolyze peptide bonds between the amino acids of proteins. They represent ~2% of all gene products and are of particular importance in medicine and biotechnology because their effector function can easily be targeted by small peptide-based inhibitors or chemical compounds. Inappropriate proteolytic activity can have devastating consequences and is the cause of numerous human diseases [1,2].

Proteinases are classified on the basis of their catalytic mechanisms into six types: serine (S), cysteine (C), threonine (T), aspartic acid (D), glutamic acid (E) and metallo catalytic types. Only a few remain uncategorized in this way. Another classification is based on the kind of reaction they catalyze. An endoproteinase hydrolyzes internal α -peptide bonds in a polypeptide chain. Exoproteinases, by contrast, typically require a free N-terminal amino group, C-terminal carboxyl group or both and hydrolyze a peptide bond not more than three residues from the terminus. Some proteins can act as endoproteinases or exoproteinases depending on the pH.

Exoproteinases are involved mainly in degradative processes, such as food digestion, proteasome phagocytosis and proteasomal digestion (protein catabolism). Endoproteinases too might be rather aspecific (e.g. calpains and cathepsins), but they might also be highly specific for certain target sequences (e.g. caspases and granzyme-B,

which cleave after an aspartate residue). Specific degradation of proteins is also used in regulatory processes, as in the degradation of ubiquitylated I κ B inhibitor, which might lead to nuclear factor- κ B (NF- κ B) activation.

Cleavage of proteins requires that they first bind to the active site of the proteinases. For proteins involved in the control of biological processes, the cleavage activates, inactivates or modifies the substrate in some way. The active site of the proteinase contains several conterminous S pockets, which accommodate consecutive amino acids (called P sites – not to be confused with proline [P]) from the substrate (Figure 1). The P1 site is defined as the amino acid that is C-terminally cleaved. N-terminal from P1 are additional sites (P2, P3, P4, etc.) that are accommodated by corresponding S sites in the catalytic pocket of the proteinase. P' sites are C-terminal from P1. The P' sites can also be accommodated by corresponding subsites in the substrate-binding pocket. The number of P and P' sites of the substrate that fit the substrate-binding pocket varies from one proteinase to another. For certain proteinases the substrate specificity is determined by the number of subsites in the active site and by the size, shape and charge of the side chains involved. The compatibility and fit between the S sites in the substrate-binding pocket and the substrate P sites is influenced by the three-dimensional structure of the substrate. The interaction between the substrate and the proteinase outside the active site, the so-called exosites, and the influence of co-factors should also be taken into account [2,3]. In the end, the accessibility of the potential cleavage site will determine whether a substrate is cleaved or not.

Much research has focused on identifying proteinase specificity and target substrates in human diseases, with the ultimate goal of designing appropriate treatments. The conventional way of identifying proteinase specificity by testing the enzyme against a library of peptides is time consuming and expensive. Other techniques are based on genetic approaches (substrate phage display, substrate display by bacteria, yeast-based screening methods, mRNA-based approaches) or on proteome analytical approaches (2D-PAGE, 2D-DIGE, PICS and COFRADIC) [4]. Bioinformatics can be used to predict on the basis of the available data the possible substrates for the proteinases and the likely location of the cleavage sites. Some tools have already been developed for this purpose: PoPS for modeling and predicting proteinase specificity [5], PeptideCutter for prediction of potential cleavage sites for proteases [6], GraBCas for prediction of sites cleaved by granzyme B and caspases [7], and CaSPredictor for

Corresponding author: Vandenabeele, P. (peter.vandenabeele@dmb.rugent.be)

* These authors shared senior co-authorship.

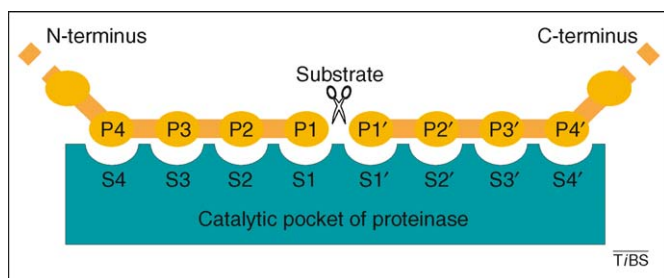


Figure 1. Representation of the proteinase-substrate interaction.

caspase substrate prediction [8]. Most tools are proteinase specific (GraBCas and CaSPredictor) and therefore have limited applicability. Furthermore, some tools are very complicated or have a cumbersome user interface. Recently, we developed 'SitePrediction' to provide researchers with a user-friendly tool to predict possible cleavage sites in candidate substrates based on cleavage sites found in the literature or identified in their own experiments. Some extra features, such as secondary structure prediction, solvent accessibility and PEST sequence occurrence, have been integrated into SitePrediction. These features will be explained and discussed in this article, which also offers a comparative analysis of the tools available for cleavage site prediction, which might be of interest for many protease researchers. SitePrediction is publicly available as a web application at <http://www.dmbr.ugent.be/prx/bioit2-public/SitePrediction/>.

Comparative description of the cleavage site prediction tools

The main goal of the tools under consideration is to predict the location of cleavage sites in candidate substrates. [Supplementary Table S1](#) provides an overview of the prediction programs we compare in this article: PoPS, GraBCas, CaSPredictor, PeptideCutter and SitePrediction. In this section, we give a brief overview of the three main steps that determine the applicability and reliability of these prediction tools. First, the user needs to define which protease specificity should be used and which protein sequences need to be analyzed (= user input). Second, the different calculation methods used will determine the reliability of the prediction. Third, extra features, such as statistical analysis, secondary structure, solvent accessibility and PEST-sequence prediction, can increase the prediction value.

User input

Two main decisions have to be made about the input: which protease are we interested in and in which candidate substrates do we want to predict possible cleavage sites? Most tools allow the user to enter the substrate sequences one by one in FASTA format (GraBCas, POPS and PeptideCutter), whereas others allow the entry of a list of FASTA format sequences (CaSPredictor and SitePrediction). SitePrediction also allows the user to enter a list of common IDs, such as SwissProt identifiers (<http://www.expasy.ch/sprot/>), NCBI Accession numbers or GenBank identifiers (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=Protein>, see [Supplementary Table S1](#)).

One of the most crucial steps in cleavage site prediction is to clearly define the specificity of a chosen protease. Some tools, GraBCas and CaSPredictor, only offer selection of a few protease specificities, such as granzyme B and some caspases, respectively. PeptideCutter allows the selection of more protease specificities, but only offers the use of fixed consensus sites. Therefore, PeptideCutter is not able to identify non-canonical cleavage sites. Some tools, such as PoPS, offer the choice of a list of predefined protease specificity profiles or allow insertion of a custom cleavage site profile. This profile indicates the frequency of every amino acid at the different P and P' positions. Because most users do not have a ready-to-use cleavage site profile, SitePrediction allows entry of so-called 'known cleavage sites' from a list based on the literature, databases or experimental protease degradomics data [9]. Such cleavage-site lists are then used to calculate a statistically relevant profile. The cleavage site analysis can be visualized by SitePrediction in two ways, a logo ([Figure 2](#)) and a histogram. For many proteinases, the site specificity is already included, whereas for others, the 'getMerops' feature helps the user to find sites that have been entered in the MEROPS database [1], a proteinase database of numerous proteinases, including experimentally derived cleavage sites. It should be kept in mind that the quality of such a 'training set' is a major determining factor for the outcome and the accuracy of the prediction. However, we simply cannot assess whether all literature reports provide equivalent confidence of reported results and can thus be treated equally [3]. By allowing the user to decide which input to use, the user is more aware of which data led to the prediction. This can help in correctly interpreting the results. The user can also decide which species-specificity to use, which is an important issue because orthologue proteases might have different substrates [10].

Scoring methods

A very important distinction among the different tools is the scoring method used to predict cleavage sites in a substrate. The most rudimentary tool (PeptideCutter) just looks for occurrences of fixed consensus cleavage sites in the substrate sequence. This approach can overlook cleavage sites deviating from the consensus sequence. All the other tools use a frequency score that indicates whether the amino acids of the potential cleavage site are likely to occur at that position. The implementation of this score, however, might not always be similar or might even contain errors. For instance, CaSPredictor adds the frequencies of each position instead of multiplying them. SitePrediction, as well as CaSPredictor, uses a second score that is based on the similarity of the potential cleavage sites to the known cleavage sites. This comparison is done by using an amino acid substitution matrix (like the BLOSUM 62 matrix [11]), and the actual used score is then arbitrarily chosen as the product of both scores. CaSPredictor also uses a PEST score to calculate the final score, but because the influence of a PEST sequence on proteolysis specificity is not evident from experimental data, this additional feature should not be used by default as a crucial determining factor.

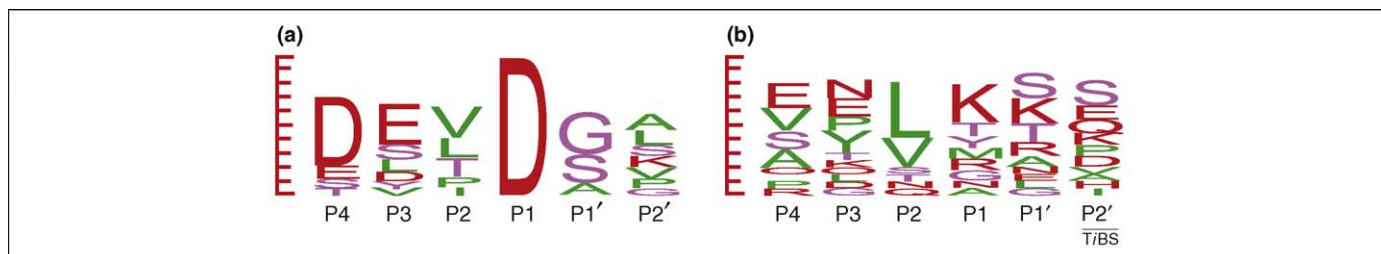


Figure 2. Distribution of the amino acids at each position. The logos were generated using SitePrediction. (a) Logo for human caspase-3. (b) Logo for human calpain-2. The experimental input cleavage sites (P4 to P2') for human substrates of caspase-3 or calpain-2 were taken from the MEROPS database.

Extra features: statistics, PEST sequences, solvent accessibility and secondary structure

Besides the main goal of predicting potential proteinase cleavage sites in a protein sequence, SitePrediction and PoPS offer some extra features that can be useful for interpreting the prediction results. An important feature in SitePrediction is the statistics calculation, which gives the user better insight into the quality of the input sites (i.e. known or expected protease specificities) and threshold scores of predicted cleavage sites by comparing the scores of experimentally known sites with those of random sequences. Both SitePrediction and PoPS also offer extra predictions, such as PEST region, solvent accessibility and secondary structure prediction. Regions rich in P, E, S and T could indicate that they are more susceptible to proteolysis [12] and might affect cleavage site prediction results by forming an unstructured loop [13]. Because appropriate presentation of a cleavage site in an exposed and unstructured region is crucial for efficient hydrolysis, we included an extra feature in SitePrediction by integrating the SSPro package, which predicts the solvent accessibility and secondary structure of a protein sequence [14]. The user runs the program on selected sequences and receives additional information on whether or not the predicted cleavage sites are accessible in solvent or are part of or close to an α helix, an extended β strand or in an unstructured region. As it is not proven that the presence or absence of secondary structures always has an important role in cleavage site predictions, results should be interpreted with care and never be considered conclusive.

Comparative analysis of prediction tools

Cleavage site prediction

We compared SitePrediction to four other cleavage site prediction tools (PoPS, GraBCas, CaSPredictor and PeptideCutter) on two different proteinases (caspase-3 and calpain-2). Caspases constitute an important family of proteinases involved in metazoan programmed cell death and inflammation [15]. Like all caspases, caspase-3 is characterized by cleaving specifically C-terminal of aspartic acid residues that occur in specific substrate recognition motifs (Figure 2a). We found 202 different sites in 135 substrates in the MEROPS database, and the logo generated by SitePrediction shows that the most frequent residues at the P4 to P1' positions are DEVD↓G (Figure 2a), as reported previously [16–18]. In general, the P1' position of the generated cleavage products was enriched for G, A or S residues, which correspond to the N-end rule of stabilizing residues [19], suggesting that most of the generated C-

terminal fragments would be stable in this respect. All the programs we compared can predict caspase cleavage sites, but CaSPredictor does not allow a distinction between different caspases.

The experimental input cleavage sites (P4 to P2') for caspase-3 were taken from the MEROPS database, and the corresponding *Homo sapiens* substrates were analyzed using the different cleavage site prediction tools. Because all tools generate their own scores and have different methods for determining whether or not a defined site is a potential cleavage site, it is not possible to use these theoretical scores as a comparative validation parameter. Therefore, the ranking of each experimentally known cleavage site was used as a first validation parameter, as it can be expected that the experimentally identified sites would rank highest in a list of possible cleavage sites. The percentages of these experimentally defined cleavage sites that are ranked first or within the top three are depicted in Supplementary Table S2. Finally, we compared the percentage of experimental input sites that are predicted by the tool ('true positives' in Supplementary Table S2).

Because PeptideCutter only takes the consensus cleavage site sequences into account for prediction, it only detects ~2.5% of all caspase-3 cleavage sites (Supplementary Table S2). Therefore, we did not further include this prediction program in our comparative statistical analysis. For SitePrediction, 39.6% of all known sites were ranked first and 63.9% were in the top three, which is considerably higher than the scores for GraBCas but similar to those for PoPS and CaSPredictor (Supplementary Table S2). The good ranking results can be explained by the absolute D specificity at the P1 position. SitePrediction (using a sensitivity cut-off of 95%) indicates all the known-sites (100%) as being possible cleavage sites, PoPS 81.7%, CaSPredictor 63.9% and GraBCas 41.6%, demonstrating that SitePrediction is superior in covering all experimentally defined data in this instance.

To further investigate the validity of the cleavage site prediction programs when using proteinases that do not have 100% specificity at a defined position, we examined known substrates of calpain-2 (Figure 2b). Because only PoPS and SitePrediction allow the definition of any peptide-cleaving motif, only these two tools were compared. Calpain-2 is a cysteine proteinase that is activated by calcium during endoplasmic reticulum stress and during anoxic neuronal cell death in stroke and spinal injuries. We used the 36 sites in 12 human substrates described in the MEROPS database to generate the cleavage site motif.

The top-one and top-three ranking parameters show drastically lower values as compared to the caspase-3 results (Supplementary Table S3). This could be expected because of the lower number of input sites and the absence of a 100% specificity position. By contrast, PoPS ranked only 5.6% and 8.3% of these cleavage sites in the top one and top three, respectively. Overall, SitePrediction detected 75% of the experimentally identified sites, whereas PoPS predicted only 41.7%. For all scoring parameters, SitePrediction gave better predictions of the experimental cleavage sites than PoPS. Taken together, we can conclude that SitePrediction has a broader applicability and a wider versatility than the other tools.

Evaluation of extra features of SitePrediction

Statistics

The statistics feature of the SitePrediction tool estimates the quality of the prediction and allows a threshold to be set to determine *in silico* cleavage sites. The SitePrediction statistics feature was run against input cleavage sites of both caspase-3 and calpain-2. An effective method for evaluating the performance of the tool is the receiver operating characteristic (ROC) curve, which is defined as a plot of the sensitivity versus its false positive rate (= one minus specificity) [20]. The accuracy of ROC analysis is measured by the AUC (area under the curve), and the values of 0.995 and 0.951 for caspase-3 and calpain-2, respectively, illustrate that the calculation is very accurate using the given input sites (Figure 3).

We generally assume a specificity of at least 95% for determining the threshold. This means that for caspase-3 the threshold of the average score is set at 0.05, which results in a sensitivity of 100%, meaning that all the known sites have a score higher than this threshold. For calpain-2, the 95% specificity is obtained at a threshold score of 0.375, with a sensitivity of 75%. Thresholds with higher specificity (>95%) can also be used, but this would result in a reduction in the number of predicted cleavage sites.

PEST analysis

The PEST analysis feature of SitePrediction was applied to the predicted substrates from caspase-3, calpain-2, granzyme-B and cathepsin-D. Substrates of granzyme-B and cathepsin-D have been identified extensively (31 and 17 substrates, respectively, were found in MEROPS [1]).

The percentage of amino acids in PEST regions was calculated for all substrates. These values were compared to those of all translated transcripts of the human genome currently available at NCBI. In the theoretical human proteome, 3.46% of all amino acids are calculated to be in PEST sequences, whereas for known proteinase substrates only, slightly different percentages are found (Supplementary Table S4). Looking at the percentage of the substrates that contain PEST sequences, a clearer difference emerges. Only 33% of the human proteins are predicted to contain PEST sequences, compared with 58%, 50%, 58% and 39% of the caspase-3, calpain-2, granzyme-B and cathepsin-D substrates, respectively. To further examine a possible correlation among the occurrences of PEST sequences containing potential proteinase cleavage sites, we calculated the percentage of amino acids of the experimentally known sites that are situated in a PEST sequence (Supplementary Table S4). Only caspase-3 and granzyme-B substrates seem to have a percentage of amino acids of known cleavage sites within PEST sequences, whereas known calpain-2 and cathepsin-D cleavage sites seem to negatively correlate with their presence within or partly within PEST sequences (Supplementary Table S4). Taken together, these results indicate that for caspase-3 and granzyme-B there is a higher frequency of cleavage sites in PEST sequence regions.

The presence of PEST sequences in the C- or N-terminus of the generated fragments might determine their turnover and stability, and thereby their function. For calpain-2 and cathepsin-D, there is no increase of the frequency of PEST sequences in the cleavage sites observed. This is in line with the role of caspases and granzyme-B in signal transduction events, as opposed to the non-specific functions of degrading proteases. The calculation of the PEST sequence might thus be more indicative for predicting a substrate than for predicting the position of a cleavage site.

Solvent accessibility and structure prediction

We analyzed the substrates of the proteinases with the SSPro package incorporated into SitePrediction to predict solvent accessibility and the secondary protein structure at the cleavage site. For caspase-3, 53.9% of all amino acids of the known substrates were predicted to be 'exposed', compared with 68.15% for the experimental cleavage sites (Supplementary Table S4). Predictions on the substrates

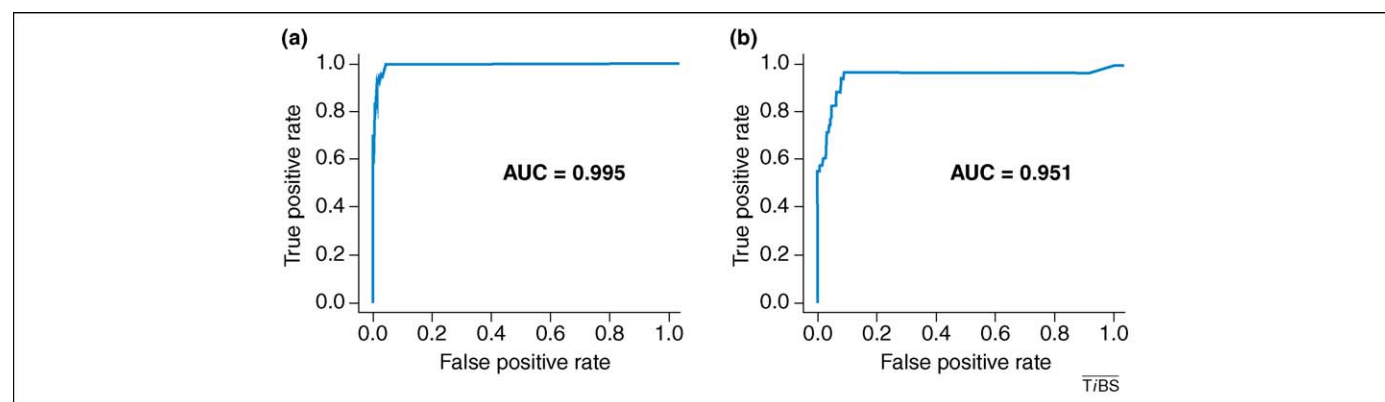


Figure 3. ROC curves generated by SitePrediction for two proteinases. (a) Caspase-3 ROC curve. (b) Calpain-2 ROC curve. The ROC curve is defined as a plot of the sensitivity versus its false positive rate (= one minus specificity). The area under the curve (AUC) gives an indication on the quality of the prediction.

of calpain-2 and granzyme-B also showed an increase of exposed residues, but the cathepsin-D results showed a slight decrease. This shows that cleavage sites might be present in protein regions that are more accessible to solvent than the rest of the substrate, but when solvent accessibility is used as an extra decision factor, caution and prior analysis of known substrates is needed.

SitePrediction also contains a secondary structure prediction feature that can analyze whether the presence or absence of a secondary structure would affect the prediction of potential cleavage sites. The cleavage sites of caspase-3 and granzyme-B are preferentially situated in the unstructured sequences (Supplementary Table S4), which agrees with experimental data [21], whereas calpain-2 cleavage sites are present equally in structured and non-structured protein sequences. By contrast, cathepsin-D cleavage sites are preferably present in structured regions. These observations are in agreement with the functions of the different proteases. Hence, depending on the proteinase, prediction of the secondary structure could be an additional feature for evaluating the likelihood of potential cleavage sites. More extensive analysis on large datasets using different proteases could reveal the general applicability of parameters such as solvent accessibility, PEST region prediction and secondary structure prediction in delineating *in silico* cleavage sites.

Conclusions

A comparative analysis of five publicly available protease cleavage site prediction tools shows that only two of them can make accurate predictions for a wide range of proteases (PoPS and SitePrediction). PeptideCutter scored worst in defining caspase-3 cleavage sites because it detected only 2.5% of the cleavage sites reported in the literature. GraBCas and CaSPredictor are limited by the number of proteases for which they can be used. The advantages of SitePrediction are that it is user-friendly, fast and allows the advanced user to gather additional features on the physicochemical environment of the potential cleavage sites (solvent accessibility, secondary structure and PEST sequences).

Acknowledgements

We thank Amin Bredan for editing the manuscript. P.V. is holder of a Methusalem grant from the Flemish Government. This work has been supported by the Flanders Institute for Biotechnology (VIB) and several Grants from the European Union (EC Marie Curie Training and Mobility Program, FP6, ApopTrain, MRTN-CT-035624; EC RTD Integrated Project, FP6, Epistem, LSHB-CT-2005-019067; EC RTD Integrated Project, FP7, APO-SYS, Health-F4-2007-200767), the Interuniversity Poles of Attraction-Belgian Science Policy (IAP6/18), the Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (3G.0218.06), and the Special Research Fund of Ghent University (Geconcerteerde Onderzoekstacties 12.0505.02). The Department of Medical Protein Research further acknowledges support by research grants from the Fund for Scientific

Research-Flanders (Belgium) (project numbers G.0156.05, G.0077.06 and G.0042.07), the Concerted Research Actions (project BOF07/GOA/012) from Ghent University, the Interuniversity Attraction Poles (IUAP06) and the European Union Interaction Proteome (6th Framework Program).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.tibs.2009.04.001](https://doi.org/10.1016/j.tibs.2009.04.001).

References

- Rawlings, N.D. *et al.* (2004) MEROPS: the peptidase database. *Nucleic Acids Res.* 32, D160–D164
- Turk, B. (2006) Targeting proteases: successes, failures and future prospects. *Nat. Rev. Drug Discov.* 5, 785–799
- Timmer, J.C. and Salvesen, G.S. (2007) Caspase substrates. *Cell Death Differ.* 14, 66–72
- Van Damme, P. *et al.* (2008) Disentanglement of protease substrate repertoires. *Biol. Chem.* 389, 371–381
- Boyd, S.E. *et al.* (2005) PoPS: a computational tool for modeling and predicting protease specificity. *J. Bioinform. Comput. Biol.* 3, 551–585
- Gasteiger, E. *et al.* (2003) ExpASY: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31, 3784–3788
- Backes, C. *et al.* (2005) GraBCas: a bioinformatics tool for score-based prediction of caspase- and granzyme-B-cleavage sites in protein sequences. *Nucleic Acids Res.* 33, W208–W213
- Garay-Malpartida, H.M. *et al.* (2005) CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics* 21 (Suppl. 1), i169–i176
- Lopez-Otin, C. and Overall, C.M. (2002) Protease degradomics: a new challenge for proteomics. *Nat. Rev. Mol. Cell Biol.* 3, 509–519
- Cullen, S.P. *et al.* (2007) Human and murine granzyme B exhibit divergent substrate preferences. *J. Cell Biol.* 176, 435–444
- Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919
- Rechsteiner, M. and Rogers, S.W. (1996) PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* 21, 267–271
- Belizario, J.E. *et al.* (2008) Coupling caspase cleavage and proteasomal degradation of proteins carrying PEST motif. *Curr. Protein Pept. Sci.* 9, 210–220
- Cheng, J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* 33, W72–W76
- Lamkanfi, M. *et al.* (2002) Alice in caspase land. A phylogenetic analysis of caspases from worm to man. *Cell Death Differ.* 9, 358–361
- Stennicke, H.R. *et al.* (2000) Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1, 3, 6, 7 and 8. *Biochem. J.* 350, 563–568
- Talanian, R.V. *et al.* (1997) Substrate specificities of caspase family proteases. *J. Biol. Chem.* 272, 9677–9682
- Thornberry, N.A. *et al.* (1997) A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J. Biol. Chem.* 272, 17907–17911
- Varshavsky, A. (1992) The N-end rule. *Cell* 69, 725–735
- Park, S.H. *et al.* (2004) Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J. Radiol.* 5, 11–18
- Mahrus, S. *et al.* (2008) Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* 134, 866–876