Published in final edited form as:

Proteomics. 2014 January ; 14(1): 42–50. doi:10.1002/pmic.201300416.

SEQUENCE-DERIVED STRUCTURAL FEATURES DRIVING PROTEOLYTIC PROCESSING

Alexander A. Belushkin¹, Dmitry V. Vinogradov², Mikhail S. Gelfand^{1,2}, Andrei L. Osterman³, Piotr Cieplak^{3,*}, and Marat D. Kazanov^{2,*}

¹M.V. Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Russia

²A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Science, Russia

³Sanford-Burnham Medical Research Institute, La Jolla, CA, USA

Abstract

Proteolytic signaling, or regulated proteolysis, is an essential part of many important pathways such as Notch, Wnt, and Hedgehog. How the structure of the cleaved substrate regions influences the efficacy of proteolytic processing remains underexplored. Here, we analyzed the relative importance in proteolysis of various structural features derived from substrate sequences using a dataset of more than five thousand experimentally verified proteolytic events captured in CutDB. Accessibility to the solvent was recognized as an essential property of a proteolytically processed polypeptide chain. Proteolytic events were found nearly uniformly distributed among three types of secondary structure, although with some enrichment in loops. Cleavages in α -helices were found to be relatively abundant in regions apparently prone to unfolding, while cleavages in β -structures tended to be located at the periphery of β -sheets. Application of the same statistical procedures to proteolytic events divided into separate sets according to the catalytic classes of proteases proved consistency of the results and confirmed that the structural mechanisms of proteolysis are universal. The estimated prediction power of sequence-derived structural features, which turned out to be sufficiently high, presents a rationale for their use in bioinformatic prediction of proteolytic events.

INTRODUCTION

Proteolysis, which is irreversible post-translational modification via hydrolysis of a peptide bond, has been known mainly for its primary role in protein degradation relevant to food digestion and intracellular protein turnover [1]. Recently, this viewpoint has been revised following a demonstration of the important role of proteolysis as a signaling mechanism in numerous vital biological processes [2]. Proteases, enzymes catalyzing hydrolysis of peptide bonds, together with their substrates compose complex proteolytic signaling networks in many eukaryotic organisms [3]. Signal transduction via the proteolytic network implies activation or deactivation of physiological substrates by proteases through one or several proteolytic cleavages. This process is known as proteolytic processing or regulated/limited proteolysis [4]. The size of the proteolytic network in an organism can be large; for example, in humans more than 570 proteases have been identified to date [5,6]. However, most protease substrates are still unknown. Identification of members of the proteolytic network is important due to protease involvement in many biological processes, such as apoptosis, development, and cell proliferation. However, despite recent advances in technology [7,8],

*Corresponding authors. pcieplak@sanfordburnham.org; mkazanov@gmail.com.

experimental searches for and validations of protease substrates are still very labor-intensive. These processes can be substantially facilitated by a hypothesis-driven search strategy, guided by a bioinformatic prediction of protease substrates. Bioinformatic prediction implies calculation of the probability of a proteolytic event for a particular protease and its candidate substrate based on information about protease specificity and substrate sequence and/or 3D structure.

An accurate bioinformatic prediction of proteolytic events requires a deep understanding of the proteolysis mechanisms. Traditionally, the main attention in bioinformatic prediction methods was devoted to exploiting the primary specificity of proteases, that is, the specific amino acid content of the substrate's polypeptide chain around the cleaved bond. The primary specificity, which is a constant and unique characteristic of each protease [9], can be recognized by a variety of experimental methods [10] and captured in the form of predictive models. Primary specificity models such as position-specific scoring matrices (PSSM) [11,12] successfully proved their applicability for the prediction of proteolytic events [13–15], especially for proteins in denatured conditions. However, it has become clear that for a substrate in its native state, specific structural properties of cleaved regions are at least equally important for the proteolysis to occur [16]. To date, the impact on proteolysis of structural features of the substrate's peptide bonds and the relative importance of these features remains insufficiently explored. Indeed, one of the key studies on this topic was published more than 15 years ago and analyzed a limited set of known proteolytic events [17], whereas recently the number of experimentally validated proteolytic events has substantially increased. Several bioinformatic methods for prediction of proteolytic events, which mainly rely on primary specificity, additionally include various structural features of substrates to the prediction model [18,19]. However, most relevant structural features could neither be preselected due to the lack of sufficient knowledge nor extracted from the trained predictors, as most machine-learning models are the “black box” predictors.

Our recent study [20] investigated the importance of substrate structure in proteolysis using a large collection of experimentally verified proteolytic events captured in CutDB [21]. While in that study we analyzed only substrates with known 3D structures, the size of the analyzed dataset, ~200 substrates, allowed us to consistently estimate the significance of structural features and their relative ranking. Recent advances in bioinformatic prediction methods encouraged us to reinvestigate the conclusions from [20] using the complete CutDB set of substrates (2,062 entries) with deriving structural features from substrate sequences. Indeed, it has been reported that current methods of structural prediction have achieved up to >80% quality of prediction [22]. Here, we applied 11 state-of-the-art bioinformatic methods to predict structural properties of all substrates available in CutDB. We investigated a secondary structure context of proteolytic events; calculated correlations of proteolytic events and sequence-derived structural features of cleaved regions; and estimated their capabilities in prediction of proteolytic events. We also assessed whether the structural determinants of proteolysis that were found are universal across protease classes. To our knowledge, this study is the largest-to-date statistical analysis of the structural context of experimentally verified proteolytic events.

MATERIALS AND METHODS

Data

The analyzed set of 5,704 experimentally observed proteolytic events was taken from CutDB [21]. CutDB is a part of the PMAP web-environment that integrates five databases reflecting different aspects of proteolysis, such as proteolytic events, pathways, enzymes, substrates, and profiles. Each proteolytic event taken from CutDB was uniquely identified by the substrate sequence, protease, and position of the cleaved peptide bond. All structural

features of substrates were derived based on sequence information only. The entire dataset including predicted structural features is available at http://mg.uncb.iitp.ru/SupplementalData/Belushkin_et_al_Proteomics_2014/rawdata.tgz.

Bioinformatic tools

Due to the large size of the analyzed dataset, we used only software packages that had versions adapted for local installation. Solvent accessibility was predicted using Sable [23], NetSurfP [24], SPINE X [25], ACCpro [26] and PROFphd [27]. Secondary structure was predicted by PSIPRED [28], PROFphd [27], Sable [23], NetSurfP [24], and SPINE X [25]. Disordered regions were predicted using DISOPRED [29], DisEMBL [30], and IUPred [31]. All these programs are freely available. Computations were performed with the programs' default parameters on a Linux-based computational cluster. The cluster consisted of 10 nodes, each with 8 CPU cores (2 Intel L5520 CPUs) and 24GB of RAM.

Statistical analysis

Statistical significance was estimated by the chi-squared test, as most structural features were categorical. Numerical features were transformed into nominal variables using thresholds that were selected by minimization of the significance values. The chi-squared test was applied to the combined set of substrates, i.e., the number of cleavage sites and the ratios of particular values of structural features were summed over all substrates. Statistical significance was calculated independently for each applied bioinformatic method without any statistical filters within the respective program. For a particular cleavage event, all associated structural features of the cleaved substrate region were defined with respect to the P1 position (in notation of Schechter and Berger [32]), as this position was recognized earlier as the most important in the structural sense [20]. The significance of differences in frequencies of cleavage sites between central and peripheral parts of the β -strands were estimated using the Student's T-test.

Prediction capabilities of structural features were estimated using conventional machine-learning evaluation metrics. The cleaved and uncleaved peptide bonds were considered as positive and negative classes, respectively, of a binary classification problem. We did not use Accuracy, as it is not an appropriate evaluation metric when the size of the negative class is larger than the size of the positive class. Instead, we used positive class-oriented evaluation metrics, Precision and Sensitivity, and the generic threshold-less Area Under ROC curve metric. All estimates were calculated for individual substrates and were then averaged.

RESULTS

The analyzed dataset of documented proteolytic events taken from CutDB [21] comprised 5,704 unique events. The majority of events were linked to the regulatory proteases of mammalian origin belonging to the seven catalytic protease classes. The distribution of the analyzed proteolytic events among the protease classes is presented in Figure 1.

Secondary structure context of proteolytic events

The question of what type of a protein secondary structure is susceptible to proteolysis still remains controversial. Thus, the long-established opinion that cleavages are possible only in loops was recently challenged by findings that registered the occurrence of cleavage sites in α -helices [16,33] and β -sheets [34,35]. To clarify this issue, we analyzed the secondary structure context of 4,576 unique cleavage positions (fewer than the total number of proteolytic events, as more than one protease could cleave the same site) taken from CutDB. To elucidate a general trend and to filter out biases that could be introduced by prediction

methods, we applied six different methods of secondary-structure prediction (Figure 2, Table S1a–c). While the methods differ in the predicted ratios of loops, α -helices, and β -sheets, the distribution of proteolytic events by types of secondary structure is consistent. As expected, we found that cleavages tend to occur more often in loops (3.08 ± 0.15 events per 1K residues, averaged across methods) than in two other types of secondary structure (Table S1c). Nevertheless, the frequencies of proteolytic events in predicted α -helical and β -sheet regions are comparable with the frequency observed in loops. Indeed, the values averaged across the prediction methods are 2.05 ± 0.23 and 1.9 ± 0.41 per 1K residues, respectively. This confirms recent observations [16,33,35] that proteolytic events in α -helices and β -sheets are common, contrary to earlier opinions that they are very rare or even impossible [36,37].

We further analyzed whether regions of α -helices where cleavages occur differ from other regions of α -helices. Modeling studies argue that local unfolding of a protein chain around the cleavage site is required for proteolysis [37]. To check whether proteolytic events in α -helices occur in regions susceptible to unfolding, we compared the confidence values of the secondary-structure prediction in known cleaved and uncleaved positions of α -helices. These confidence values, which are provided along with the predicted type of secondary structure by most prediction methods, apparently reflect the tendency of a particular helical region to unfold. We observed that the confidence values in the cleavage positions, predicted by most of the prediction methods, are significantly lower than in other positions of helical regions. Indeed, the p-values of these differences calculated from the Student's T-test for different prediction methods ranges from $3.91\text{e-}5$ for PROFphd to $1.02\text{e-}42$ for SPINE X (Figure 3). This result suggests that the occurrence of proteolytic events in helices is possible in regions of helices amenable to unfolding.

In our previous study performed on substrates with known 3D structure [20], we observed that cleavages in β -sheets occur almost exclusively in the periphery of β -sheets if they were exposed to the solvent. There were only rare cleavages in internal areas of β -sheets; moreover, all of them were close to the C- or N-terminus of the substrate. Unfortunately, to our knowledge, no bioinformatic methods are able to predict the relative order of β -strands within a β -sheet except for 3D structure modeling, which we are not considering here. Thus, we cannot predict which strands belong to the periphery of β -sheets and which are internal ones. However, we were able to analyze whether cleavages tend to be located in the terminal or internal regions of a strand. Each predicted β -strand was divided into equal-length terminal and central parts, and the distribution of cleavage sites in these parts of the β -strands was analyzed. The results presented in Figure 4 and Table S2 clearly show that cleavages tend to be located closer to the β -strand termini but not in the central part. Although the same effect was also clearly observed for substrates with known 3D structures [20], there is a possibility that in this analysis the effect, at least in part, stems from the loops' misprediction. To clarify this issue, this analysis should be repeated when more 3D structures of substrates are available from the X-ray crystallography or NMR spectroscopy experiments.

Correlation between proteolytic events and sequence-derived structural features of substrates

Along with the secondary-structure context of proteolytic events, we characterized cleavage regions by other sequence-derived structural features that may determine susceptibility of a polypeptide chain to proteolysis. We analyzed correlations between the proteolytic events and the degree of solvent accessibility and disorder in the surrounding region. We calculated values of statistical significance of observed effects and also compared them with the significance values of distribution of cleavage regions by secondary-structure type. Again, to avoid biases of any single approach, we calculated sequence-derived structural features

using several bioinformatic methods. The calculated correlations of proteolytic events with the structural properties of polypeptide chain in the vicinity of cleavage sites are presented in Figure 5 and Table S3. The highest correlation is observed for the absolute solvent accessibility of cleaved regions, followed by an approximately equal significant relative solvent accessibility and disordered regions. A lower but still substantial significance is observed for loops, whereas α -helices and β -sheets demonstrate only a modest significance. While significance values obtained by different methods from the same category are not similar, they are sufficiently close enough to support the ordering of the features. Indeed, it can be seen from Figure 5 that all significance values from any category of structural features are strictly smaller or larger than values in another category. The only exception is a relative solvent accessibility and disorder that exhibit an approximately similar significance. The methods that demonstrate the highest significance values in their categories for solvent accessibility, disorder, loops, α -helices, and β -sheets are PROFphd, DISOPRED, NetSurfP, Sable, and again NetSurfP, respectively.

Cleavage site prediction capabilities of sequence-derived structural features

While statistical analysis gives us a general picture of the relative importance of structural features in the susceptibility of polypeptide chain to proteolysis, a more-detailed view on the capabilities of sequence-derived structural features in the prediction of cleavage sites can be obtained using machine-learning metrics. For the same set of structural features, we calculated four machine-learning metrics—Area Under ROC Curve (AUC), F-score, Sensitivity, and Specificity—which allowed us to comprehensively estimate the power of sequence-derived structural features in the prediction of cleavage sites. The relative ranking of structural features obtained using AUC (Figure 6, Table S4), which is the most general threshold-less machine-learning metric, is completely consistent with the results of the statistical analysis presented above. The calculated AUC values range from the highest value of 0.67 for the absolute solvent accessibility, to the lowest value of 0.52 (0.48 before reversion ($AUC_{rev} = 1 - AUC$)) for β -sheets. In this analysis, we accounted for the direction of an observed effect. As expected from the analysis of the secondary-structure preferences of cleavage sites, we found lower than 0.5 AUC values, i.e., negative-class (non-cleavage site) prediction abilities for α -helices and β -sheets. This result means that cleavages are more probable outside of such structured regions and, hence, negations of these structural features should be used for prediction of cleavage sites.

Additional details can be provided by the analysis of Sensitivity and Specificity metric values and their aggregate metric, F-score (Table S4b–d). These are threshold-based metrics; thus, for continuous features, we applied the thresholds optimizing the F-score metric, while for categorical features a threshold implicitly exists. For example, the sensitivity and specificity of the absolute solvent accessibility as a prediction property are ~87% and ~70%, respectively, which means that using this structural feature alone it is possible to predict the majority of cleavage sites with ~30% false-positive rate. It is worth noting that in spite of a sufficiently large number of false positives, in the course of the development of a bioinformatic prediction method they can be filtered further using a known primary specificity of proteases. A primary specificity of regulatory proteases involved in proteolytic processing is usually very high.

Most methods that predict binary protein structural features accompany them with the quantitative estimates of the prediction confidence. We analyzed whether using these confidence values instead of binary features themselves would improve the prediction quality. The results for the threshold-based metrics (F-score, Sensitivity, and Specificity) demonstrated that using the confidence values instead of corresponding binary features improves the prediction of cleavage sites (Table S4b–d). Thus, all threshold-based metrics calculated using confidence values for disorder, loops, and negations of α helices and β -

sheets exceeded metrics calculated for corresponding binary features. Values of AUC, the only threshold-less metric, calculated for predictions that used confidence values were larger than AUC values calculated for binary features of disorder, α -helices, and β -sheets but were smaller for loops. We can attribute this partial inconsistency to artifacts of the AUC calculation using a small number of points. Indeed, calculation of the AUC value for binary features is performed based on three points only.

The methods demonstrating the best results for individual structural categories in terms of an AUC metric are SPINE X (absolute solvent accessibility), Sable (relative solvent accessibility), DISOPRED (disorder), PSIPRED (loops), PSIPRED (α -helices), and NetSurfP (β -sheets). In the case of utilization of binary feature's confidence values for cleavage site prediction, the best methods for disorder, loops, α -helices, and β -sheets are IUPred, Sable, PSIPRED, and NetSurfP, respectively. Notably, although statistical and machine-learning analyses were absolutely consistent in the ranking of structural categories in general, the ranking of methods within the particular categories varied.

Universality of proteolytic structural determinants across different protease catalytic classes

To date, seven types of protease catalytic mechanisms are known [38]. *A priori*, it is not clear whether significance and ranking of proteolytic structural determinants coincide across protease catalytic classes. The size of datasets used in this study was sufficient for separate analysis of four protease classes, namely, aspartic, cysteine, serine, and metalloproteases. We recalculated the machine-learning metrics used above for the same structural features using at least two bioinformatic methods for each feature. The results presented in Figure 7 do not completely answer the raised question. At first sight, one can conclude that the relative importance of structural features for the aspartic and metallo-proteases is different from the general trend observed above. However, careful examination of the lists of aspartic and metallo-protease substrates showed a prevalence of substrates that could introduce biases to the results due to lack of stable three-dimensional structure, i.e., substrates that are unstructured and transmembrane proteins as well as proteins that were incubated with proteases during a prolonged period. To exclude the influence of unstructured substrates we selected aspartic and metallo-protease substrates with known 3D structures and calculated machine-learning metrics for this small subset of substrates (data not shown). Obtained results were in accordance with the earlier observed trend. Thus, this supports our initial hypothesis about the universality of structural determinants of regulatory proteolysis across protease classes. Further investigation is required and would be done only when more substrates of aspartic and metallo-proteases with known 3D structure are available.

Discussion

In this study, we aimed to elucidate structural properties of protein regions susceptible to proteolysis by performing statistical analysis on the largest existing dataset of documented regulatory proteolytic events in mammalian proteins. Only a small fraction of this dataset (~10%) contains substrates with known 3D structure, which we have analyzed in a previous study [20]. We relied on bioinformatic tools to analyze the structure of cleavage regions based only on sequences of substrates. Firstly, we analyzed sequence-derived secondary structures of cleaved protein regions. We found that the prevalence of proteolytic events occurred in loop regions. However, we observed a substantial number of cleavages in α -helices and β -sheets, thus confirming the recent findings [16,33,35] and contradicting earlier opinions about the distribution of proteolytic events among different types of secondary structure. Additionally, we showed for α -helices that the cleavages seem to occur in regions prone to unfolding and for β -sheets that the cleavage sites apparently are possible only in the

periphery of these structures. The obtained results suggest a nearly uniform distribution of proteolytic events among the secondary structure types, thereby indicating a modest usefulness of secondary structure in prediction of cleavage sites.

To precisely estimate the capability of other sequence-derived structural descriptors in prediction of proteolytic events, we applied a statistical approach. Solvent accessibility is the most significant structural feature associated with susceptibility of protein regions to proteolysis. Other structural features also showed the same ranking as in [20] with the exception of α -helices. This type of secondary structure demonstrated low negative association with the cleaved regions.

The large size of the analyzed dataset allowed us to apply the same statistical procedure to subsets of the original dataset corresponding to four protease catalytic classes in order to check, for the first time, whether the observed correlation depends on the type of protease catalytic mechanism. Substrates of cysteine and serine proteases seem to follow all general trends, unlike substrates of aspartic and metallo-proteases. However, the results obtained for the latter may be biased by a large fraction of transmembrane and unstructured proteins. Analysis of substrates with known 3D structures for aspartic and metallo-proteases supports the existence of universal structural determinants of proteolysis.

We confirmed at a larger scale and refined the conclusions from our earlier study, performed for a limited dataset of substrates with known 3D structure [20]. This is important, as some of these observations contradicted the established views. We demonstrated that the primary indicator of susceptibility of protein regions to proteolysis is accessibility of a polypeptide chain to protease but not local interactions (secondary structure). The precise estimates of the prediction capabilities of sequence-derived features were obtained, showing that these features alone are capable of excluding from candidate cleavage bonds at least half of the peptide bonds that are inaccessible for proteases. Thus, even lacking the 3D structure of a substrate, it is theoretically possible to achieve an acceptable quality of cleavage site prediction by combining knowledge about the primary specificity of the protease and sequence-derived structural properties of substrates. Overall, this study shed light on the structural mechanism of proteolytic processing and set the foundation for the development of bioinformatic methods for prediction of proteolytic events that may drastically reduce the amount of effort required for the experimental identification of the members of a proteolytic network.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Yoshinobu Igarashi, Boris Ratnikov, Alexey Eroshkin, and Jeff Smith for useful discussions. AAB, DVV, MSG, and MDK were partially supported by the Ministry of Education and Science of the Russian Federation, projects #8049 and #8135. PC is supported by NIH grant R01GM098835.

References

1. Barrett, AJ.; Rawlings, ND.; Woessner, JM. Handbook of Proteolytic Enzymes: Cysteine, serine and threonine peptidases. Vol. 2. Elsevier Acad. Press; 2004.
2. Turk B. Targeting proteases: successes, failures and future prospects. Nat Rev Drug Discov. 2006; 5:785–799. [PubMed: 16955069]
3. Turk B, Turk D, Turk V. Protease signalling: the cutting edge. EMBO J. 2012; 31:1630–1643. [PubMed: 22367392]

4. Lopez-Otin C, Overall CM. Protease degradomics: a new challenge for proteomics. *Nat Rev Mol Cell Biol.* 2002; 3:509–519. [PubMed: 12094217]
5. Quesada V, Ordóñez GR, Sánchez LM, Puente XS, López-Otín C. The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res.* 2009; 37:D239–43. [PubMed: 18776217]
6. Puente XS, Sanchez LM, Overall CM, Lopez-Otin C. Human and mouse proteases: a comparative genomic approach. *Nat Rev Genet.* 2003; 4:544–558. [PubMed: 12838346]
7. Kasperkiewicz P, Gajda AD, Drg M. Current and prospective applications of non-proteinogenic amino acids in profiling of proteases substrate specificity. *Biol Chem.* 2012; 393:843–51. [PubMed: 22944686]
8. Van den Berg BHJ, Tholey A. Mass spectrometry-based proteomics strategies for protease cleavage site identification. *Proteomics.* 2012; 12:516–29. [PubMed: 22246699]
9. Beynon, RJ.; Bond, JS. *Proteolytic Enzymes: A Practical Approach.* Oxford University Press; 2000.
10. Diamond SL. Methods for mapping protease specificity. *Curr Opin Chem Biol.* 2007; 11:46–51. [PubMed: 17157549]
11. Turk BE, Huang LL, Piro ET, Cantley LC. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat Biotechnol.* 2001; 19:661–667. [PubMed: 11433279]
12. Boyd SE, Pike RN, Rudy GB, Whisstock JC, de la Banda M. PoPS: a computational tool for modeling and predicting protease specificity. *J Bioinform Comput Biol.* 2005; 3:551–585. [PubMed: 16108084]
13. Backes C, Kuentzer J, Lenhof HP, Comtesse N, Meese E. GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Res.* 2005; 33:W208–213. [PubMed: 15980455]
14. Garay-Malpartida HM, Occhiucci JM, Alves J, Belizario JE. CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics.* 2005; 21 (Suppl 1):i169–176. [PubMed: 15961454]
15. Wee LJ, Tan TW, Ranganathan S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics.* 2007; 23:3241–3243. [PubMed: 17599937]
16. Timmer JC, Zhu W, Pop C, Regan T, et al. Structural and kinetic determinants of protease substrates. *Nat Struct Mol Biol.* 2009; 16:1101–1108. [PubMed: 19767749]
17. Hubbard SJ, Beynon RJ, Thornton JM. Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.* 1998; 11:349–359. [PubMed: 9681867]
18. Barkan DT, Hostetter DR, Mahrus S, Pieper U, et al. Prediction of protease substrates using sequence and structure features. *Bioinformatics.* 2010; 26:1714–1722. [PubMed: 20505003]
19. Song J, Tan H, Shen H, Mahmood K, et al. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics.* 2010; 26:752–760. [PubMed: 20130033]
20. Kazanov MD, Igarashi Y, Eroshkin AM, Cieplak P, et al. Structural determinants of limited proteolysis. *J Proteome Res.* 2011; 10:3642–3651. [PubMed: 21682278]
21. Igarashi Y, Eroshkin A, Gramatikova S, Gramatikoff K, et al. CutDB: a proteolytic event database. *Nucleic Acids Res.* 2007; 35:D546–549. [PubMed: 17142225]
22. Zhang H, Zhang T, Chen K, Kedarisetti KD, et al. Critical assessment of high-throughput standalone methods for secondary structure prediction. *Brief Bioinforma.* 2011; 12:672–688.
23. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins.* 2005; 59:467–475. [PubMed: 15768403]
24. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol.* 2009; 9:51. [PubMed: 19646261]
25. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem.* 2012; 33:259–267. [PubMed: 22045506]
26. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* 2005; 33:W72–76. [PubMed: 15980571]

27. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*. 1994; 19:55–72. [PubMed: 8066087]
28. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics*. 2000; 16:404–405. [PubMed: 10869041]
29. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 2004; 337:635–645. [PubMed: 15019783]
30. Linding R, Jensen LJ, Diella F, Bork P, et al. Protein disorder prediction: implications for structural proteomics. *Structure*. 2003; 11:1453–1459. [PubMed: 14604535]
31. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. 2005; 21:3433–3434. [PubMed: 15955779]
32. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun*. 1967; 27:157–162. [PubMed: 6035483]
33. Mahrus S, Trinidad JC, Barkan DT, Sali A, et al. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell*. 2008; 134:866–876. [PubMed: 18722006]
34. Massucci MT, Giansanti F, Di Nino G, Turacchio M, et al. Proteolytic activity of bovine lactoferrin. *Biometals*. 2004; 17:249–255. [PubMed: 15222473]
35. Van Damme P, Maurer-Stroh S, Plasman K, Van Durme J, et al. Analysis of protein processing by N-terminal proteomics reveals novel species-specific substrate determinants of granzyme B orthologs. *Mol Cell Proteomics*. 2009; 8:258–272. [PubMed: 18836177]
36. Hubbard SJ, Campbell SF, Thornton JM. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J Mol Biol*. 1991; 220:507–530. [PubMed: 1856871]
37. Hubbard SJ, Eisenmenger F, Thornton JM. Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci*. 1994; 3:757–768. [PubMed: 7520312]
38. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res*. 2012; 40:D343–350. [PubMed: 22086950]

Figure 1.

Figure 2.

Figure 3.

Figure 4.

Figure 5.

Figure 6.

Figure 7.