Published in final edited form as:
J Proteome Res. 2011 August 5; 10(8): 3642–3651. doi:10.1021/pr200271w.

Structural determinants of limited proteolysis

Marat D. Kazanov[1,2,#], **Yoshinobu Igarashi**[3,#], **Alexey M. Eroshkin**[1], **Piotr Cieplak**[1], **Boris Ratnikov**[1], **Ying Zhang**[1], **Zhanwen Li**[1], **Adam Godzik**[1], **Andrei L. Osterman**[1,*], and **Jeffrey W. Smith1**[*]

[1] Sanford–Burnham Medical Research Institute, 10901 North Torrey Pines Rd, 92037 La Jolla, CA, USA

[2] Institute for Information Transmission Problems RAS, Bolshoi Karetnyi per. 19, Moscow, 127994, Russia

[3] National Institute of Biomedical Innovation, 7-6-8 Asagi Saito, Ibaraki-City, Osaka, Japan.

Abstract

Limited or regulatory proteolysis plays a critical role in many important biological pathways like blood coagulation, cell proliferation, and apoptosis. A better understanding of mechanisms that control this process is required for discovering new proteolytic events and for developing inhibitors with potential therapeutic value. Two features that determine the susceptibility of peptide bonds to proteolysis are the sequence in the vicinity of the scissile bond and the structural context in which the bond is displayed. In this study we assessed statistical significance and predictive power of individual structural descriptors and combination thereof for the identification of cleavage sites. The analysis was performed on a dataset of >200 proteolytic events documented in CutDB for a variety of mammalian regulatory proteases and their physiological substrates with known 3D structures. The results confirmed the significance and provided a ranking within three main categories of structural features: exposure > flexibility > local interactions. Among secondary structure elements, the largest frequency of proteolytic cleavage was confirmed for loops and lower but significant frequency for helices. Limited proteolysis has lower albeit appreciable frequency of occurrence in certain types of β-strands, which is in contrast with some previous reports. Descriptors deduced directly from the amino acid sequence displayed only marginal predictive capabilities. Homology-based structural models showed a predictive performance comparable to protein substrates with experimentally established structures. Overall, this study provided a foundation for accurate automated prediction of segments of protein structure susceptible to proteolytic processing and, potentially, other post-translational modifications.

Keywords

proteolysis; proteolytic processing; limited proteolysis; regulatory proteolysis; protease; cleavage site; cleavage site prediction

[*] Corresponding authors: Andrei L. Osterman 10901 North Torrey Pines Rd, 92037 La Jolla, CA, USA; 858-646-3100 ext. 5296 (phone) 858-795-5249 (fax) osterman@sanfordburnham.org Jeffrey W. Smith 10901 North Torrey Pines Rd, 92037 La Jolla, CA, USA; 858-646-3121 (phone) 858-646-3192 (fax) jsmith@sanfordburnham.org.

[#] These authors contributed equally to this work.

Introduction

Proteolysis is one of the few irreversible post-translational modifications. Limited proteolysis, also known as proteolytic processing [1], has a regulatory role in almost all biological pathways, including blood coagulation, cell proliferation, and cell death [2]. Proteolytic events are involved in progression of many diseases including cancer, inflammation, and atherosclerosis [3], so many proteases and their substrates are important diagnostic and therapeutic targets [4]. In contrast to exhaustive protein degradation, which is required for intracellular protein turnover or utilization of dietary proteins, regulatory proteolysis typically affects a specific peptide bond in a target protein and modulates the biological activity of the resulting fragments. Regulatory proteolytic events may lead to a loss or gain of function, as it was initially recognized for the proteolytic activation of zymogens [5].

Our knowledge of mechanisms that control the precision and extent of proteolytic processing is far from complete. The key features of native proteins that determine their susceptibility to limited proteolysis are: (i) the presence of an amino acid motif (typically a stretch of 1 to 5 contiguous residues) that conforms to the recognition specificity of the enzymatic cleft; (ii) the 3D structural context of these sites; (iii) required by some regulatory proteases, the presence of the so-called exosites that are located away from the primary cleavage site but which contribute additional specific enzyme–substrate interactions [6]; and (iv) finally, the co-expression of protease and substrate in both space and time, an important factor in controlling regulatory proteolysis *in vivo*. Among these determinants, sequence-based substrate specificity has received the most attention. Several established methodologies, such as phage display, peptide libraries, and genome-scale degradomics, have been successfully applied for the detailed mapping of primary substrate specificity of many regulatory proteases [7]. The results of these studies captured in the form of consensus motifs or positional-specific scoring matrices (PSSM) [8, 9] are broadly used for the prediction and interpretation of proteolytic events [9,10,11].

Despite their apparent utility, these sequence-based models, when used in isolation, often fall short of accurately describing and predicting proteolysis of proteins in their native conformation. In fact, the position of cleavage sites within the substrates’ intact 3D structure plays an equally, if not a more, prominent role [12]. Moreover, among the features listed above, such structural requirements appear to be of a truly universal nature, shared by many distinct proteases. Therefore, the objective of this study was to elucidate the structural requirements of protein substrates for cleavage by a representative subset of proteases, and thereby enhance our predictive capabilities for the entire spectrum of regulatory proteolytic enzymes.

Several structural features of protein substrates are thought to contribute to their susceptibility to proteolysis [13]. However, the relative importance of individual features remains a subject of controversy. In one of the early studies of limited proteolysis by subtilisin, a bacterial protease with broad substrate specificity, surface regions of a protein substrate having high segmental mobility were found to be particularly susceptible to proteolysis [14]. In a different study, accessibility rather than flexibility was recognized as an essential structural determinant of limited proteolysis [15]. Although both studies noted the tendency of proteolytic events to occur outside of regions with well-defined secondary structure, several cleavage sites in alpha-helices have been reported [15]. Early modeling efforts led to a conclusion that local conformational changes required for proteolysis are readily achievable for extended flexible loops, are theoretically possible for helices, but are implausible for β-sheets [16, 17]. In the first systematic survey of 32 published proteolytic events, Hubbard *et al.* [13] evaluated the significance of many structural descriptors in three general categories: exposure (solvent accessibility, protrusion, and packing), flexibility (B-Factor), and local interactions (secondary structure and hydrogen bonding). Although this analysis was performed on a relatively small dataset representing 8 proteases (mostly bacterial and digestive, rather than regulatory, enzymes) and 17 protein substrates, it provided a conceptual framework on the structural features of substrates that regulate proteolysis, which is still the gold standard today.

Since Hubbard's study in 1998, there have been massive increases in information and considerable advances in computational biology. These include a sevenfold increase in the number of 3D structures in PDB [18] and vastly improved structural modeling techniques [19]. In addition, we established CutDB, a database of >3,000 proteolytic events captured from original publications with a strong emphasis on mammalian regulatory proteases [20]. Together, these advances motivated us to revisit the conclusions that have been drawn on proteolysis using much smaller sets of data. Our main goals were to assess the contribution of structural features of substrates to proteolysis and to estimate the predictive capabilities of these features (or combinations of features).

While most of the conclusions we draw are generally consistent with those of prior studies, our analysis, for the first time, provides a ranking within the three main categories of structural features: exposure, flexibility, and local interactions. The detailed analysis of secondary structure context of proteolytic events confirmed their highest frequency in loops followed by helices. However, in contrast with some previous reports, we observe an appreciable (albeit lower) frequency of cleavage in β-strands. We built a gallery of 3D structures of analyzed protein substrates with visualized cleavage sites, which allowed us to reveal special structural features of such “non-canonical” proteolytic events. Overall, this study provides a sound statistical foundation for the accurate automated prediction of which segments of proteins are susceptible to proteolysis (and, potentially, other post-translational modifications [21]).

Methods

Datasets representing documented regulatory proteolytic events in mammalian proteins with known or modeled 3D structures were generated as follows. Information about documented proteolytic events was extracted from CutDB [20] and defined as a unique combination of three elements: (i) protein substrate (NCBI Accession [22]), (ii) protease (MEROPS ID [23]), and (iii) position of a cleavage site (from original publications available in PubMed). Structural information for a subset of protein substrates with experimentally solved 3D structure was obtained by scanning the PDB [24] using BLAST [25] with a 95% identity threshold. Homology-based structural models were obtained by applying the automated modeling protocol as implemented at the JCMM server [26]. All proteolytic sites were mapped by a script (available by request) onto respective 3D structures and models and visualized by highlighting P1 positions (according to Schechter and Berger notation [27]) using Chimera software (see Fig. 1 for examples) [28]. Based on visual inspection of these graphical representations, several outliers with numerous cleavage sites by the same protease within the same protein were identified and excluded from datasets as likely resulting from denaturation (and/or overdigestion) under experimental conditions of respective studies. Indeed, most of such cases were observed for hemoglobin, myoglobin, and similar model protein substrates historically used to assess primary specificity of proteases under the conditions of complete or partial denaturation. After manual curation, our final datasets included 80 non-redundant protein substrates with experimentally solved structures (217 associated proteolytic events) and 43 proteins with high-quality structural models (98 proteolytic events). A complete list of documented and analyzed proteolytic events is provided in Supplemental Files 1 and 2. A gallery of respective 3D structures and structural models with visualized cleavage sites is provided in Supplemental Files 3 and 4.

Structural descriptors derived from experimentally solved and modeled structures included solvent accessibility, hydrogen bonding, torsion angles, and secondary structures calculated using DSSP software [29]. DSSP secondary structure notations (8 classes) were transformed to a simplified "loop/helix/β-sheet" classification (3 classes) following the approach applied in CASP [30]. Naccess [31] was used as an alternative tool for calculating solvent accessibility. This program also allowed us to compute solvent accessibility for main chain, side-chain, non-polar side-chain, and polar side-chain atoms separately. A similar but distinct descriptor, molecular surface accessibility, was calculated using the MSMS package [32]. Packing was calculated using an original method by Nishikawa and Ooi [33] for 8-Å and 14-Å spheres (default value for most calculations). Protrusion and depth indexes were calculated using CX [34] and DPX [35] algorithms, respectively. A complete list of structure-derived descriptors is provided in Supplemental File 5. The only descriptors that could not be assessed for the structure models dataset were B-Factors and disordered regions obtained directly from PDB files. Disordered regions were defined by comparing SEQRES and ATOM records of the PDB files as described in [36]. A smaller set of structural descriptors derived solely from amino acid sequences of all proteins in both datasets included: secondary structures predicted using Psipred [37], predicted disordered regions computed by Disopred [38], and predicted solvent accessibility calculated using Sable [39]. Normalization of values of all types of descriptors for all individual protein substrates was performed using an expression:

$$z_{i,j}=\frac{x_{i,j}-\mu_j}{\sigma_j}$$

where $z_{i,j}$ is the normalized value of the structural descriptor calculated for the *i*-th residue of the *j*-th substrate, $x_{i,j}$ is the raw structural descriptor value of the *i*-th residue of the *j*-th substrate, μ_j is the mean value of the descriptor over all residues of the *j*-th substrate, and σ_j is the standard deviation. Calling and parsing of needed tools was performed using custom Python scripts making use of BioPython libraries [40]. Aggregated data were stored in Oracle XE [41]. Categorial (nominal) descriptors, e.g. secondary structure, were transformed into binary indicator variables. Closeness of the descriptor distributions to normality was checked by Kolmogorov-Smirnov test and by visual analysis of histograms. Statistical analysis was performed using Oracle XE built-in statistical functions.

Evaluation metrics

To estimate the predictive power of each structural descriptor, we used a single-variable binary classification approach where the independent variable is a particular descriptor and two classes of objects are cleavable and uncleavable peptide bonds (positive and negative class, respectively). Our choice of the specific metrics for evaluation of the quality of classification was dictated by two special features of the considered datasets.

The first feature is the size imbalance of the two classes. Indeed, of the total ~26,500 peptide bonds in the dataset of substrates with solved structures, only 158 (0.5%) belong to the positive class based on the available data. The same bias was observed for the dataset of structural models (83 of the total 15,224). Therefore, of the standard threshold-associated metrics, only three, Precision (a fraction of true positives versus all positive predictions), Recall (a fraction of correctly predicted positive cases relative to the entire positive class), and F-score (a harmonic mean of Precision and Recall with an option to assign different weights for each of the two metrics), were taken into consideration [42]. As our main objective was to establish a capability to precisely classify nearly all known cleavage sites while minimizing the total number of positive predictions, we could use F-score as the sole evaluation metric for the simultaneous optimization of both Precision and Recall. Accuracy is recognized as an inadequate metric for the imbalanced classification problem [43], and it was not used in our study.

The second feature is the intrinsic underrepresentation of the positive class. The extent of this underrepresentation is expected to be very large as the data were collected from published studies where a given protein substrate was tested with only one or several proteases (but not with the entire range of 94 proteases in our datasets). Even more importantly, whereas our analysis seeks to deduce structural features of cleavage sites independently of sequence motifs, actual regulatory proteases, even if all of them could be tested, have well-defined sequence preferences covering only a small fraction of all potentially cleavable peptide bonds. Therefore, a very large number of cryptic (or “hidden”) cleavage sites would appear in the negative class. Due to these intrinsic limitations of the dataset, the apparent optimal classification threshold would be significantly shifted toward the positive class. In an attempt to partially compensate for this bias, we introduced the additional F_{10}-score metric assigning a tenfold higher weight for Recall compared to Precision. This is equivalent to lowering a penalty for false positives compared to false negatives. In addition to these two threshold-based evaluation metrics (F-score and F_{10}-score), we used a more general AUC (Area Under the ROC Curve) metric [44], which aggregates the behavior of the classification model over all possible thresholds.

Probability of the presence of a particular type of secondary structure at proteolytic sites was assessed using a maximum likelihood method [45]. For this purpose, probability values for loops, helices, and β-sheets were treated as a vector of unknown parameters $\mathbf{p} = (p_l , p_h , p_\beta)$, where $p_l + p_h + p_\beta = 1$. Then, because secondary structure content is different in each considered protein, the probabilities of the particular type of secondary structure in the cleavage sites for each given protein were derived as:

$$p_{s,i}=\frac{p_s\frac{n_{s,i}}{N_i}}{\sum_s p_s\frac{n_{s,i}}{N_i}}$$

where i is the protein index, $s = \{l, h, \beta\}$ reflects the secondary structure type, $n_{s,i}$ is the number of residues of particular secondary structure type s in the i-th protein, and N_i is the total number of residues in the i-th protein. The estimates $\hat{p} = (\hat{p}_l, \hat{p}_h, \hat{p}_\beta)$ of probabilities were obtained by finding the values, which are maximizing likelihood function:

$$\widehat{p}=\underset{\{p_s\}\in[0,1]}{\arg\max}\prod_{j=1}^{M}p_{s(j),i(j)}$$

where j is the index of proteolytic event, M is the total number of proteolytic events, $s(j)$ is the type of the secondary structure for the j-th proteolytic event and $i(j)$ is the index of the substrate of the j-th proteolytic event. Probability values were found by an exhaustive search on a grid constructed within p_l, p_h and p_β parameters space [0,1]×[0,1]×[0,1] with 0.01 step.

Combined set of descriptors

The combined set was defined as a set of 25 descriptors, which were estimated as statistically significant (p-value < 0.05) by the t-test applied to the solved structure dataset for the case of raw values of descriptors. Naïve Bayes and Decision tree algorithms were applied using Orange 2.0 [46]. LIBSVM [47] was used for classification by the SVM method involving the gaussian and linear kernel options.

Results

The approach and the overview

The datasets used for our analysis included 315 proteolytic cleavage sites that have been described in the literature and captured in CutDB [20]. The documented cleavage sites were included for analysis if, and only if, the site resided in a protein of known structure (217 sites, Supplemental File 1) or within a protein whose structure could be confidently modeled based on the known structure of a close homolog (98 sites, Supplemental File 2). These cut sites reside within a variety of proteins, and were generated predominantly by mammalian regulatory proteases from all major catalytic classes (Supplemental File 15). The analysis aimed to assess the contribution of structural features of substrates to proteolysis and to estimate the predictive capabilities of these features included in the following key stages:

i. *Structural features of the analyzed cleavage sites* were captured by values of 14 common structural descriptors (32 descriptor variants) computed or obtained directly from PDB files (Supplemental File 5). This list encompasses all descriptors used in previous studies on proteolysis [13,14,15,16,17]. It was expanded substantially by inclusion of additional descriptors (backbone torsion angles, depth index, molecular surface accessibility, etc.) and by applying novel computational tools (for example, solvent accessibility was computed using DSSP and Naccess tools). A gallery of 3D structures and structural models with visualized cleavage sites was built (Supplemental Files 3 and 4) to assist in initial selection and further exploration of structural features of proteolytic events.

ii. Statistical significance of associations of all cleavage sites with individual descriptors was evaluated by two distinct methods [48]. The first method involved a statistical analysis to determine if cleaved *vs.* uncleaved peptide bonds indeed have a different distribution across structural features (descriptors). The second method, which is derived from the machine learning theory, sought to employ structural descriptors to predict cleavable peptide bonds by treating prediction as a binary classification task. Both methods gave essentially consistent results and allowed us to rank the entire set of structural descriptors by their significance in predicting proteolytic events.

iii. Comparison of these results between the two datasets of protein substrates with known and predicted structures revealed a high level of consistency, suggesting that homology-based models can be used for structure-based prediction of cleavage sites with comparable efficiency. In contrast, structural descriptors deduced directly from the amino acid sequence showed only marginal significance.

iv. Probability of occurrence of a particular secondary structure type in proteolytic sites was assessed in the most detail as these features are commonly used in the field despite their modest predictive capabilities compared to other types of structural descriptors revealed in this study. Most notably, our analysis based on normalization for the occurrence of the three types of secondary structure elements for each protein substrate revealed an appreciable frequency of cleavage within certain types of β-strands.

v. Mutual dependencies (redundancies) of individual descriptors and a potential for improving cleavage site predictions by combining descriptors were evaluated. Based on the results obtained by three different machine learning algorithms, a combination of selected descriptors leads to at least a 15% improvement in the false prediction rate.

The key specific results obtained using this approach are presented below.

Correlation of Cleavage Sites with Individual Structural Descriptors

The Student's method (t-test) was used to test the null-hypothesis that a distribution of a particular structural descriptor is the same for cleaved *vs.* uncleaved peptide bonds. The probability that this hypothesis is true is estimated by P-value. The results of t-test for best-performing descriptors representing all three categories of structural features—*exposure, flexibility*, and *local interactions*—are shown in Table I (results for raw and normalized values of all structural descriptors are provided in Supplemental Files 6 and 7). The four descriptors with the strongest statistical significance associated with proteolytic events (P-value range 5×10^{-47} to 3×10^{-27}) are *protrusion index*, *solvent accessibility*, *packing*, and *molecular surface accessibility*. The second tier of predictors (P-value range 7×10^{-18} to 5×10^{-4}) include *disordered regions*, *depth index*, *B-Factor*, *hydrogen bonding*, and *loops* in the secondary structure. Similar results were generally observed for both raw and normalized (Supplemental File 7a) values of structural descriptors. The B-factor is the one notable exception because its significance was 17 orders of magnitude better for normalized values (4×10^{-25} compared to 1×10^{-8}).

Calculations of statistical significance of structural descriptors were performed for amino acids in P1 subsites immediately adjacent to the cleaved peptide bond. This is a standard approach based on the assumption of the key contribution of the P1 position as known for sequence-based substrate preferences of many types of proteases [49,50,51]. To test the validity of this assumption, we calculated t-test statistics for other subsites, from P5 to P5′, using the solved structures dataset and raw values of structural descriptors. Indeed, 8 top-scoring structural descriptors (out of 14 descriptors with $P\text{-value} < 0.05$) had the highest significance values for P1 subsite. This apparent trend (illustrated in Fig. 2 for the three best descriptors) confirms that P1 subsite make the largest contribution to the observed associations between structural features and proteolytic susceptibility.

The general statistical approach allowed us to elucidate the significance of the differences between distributions of the structural descriptors values among cleaved and uncleaved peptide bonds. However, it only roughly estimates the ability of individual structural features to predict cleavage sites. To assess the predictive capability more precisely, we applied methodology from the machine learning theory to each structural descriptor. We chose three evaluation metrics, F-score, F_{10}–score, and AUC, for this task because they are the most suitable for estimating datasets in which there is an *imbalance of two classes* (intrinsic underrepresentation of cleaved peptide bonds compared to uncleaved bonds). The complete results of calculations for raw and normalized values are presented in Supplemental Files 8–12.

Importantly, the structural descriptors with the highest statistical association with cleavage sites also had the best predictive capabilities (Table I). Among them, *solvent accessibility*, *protrusion index*, *packing*, *molecular surface accessibility*, and *B-factor* (calculated in normalized values) were the top performers, whereas *disordered regions*, *hydrogen bonding*, and *loops* showed moderate prediction strength. The *depth index* was found to be a moderate-strength predictor by F-score and AUC metrics but was ranked among the top-scoring descriptors by the F_{10}-score metric.

The top-ranking descriptors listed above were capable of predicting 90% of the true cleavage sites with the cost of a false prediction rate of ~50% for other peptide bonds. The high rate of false positives likely originates from the intrinsic biases (such as underrepresentation of a positive class) in our dataset. Indeed, many of the high-scoring sites in every protein may indeed be cleavable by proteases other than those for which there is actually documentation in the literature. Therefore, a substantial fraction of peptide bonds in the negative class is expected to represent "hidden positives" (potentially cleavable bonds).

Utility of homology-based structural models

In contrast to predictors deduced from 3D structures, the predictive capabilities of sequence-derived structural descriptors were relatively weak (Supplemental File 10). The t-test applied to ~80 sequences from the dataset of protein substrates with solved 3D structures revealed a marginal significance only for *predicted solvent accessibility* and *predicted disorder* (see Table I). Similar results were obtained by the second method (based on AUC and F-score metrics).

On the other hand, a comparison of the t-test statistics between solved structure and structural model datasets revealed an appreciable significance of model-derived descriptors and a high level of consistency between the results for the two datasets. Indeed, the ranking order of top-scoring descriptors was essentially the same for both datasets (Table I). This observation as well as statistically significant P-values (which may not be directly compared between the datasets due to the difference in sample size) provided the first evidence of the utility of homology-based 3D models for prediction of proteolytic susceptibility. These conclusions were further confirmed by the results of the machine learning approach (Table I).

Proteolytic events in the context of secondary structure

The results of our study qualified secondary structure features of substrate proteins as descriptors with low to moderate predictive capabilities (Table I). At the same time, these features are commonly used for the prediction and interpretation of proteolytic events. This observation prompted us to perform a more rigorous analysis of the distribution and specific properties of cleavage sites over three types of secondary structure elements (loops, helices, and sheets). A maximum likelihood approach was used to accurately estimate probabilities of occurrence of the three types of secondary structures in the sites of limited proteolysis taking into account different representations of each secondary structure type in a given protein substrate. The results of this analysis that are in good agreement for both datasets (Fig. 4) confirmed that, indeed, loops have the highest probability to be cleaved by proteases. A lower, but also significant, probability of helices to be presented in the sites of limited proteolysis was alluded to in previous reports [12, 52]. At the same time an appreciable frequency of cleavage sites in β-sheets (yet lower than in helices) is in marked contrast with the established opinion. A visual inspection of cleavage sites in β-sheets revealed that many of them are located along the perimeter of β-sheets. Thus, of 26 analyzed sites, 11 were found at the ends of β-strands (Fig. 5a), and 11 were found inside β-strands that are located at the edge of the β-sheet (Fig. 1c).

Predictive capabilities of combinations of structural descriptors

To explore and reduce the anticipated redundancy in the set of structural descriptors, we calculated Pearson correlation coefficients for pairs of descriptors (Supplemental File 13). Not unexpectedly, the strongest dependency (0.85) was observed between solvent accessibility and molecular surface accessibility. Another notable correlation was observed between protrusion index and solvent accessibility, which is consistent with previous reports [34]. All other pairwise correlations were less pronounced.

Based on the analysis described above, we have chosen a set of 25 descriptors deemed statistically significant (P-value < 0.05) by the t-test on the raw values of descriptors (calculated for the solved structure dataset). A predictive capability of this set with respect to the binary classification problem (as formulated above) was assessed by modern machine learning methods: *Linear Kernel SVM*, *Gaussian Kernel SVM*, *Naive Bayes*, and *Decision Tree*. Testing of the built classification models was performed using a tenfold cross-validation approach. The results of this analysis demonstrated that combining structural descriptors leads to a notable improvement of predictive capabilities (Fig. 3 and Supplemental File 14). The best results were obtained using a linear SVM classifier whose scoring function is the weighted sum of features (structural descriptors). Overall, this integrative approach allowed us to reduce the false prediction rate from 50% to 35% while retaining an ability to recover 90% of true positives. As already discussed, these results, in fact, point to approximately one-third of all peptide bonds in an average protein as being potentially susceptible to proteolytic processing. Although this conclusion cannot be experimentally tested (due to constrains imposed by sequence-based substrate specificity of proteases), it is in general agreement with expectations from the overall architecture of globular proteins.

Discussion

The aim of this study was to establish structural descriptors for classifying individual peptide bonds within proteins and then to estimate the power of these descriptors in predicting susceptibility to proteolysis. We reasoned that features (structural descriptors) with high predictive power reflect the aspects of protein structure that permit proteolytic cleavage and that these regions have a higher probability for containing regulatory cleavage sites. To accomplish this objective, we performed a statistical analysis on a set of ~ 315 documented proteolytic events in 123 proteins with known or predicted 3D structures.

The structural descriptors can be segregated into three main categories (*exposure*, *flexibility*, and *local interactions*) based on the features they describe. Based on results from two complementary approaches (statistical hypothesis testing and machine learning–based classification), descriptors with the highest predictive power relating protein topology are: *solvent accessibility*, *protrusion index*, *molecular surface accessibility*, *packing*, and, to some extent, *depth index*. Descriptors that reflect the flexibility of polypeptide chains (like *B-Factor* and *disordered regions*) had the second-highest predictive power. Finally, descriptors that convey the strength of local interactions in proteins (like *hydrogen bonding* and *secondary structure loops*) have the lowest predictive power. The rank order of the predictive power of these features holds for proteins of known 3D structure and for proteins for which we had only homology-based models. Taken together, these observations suggest that the ability of a protease to physically access a peptide bond is the most critical factor in determining susceptibility to proteolysis. Other structural descriptors from our list did not exhibit significant correlation with cleavage sites (Supplementary Files 6–12).

The observed consistency of all types of estimations between two protein datasets suggest that structural models [53] can be used for cleavage site prediction almost as efficiently as the solved structures (Table I). This is particularly important in the context of the rapidly improving structural modeling techniques along with the expansion of structural coverage of protein families in PDB. On the other hand, structural descriptors deduced solely from the amino acid sequences showed relatively poor performance. The only association with cleavage sites at the significance level comparable with genuine structural descriptors of moderate predictive power (from the *local interactions* category) was revealed for *predicted solvent accessibility* (Table I). This observation indicates that, despite the obvious benefits of utilizing readily available protein sequences, our current ability to deduce useful structural descriptors is limited by relatively scarce 3D structural data.

Overall, the main conclusions of our study are in general agreement with the first systematic survey by Hubbard *et al.* [13], which implicated *exposure*, *flexibility*, and *local interactions* as important structural determinants of limited proteolysis. At the same time, though, our statistical analysis for the first time afforded the ability to rank these three categories of structural features by their relative significance (Table I). Among the few contradictions between our specific findings and conclusions of Hubbard *et al.* [13], we assigned a high predictive power to protrusion index, a likely result of the difference in the calculation methods.

Another interesting finding is that B-Factor, a feature reflecting an atom's thermal motion, showed the best correlation with proteolytic susceptibility only when normalized in the context of each protein substrate, whereas for most other good descriptors, both normalized and raw values produced comparable results. This observation suggests that the flexibility observed in the actual 3D structure (which captures only a subset of many possible conformations of the protein molecule) should be considered in relative terms within a given structure rather than between the structures, which might be at least partially due to the inconsistent reporting of B-Factor between structures [54].

A relative probability of proteolytic processing within different types of secondary structures remains a subject of conflicting reports. Thus, while many early studies indicated that proteases cleaved mostly in the loops, our analysis revealed a lower but substantial probability of cleavage in helices. These conclusions are consistent with some of the recent reports [12,52]. Cleavage in β-strands is still commonly perceived as highly unlikely if not impossible [17]. Nevertheless, a more rigorous statistical analysis, which accounts for differences in the relative content of all types of secondary structure elements in different proteins, revealed an appreciable (albeit lower) frequency of limited proteolysis in β-strands.

The relevance of cleavage in β-strands is supported by many well-documented and physiologically important proteolytic events. For example, the cleavage inside the edge β-strand of the birch profilin β-sheet (Fig. 1c) was reported for mast cells alpha-chymase in conjunction with the attenuation of allergic response [55]. Two cleavage sites at the edges of the β-strands were reported for lactoferrin (Fig.5a) as a result of autoproteolytic activity of this iron binding protein associated with mammalian non-immune defense against pathogens [56]. Cleavage of an internal strand of a β-sheet, which, at the same time, is the N-terminus of the protein, was registered for actin and two different types of proteases (Fig. 5b), caspase-1 and Granzyme B [57, 58]. The latter protease was also reported to cleave the internal β-strand proximal to the N-terminal strand in alpha-enolase (Fig. 5c) [58].

The examples listed above also illustrate the tendency of cleavage sites to occur either close to the edges of β-strands or inside β-strands that are located at the edge of a β-sheet. This trend revealed by detailed examination of our 3D structural gallery of visualized proteolytic events (Suplemental Files 3 and 4) is consistent with some of the earlier suggestions [17]. It likely reflects the tendency of β-sheet perimeter residues to be exposed and have lower hydrogen bonding energy than internal residues. Interestingly, in all 4 cases (of 26 examined) of truly internal cleavage sites in β-sheets, the respective strands were located very close to the N- or C-termini (Fig. 5b and 5c).

Among significant structural descriptors assessed by us for the first time, the most interesting behavior was observed for the *depth index*, which measures the distance from the peptide bond to the surface of the protein. When evaluated by the F_{10}-score metric, *depth index* was the most important feature for cleavage, although by other metrics it was behind *accessibility*, *protrusion*, and *packing index*. Interestingly, *depth index* appears to be of particular importance for the cleavage sites with relatively low solvent accessibility. Visual inspection of representative poorly accessible cleavage sites revealed favorable values of *depth index*. Remarkably, in most of such cases, the respective peptide bonds located at relatively low depth appeared to be "shielded" by loops with high B-Factor values. It is tempting to speculate that the access of a protease to such a bond might be granted by the mobility of a loop. This interpretation is consistent with the utmost importance of accessibility (exposure) even when it is masked by "freezing" a protein in a particular crystallizable conformation. Using a combination of descriptors (as opposed to any single descriptor) opens the possibility of resolving at least some of such difficult cases, although it would likely require employment of additional rule-based approaches.

However, the conventional machine learning classification methods used in this study proved the concept that a combination of structural descriptors leads to substantial improvement of the accuracy of predicting a cleavage site. Thus, the linear SVM approach, despite the apparent simplicity of its scoring method, allowed us to predict ~90% of cleavable bonds while increasing the number of bonds excluded from consideration by 15% compared to the best individual descriptor.

Overall, this study provides a statistical foundation for the automated and accurate prediction of regions within proteins that have a high propensity for cleavage by endopeptidases. Our analysis suggests that approximately one-third of all peptide bonds in an average protein have the potential to be proteolytically processed based on their structural properties. By combining structure-based predictions common for many proteases with sequence-based preferences of a given protease, we expect to achieve more-accurate mapping of individual cleavage sites. In a general sense, this combined strategy has shown promise when applied to caspases [59,60]. The main distinction of the analysis described here is that it provides a solid statistical foundation for the extraction of structural features of general utility, potentially applicable to numerous regulatory proteases implicated in a variety of pathways and syndromes. These findings set the stage for the development of a new generation of software tools for accurate structure-based predictive modeling of regulatory proteolysis and other post-translational modifications. Such computational tools would find numerous applications in proteomics research, for example in a rapidly developing field of degradomics or N-terminomics [61,62,63,64].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Mikhail Gelfand, Gennady Fedonin for useful discussions and to Alessandro Pintar for providing CX and DPX algorithm's source codes. This work was supported by NIH grant number RR020843, "Center on Proteolytic Pathways" (to JWS).

References

1. Lopez-Otin C, Overall CM. Protease degradomics: a new challenge for proteomics. Nat. Rev. Mol. Cell Biol. 2002; 3:509–519. [PubMed: 12094217]
2. Lopez-Otin C, Bond JS. Proteases: multifunctional enzymes in life and disease. J. Biol. Chem. 2008; 283:30433–30437. [PubMed: 18650443]
3. 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]

4. Turk B. Targeting proteases: successes, failures and future prospects. Nat Rev Drug Discov. 2006; 5:785–799. [PubMed: 16955069]
5. Davie EW, Neurath H. Identification of a peptide released during autocatalytic activation of trypsinogen. J. Biol. Chem. 1955; 212:515–529. [PubMed: 14353852]
6. Overall CM. Molecular determinants of metalloproteinase substrate specificity: matrix metalloproteinase substrate binding domains, modules, and exosites. Mol. Biotechnol. 2002; 22:51–86. [PubMed: 12353914]
7. Diamond SL. Methods for mapping protease specificity. Curr Opin Chem Biol. 2007; 11:46–51. [PubMed: 17157549]
8. 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]
9. Boyd SE, Pike RN, Rudy GB, Whisstock JC, Garcia de la Banda M. PoPS: a computational tool for modeling and predicting protease specificity. J Bioinform Comput Biol. 2005; 3:551–585. [PubMed: 16108084]
10. 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]
11. Wee LJ, Tan TW, Ranganathan S. SVM-based prediction of caspase substrate cleavage sites. BMC Bioinformatics. 2006; 7(Suppl 5):S14. [PubMed: 17254298]
12. Timmer JC, Zhu W, Pop C, Regan T, Snipas SJ, Eroshkin AM, Riedl SJ, Salvesen GS. Structural and kinetic determinants of protease substrates. Nat. Struct. Mol. Biol. 2009; 16:1101–1108. [PubMed: 19767749]
13. 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]
14. Fontana A, Fassina G, Vita C, Dalzoppo D, Zamai M, Zambonin M. Correlation between sites of limited proteolysis and segmental mobility in thermolysin. Biochemistry. 1986; 25:1847–1851. [PubMed: 3707915]
15. Novotny J, Bruccoleri RE. Correlation among sites of limited proteolysis, enzyme accessibility and segmental mobility. FEBS Lett. 1987; 211:185–189. [PubMed: 3542567]
16. 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]
17. 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]
18. Dutta S, Burkhardt K, Young J, Swaminathan GJ, Matsuura T, Henrick K, Nakamura H, Berman HM. Data deposition and annotation at the worldwide protein data bank. Mol. Biotechnol. 2009; 42:1–13. [PubMed: 19082769]
19. Zhang Y. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 2008; 18:342–348. [PubMed: 18436442]
20. Igarashi Y, Eroshkin A, Gramatikova S, Gramatikoff K, Zhang Y, Smith JW, Osterman AL, Godzik A. CutDB: a proteolytic event database. Nucleic Acids Res. 2007; 35:D546–549. [PubMed: 17142225]
21. Pang CN, Hayen A, Wilkins MR. Surface accessibility of protein post-translational modifications. J Proteome Res. 2007; 6:1833–1845. [PubMed: 17428077]
22. The NCBI handbook [Online]. National Library of Medicine (US), National Center for Biotechnology Information; Bethesda (MD): 2002. http://www.ncbi.nlm.nih.gov/books/NBK21101/
23. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. Nucleic Acids Res. 2010; 38:D227–233. [PubMed: 19892822]
24. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. 2003; 10:980. [PubMed: 14634627]

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990; 215:403–410. [PubMed: 2231712]

26. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Palsson B, Osterman A, Godzik A. Three-dimensional structural view of the central metabolic network of Thermotoga maritima. Science. 2009; 325:1544–1549. [PubMed: 19762644]

27. Schechter I, Berger A. On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. Biochem. Biophys. Res. Commun. 1968; 32:898–902. [PubMed: 5682314]

28. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera – a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25:1605–1612. [PubMed: 15264254]

29. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]

30. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins. 1995; 23:ii–v. [PubMed: 8710822]

31. Hubbard, SJ.; Thornton, JM. [Mach 16, 2011] ’NACCESS’, Computer Program. http://www.bioinf.manchester.ac.uk/naccess/

32. Sanner MF, Olson AJ, Spehner J. Reduced Surface: an Efficient Way to Compute Molecular Surfaces. Biopolymers. 1996; 38:305–320. [PubMed: 8906967]

33. Nishikawa K, Ooi T. Radial locations of amino acid residues in a globular protein: correlation with the sequence. J. Biochem. 1986; 100:1043–1047. [PubMed: 3818558]

34. Pintar A, Carugo O, Pongor S. CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics. 2002; 18:980–984. [PubMed: 12117796]

35. Pintar A, Carugo O, Pongor S. DPX: for the analysis of the protein core. Bioinformatics. 2003; 19:313–314. [PubMed: 12538266]

36. Zhang Y, Stec B, Godzik A. Between order and disorder in protein structures – analysis of “dual personality” fragments in proteins. Structure. 2007; 15:1141–1147. [PubMed: 17850753]

37. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000; 16:404–405. [PubMed: 10869041]

38. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. Bioinformatics. 2004; 20:2138–2139. [PubMed: 15044227]

39. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins. 2004; 56:753–767. [PubMed: 15281128]

40. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25:1422–1423. [PubMed: 19304878]

41. [March 16, 2011] Oracle Database 10g Express Edition. http://www.oracle.com/technology/products/database/xe/index.html

42. Batuwita R, Palade V. A New Performance Measure for Class Imbalance Learning. Application to Bioinformatics Problems. Proceedings of the Fourth International Conference on Machine Learning and Applications. 2009:545–550.

43. Weiss GM, Provost F. Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction. Journal of Artificial Intelligence Research. 2003; 19:315–354.

44. Fawcett T. An introduction to ROC analysis. Pattern Recogn. Lett. 2006; 27:861–874.

45. Ewens, WJ.; Grant, GR. Statistical Methods in Bioinformatics: An Introduction. Springer; 2005.

46. [March 16, 2011] Orange 2.0 data mining package. http://www.ailab.si/orange/

47. Chang, C.; Lin, C. [March 16, 2011] LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

48. Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. Feature Extraction, Foundations and Applications. Springer; 2006.

49. Schilling O, Overall CM. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nat. Biotechnol. 2008; 26:685–694. [PubMed: 18500335]

50. Alves MF, Puzer L, Cotrin SS, Juliano MA, Juliano L, Bromme D, Carmona AK. S3 to S3' subsite specificity of recombinant human cathepsin K and development of selective internally quenched fluorescent substrates. Biochem. J. 2003; 373:981–986. [PubMed: 12733990]

51. Debela M, Magdolen V, Schechter N, Valachova M, Lottspeich F, Craik CS, Choe Y, Bode W, Goettig P. Specificity profiling of seven human tissue kallikreins reveals individual subsite preferences. J. Biol. Chem. 2006; 281:25678–25688. [PubMed: 16740631]

52. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell. 2008; 134:866–876. [PubMed: 18722006]

53. Zhang Y. Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. 2009; 19:145–155. [PubMed: 19327982]

54. Parthasarathy S, Murthy MR. Analysis of temperature factor distribution in high-resolution protein structures. Protein Sci. 1997; 6:2561–2567. [PubMed: 9416605]

55. Mellon MB, Frank BT, Fang KC. Mast cell alpha-chymase reduces IgE recognition of birch pollen profilin by cleaving antibody-binding epitopes. J. Immunol. 2002; 168:290–297. [PubMed: 11751973]

56. Massucci MT, Giansanti F, Di Nino G, Turacchio M, Giardi MF, Botti D, Ippoliti R, De Giulio B, De Giulio B, Siciliano RA, Siciliano R, Donnarumma G, Valenti P, Bocedi A, Polticelli F, Ascenzi P, Antonini G. Proteolytic activity of bovine lactoferrin. Biometals. 2004; 17:249–255. [PubMed: 15222473]

57. Kayalar C, Ord T, Testa MP, Zhong LT, Bredesen DE. Cleavage of actin by interleukin 1 beta-converting enzyme to reverse DNase I inhibition. Proc. Natl. Acad. Sci. U.S.A. 1996; 93:2234–2238. [PubMed: 8700913]

58. Van Damme P, Maurer-Stroh S, Plasman K, Van Durme J, Colaert N, Timmerman E, De Bock PJ, Goethals M, Rousseau F, Schymkowitz J, Vandekerckhove J, Gevaert K. 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]

59. Song J, Tan H, Shen H, Mahmood K, Boyd SE, Webb GI, Akutsu T, Whisstock JC. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics. 2010; 26:752–760. [PubMed: 20130033]

60. Barkan DT, Hostetter DR, Mahrus S, Pieper U, Wells JA, Craik CS, Sali A. Prediction of protease substrates using sequence and structure features. Bioinformatics. 2010; 26:1714–1722. [PubMed: 20505003]

61. Doucet A, Kleifeld O, Kizhakkedathu JN, Overall CM. Identification of Proteolytic Products and Natural Protein N-Termini by Terminal Amine Isotopic Labeling of Substrates (TAILS). Methods Mol Biol. 2011; 753:273–287. [PubMed: 21604129]

62. van Domselaar R, de Poot SA, Bovenschen N. Proteomic profiling of proteases: tools for granzyme degradomics. Expert Rev Proteomics. 2010; 7:347–359. [PubMed: 20536307]

63. Timmer JC, Salvesen GS. N-terminomics: a high-content screen for protease substrates and their cleavage sites. Methods Mol Biol. 2011; 753:243–255. [PubMed: 21604127]

64. Prudova A, auf dem Keller U, Butler GS, Overall CM. Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. Mol Cell Proteomics. 2010; 9:894–911. [PubMed: 20305284]

65. Danielsson B, Fenton A, Jornvall J. The site in human antithrombin for functional proteolytic cleavage by human thrombin. FEBS Lett. 1981; 126:257–260. [PubMed: 7238875]

66. Moncrief JS, Obiso R, Barroso LA, Kling JJ, Wright RL, Van Tassell RL, Lyerly DM, Wilkins TD. The enterotoxin of Bacteroides fragilis is a metalloprotease. Infect. Immun. 1995; 63:175–181. [PubMed: 7806355]

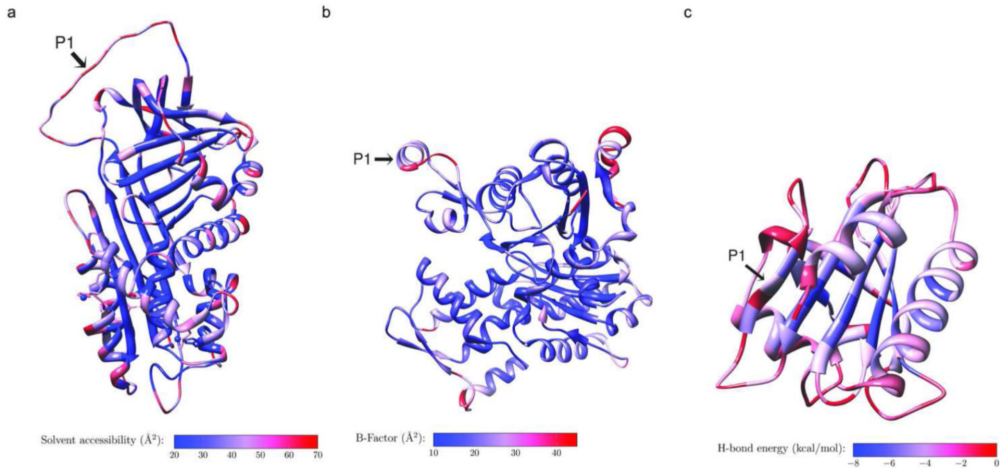

Figure 1.

Examples of the mapping of proteolytic events into 3D structure of substrate along with visualization of selected structural descriptors. Values of the solvent accessibility, B-Factor, and hydrogen bonding energy were color-mapped into substrate's structure for the cases of proteolytic processing of antithrombin by thrombin [65] (a), actin by Bacteroides fragilis enterotoxin [66] (b), and human profilin by mast cell alpha-chymase [55] (c).

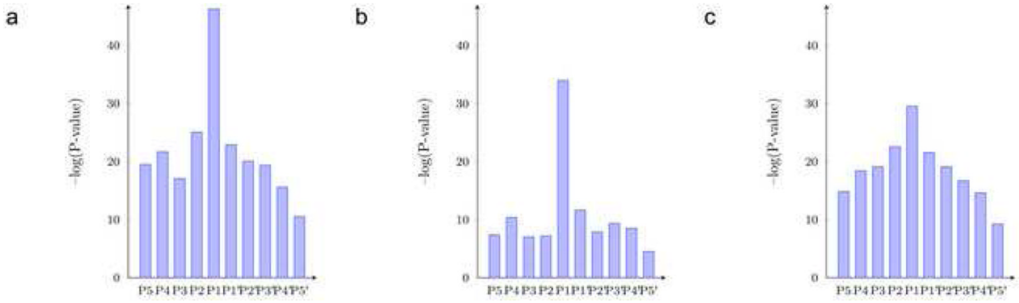

Figure 2.

Structural importance of the subsites around the cleavage site. Three bar plots demonstrate distribution of P-values over subsites for three structural descriptors: solved accessibility (a), protrusion index (b), and packing (c). The significances of P5–P5′ subsites were estimated by t-test for the raw values of structural descriptors calculated for the solved structure dataset.

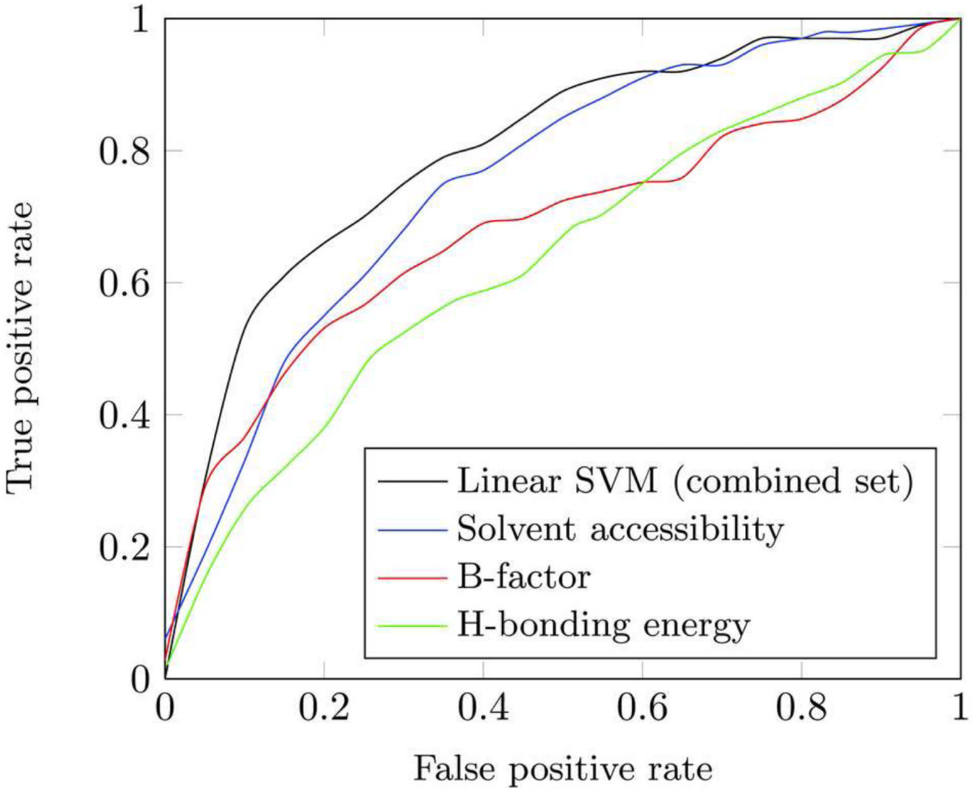

Figure 3.

Estimations of the prediction capabilities of the three individual structural descriptors and the combined descriptors set, presented by ROC curves of the corresponded classifiers.

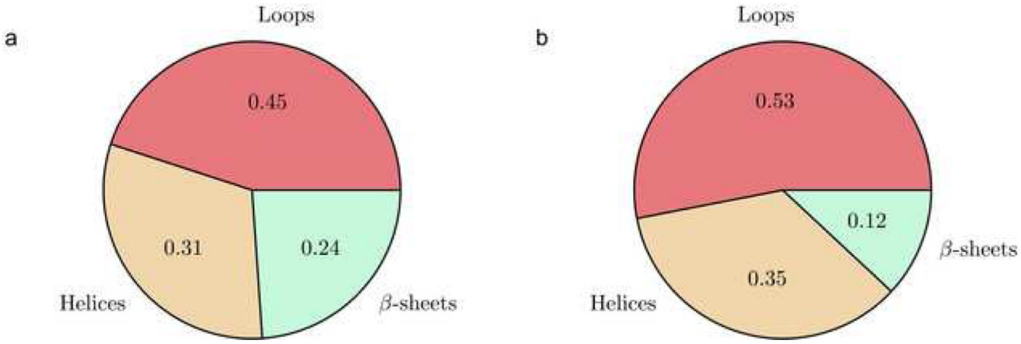

Figure 4.
Probabilities of the particular type of secondary structure in the sites of limited proteolysis calculated by maximum likelihood method for solved structure (a) and structure models (b) datasets.

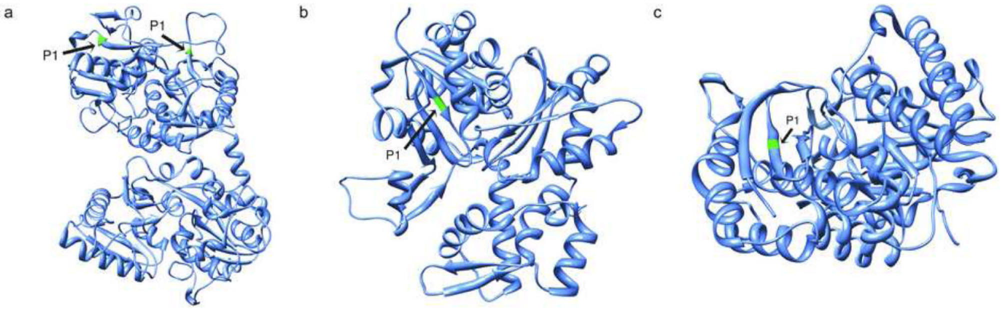

Figure 5.

Examples of cleavage sites in β structures. (a) Two proteolytic events located at the ends of the β strands were reported to be connected with the autoactivation of lactoferrin [56]. (b) The cleavage site in the buried internal β-strand of the β-sheet of actin protein, which is located close to the N-termini, was registered both for caspase 1 and for Granzyme B proteases [57, 58]. (c) Another example of the internal β-sheet cleavage, which is located inside the β-strand and next to the N-terminal β-strand of the alpha-enolase protein [58].

Table I

Significance of relevance of structural descriptors to proteolytic susceptibility estimated by statistical (t-test) and machine learning approaches (Area Under ROC-curve, F-score). Estimation methods were applied to the structural descriptors, which were calculated based on three different source of information: solved structures, computational structure models, and substrate sequences.

		Source of calculation								
Structural category	**Structural descriptor**	**Solved structures**			**Structure models**			**Sequence**		
		P-value	**AUC**	**F-score**	**P-value**	**AUC**	**F-score**	**P-value**	**AUC**	**F-score**
Exposure	Solvent accessibility	1.15e-34	0.752	7.1e-2	5.14e-12	0.688	5.7e-2	1.01e-06	0.622	1.9e-2
	Depth index	1.61e-12	0.744	3.3e-2	4.24e-06	0.690	2.4e-2	N/A	N/A	N/A
	Protrusion index	5.35e-47	0.766	6.4e-2	3.17e-13	0.712	5.1e-2	N/A	N/A	N/A
	Packing	3.03e-30	0.768	6.7e-2	3.30e-10	0.673	4.3e-2	N/A	N/A	N/A
	Molecular surface accessibility	3.19e-27	0.741	7.0e-2	1.28e-07	0.668	3.2e-2	N/A	N/A	N/A
Flexibility	B-factor	1.11e-08	0.632	2.8e-2	N/A	N/A	N/A	N/A	N/A	N/A
	Disordered regions	7.08e-18	0.536	N/A	N/A	N/A	N/A	1.90e-03	0.534	2.0e-2
Local interactions	Hydrogen bonding	3.35e-07	0.631	3.1e-2	4.25e-05	0.645	2.3e-2	N/A	N/A	N/A
	Secondary structure (loops)	4.72e-04	0.572	1.4e-2	2.81e-03	0.582	1.4e-2	0.01	0.553	1.3e-2
	Secondary structure (helices)	0.06	0.537	1.2e-2	0.34	0.524	1.2e-2	0.09	0.533	1.2e-2
	Secondary structure (β-sheets)	0.04	0.536	1.2e-2	5.59e-03	0.558	1.2e-2	0.22	0.521	1.2e-2