Cleavage Entropy as Quantitative Measure of Protease Specificity

Julian E. Fuchs, Susanne von Grafenstein, Roland G. Huber, Michael A. Margreiter, Gudrun M. Spitzer, Hannes G. Wallnoefer, Klaus R. Liedl*

Institute of General, Inorganic and Theoretical Chemistry, and Center for Molecular Biosciences Innsbruck (CMBI), University of Innsbruck, Innsbruck, Austria

Abstract

A purely information theory-guided approach to quantitatively characterize protease specificity is established. We calculate an entropy value for each protease subpocket based on sequences of cleaved substrates extracted from the MEROPS database. We compare our results with known subpocket specificity profiles for individual proteases and protease groups (e.g. serine proteases, metallo proteases) and reflect them quantitatively. Summation of subpocket-wise cleavage entropy contributions yields a measure for overall protease substrate specificity. This total cleavage entropy allows ranking of different proteases with respect to their specificity, separating unspecific digestive enzymes showing high total cleavage entropy from specific proteases involved in signaling cascades. The development of a quantitative cleavage entropy score allows an unbiased comparison of subpocket-wise and overall protease specificity. Thus, it enables assessment of relative importance of physicochemical and structural descriptors in protease recognition. We present an exemplary application of cleavage entropy in tracing substrate specificity in protease evolution. This highlights the wide range of substrate promiscuity within homologue proteases and hence the heavy impact of a limited number of mutations on individual substrate specificity.

Citation: Fuchs JE, von Grafenstein S, Huber RG, Margreiter MA, Spitzer GM, et al. (2013) Cleavage Entropy as Quantitative Measure of Protease Specificity. PLoS Comput Biol 9(4): e1003007. doi:10.1371/journal.pcbi.1003007

Editor: James M. Briggs, University of Houston, United States of America

Received October 2, 2012; **Accepted** February 7, 2013; **Published** April 18, 2013

Funding: The research of the manuscript was supported by funding of the Austrian Science Fund FWF: project "Targeting Influenza Neuraminidase" (P23051). Julian E. Fuchs is a recipient of a DOCfellowship of the Austrian Academy of Sciences at the Institute of General, Inorganic and Theoretical Chemistry at University of Innsbruck. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: klaus.liedl@uibk.ac.at

Introduction

Proteases catalyze cleavage of peptide bonds and are involved in virtually all fundamental cellular processes [1] turning proteases into central drug targets [2]. Far over 500 proteases with unique substrate cleavage patterns have been identified in the human genome [3]. These patterns reach from specificity for a single peptide to broad spectra of cleaved peptides. For instance, digestive enzymes are known to process a wide range of substrate sequences in contrast to proteases involved in signaling pathways cleaving only very distinct peptide bonds [1]. These signaling cascades include the blood-clotting cascade [4], apoptosis pathways [5] and regulatory activation steps of digestive proteases [6]. Specificity of a protease is determined by interactions in the protein-protein interface of protease and substrate. The spectrum of substrates to be cleaved is classified by subpocket-wise interactions following the convention of Schechter and Berger [7]: The peptide's scissile bond is designated between N-terminal P1 and C-terminal P1′. These subpocket indices are incremented over sequential amino acids. Protease interface residues are numbered accordingly over all subpockets Sn-Sn′, thus ensuring that interacting residues are indexed with the same number. Binding modes of processed polypeptides are highly similar due to the fact that the substrate is locked in an extended beta conformation within the protease binding site [8,9]. This canonical conformation usually includes residues in the P3-P3′ substrate region, at most extended to P5, in serine protease elastase [10].

Cleavage specificity is generally originating from distinct molecular interactions between substrate and enzyme. Simple cleavage rules for serine proteases only rely on the prominent P1-S1 interactions. For instance, the hydrophobic S1 pocket of chymotrypsin causes specificity for substrates providing hydrophobic residues at their P1 position. In contrast, an Asp residue in the S1 site of the homologous trypsin determines specificity for Arg and Lys at P1 [11]. Limitations of such simple models are evident, as S1-directed mutation does not allow transposition of trypsin specificity to chymotrypsin [12]. Moreover, complex adjacent protein-loop interactions and dynamics were found to determine substrate specificity [13,14].

Interactions between enzyme and substrate span several subpockets in the protease binding site. Experimental data shows that S2–S3 sites hardly affect substrate specificity in chymotrypsin [15], but account for specificity of the homologous elastase [16]. Especially chymotrypsin-like enteropeptidase shows exceptional specificity in the S5-S1-region cleaving only substrates containing the sequence Asp-Asp-Asp-Asp-Lys as trypsinogen [17]. P4-S4 interactions are found to be highly specific in case of the non-homologous subtilisin serine proteases [18]. Especially in the S1-S4-region, closely homologous serine proteases show significant differences in respective cleavage specificity reaching from limited proteolysis to almost unspecific substrate cleavage. Several

Author Summary

Proteases show a broad range of cleavage specificities. Promiscuous proteases as digestive enzymes unspecifically degrade peptides, whereas highly specific proteases are involved in signaling cascades. As a quantitative index of substrate specificity was lacking, we introduce cleavage entropy as a measure of substrate specificity of proteases. This quantitative score allows for straight-forward rationalization of substrate recognition by a subpocket-wise assessment of substrate readout leading to specificity profiles of individual proteases as well as an estimate of overall substrate promiscuity. We present an exemplary application of the descriptor 'cleavage entropy' to trace substrate specificity through the evolution of different protease folds. Our score highlights the diversity of substrate specificity within evolutionary related proteases and hence the complex relationship between sequence, structure and substrate recognition. By taking into account the whole distribution of known substrates rather than simple substrate counting, cleavage entropy provides the unique opportunity to dissect the molecular origins of protease substrate specificity.

cleavage site prediction tools are based on such simple and intuitive rules and are available online [19].

A plethora of experimental cleavage data for proteases is available in several databases. Cleavage information is generated experimentally by several methods reviewed by Diamond [20] and Poreba and Drag [21] reaching from fluorescence-based assays [22], isotopic labeling techniques [23], biotinylation schemes [24] over phage display [25], library-based approaches [26], microarray-based methods [27,28] and combinations thereof to modern high-throughput techniques as proteomic identification of cleavage sites (PICS) [29,30]. Cleavage data is accessible in several public databases including the MEROPS database [31,32] linking structural protease data to cleavage activity.

Although cleavage information for known proteases is easily accessible, by now no attempt has been made to develop a quantitative measure for subpocket-wise and total protease specificity in contrast to pure feature extraction techniques as for example cascade detection [33]. Analysis of protease cleavage data was mostly limited to qualitative interpretation by conversion into consensus recognition motives and visualization by sequence logos [34], iceLogo [35] or heat maps [29]. We propose the usage of information entropy to merge experimental cleavage data into an easily interpretable score for subpocket specificity as well as overall protease specificity. Following the idea of information entropy [36], which is consistent with entropy in statistical mechanics [37], we developed an information theory-based specificity score named "cleavage entropy". These cleavage entropy values depict a measure for uncertainty, and hence strictness of substrate readout, directly related to the information content of each amino acid position in a cleavage motif. A similar approach was successfully applied for description of sequence specificity of DNA binding proteins [38] and substrate promiscuity of whole enzyme families [39], including the P-region of proteases as an example [40]. DuVerle and Mamitsuka used information entropy for selection of a set of proteases showing diverse cleavage patterns and hence substrate promiscuity [41].

Methods

Extraction and Selection of Cleavage Data

To generate subpocket-wise specificity entropies, cleavage data were extracted from the MEROPS database [31]. Comparable cleavage databases as the CutDB [42] or Proteolysis MAP [43] were found to provide less cleavage information. Proteases of diverse families containing at least 100 substrate entries form a data set of 47 proteases. Methionyl aminopeptidases were excluded from the analysis, as positions P4-P2 remain unoccupied by the substrate upon cotranslational removal of N-terminal methionine residues. A complete sequence matrix containing the absolute occurrence of 20 amino acids at eight subpockets P4′ to P4 was compiled for each protease.

Calculation of Subpocket-wise Cleavage Entropy

Protease-wise cleavage sequence matrices were normalized according to the natural abundance of individual amino acids [44]. Subsequently, a second normalization to 1 at each subpocket yielded a data matrix containing probabilities for each substrate amino acid at each protease subpocket. Information theory-based cleavage entropy is defined according to Formula 1 taking into account the whole distribution of amino acids at each position rather than a single peak of elevated amino acid abundance. Substrate information is purely incorporated as sequence, not covering any kind of secondary structure information. Derived dimensionless subpocket-wise entropy values, measure the broadness of distribution of cleaved substrates, range from 0 for a perfectly conserved single amino acid to 1 for an equal distribution of substrates, reflecting complete unspecific substrate binding.

$$S_i = -\sum_{a=1}^{20} p_{a,i} * {}^{20}\log\ p_{a,i}$$

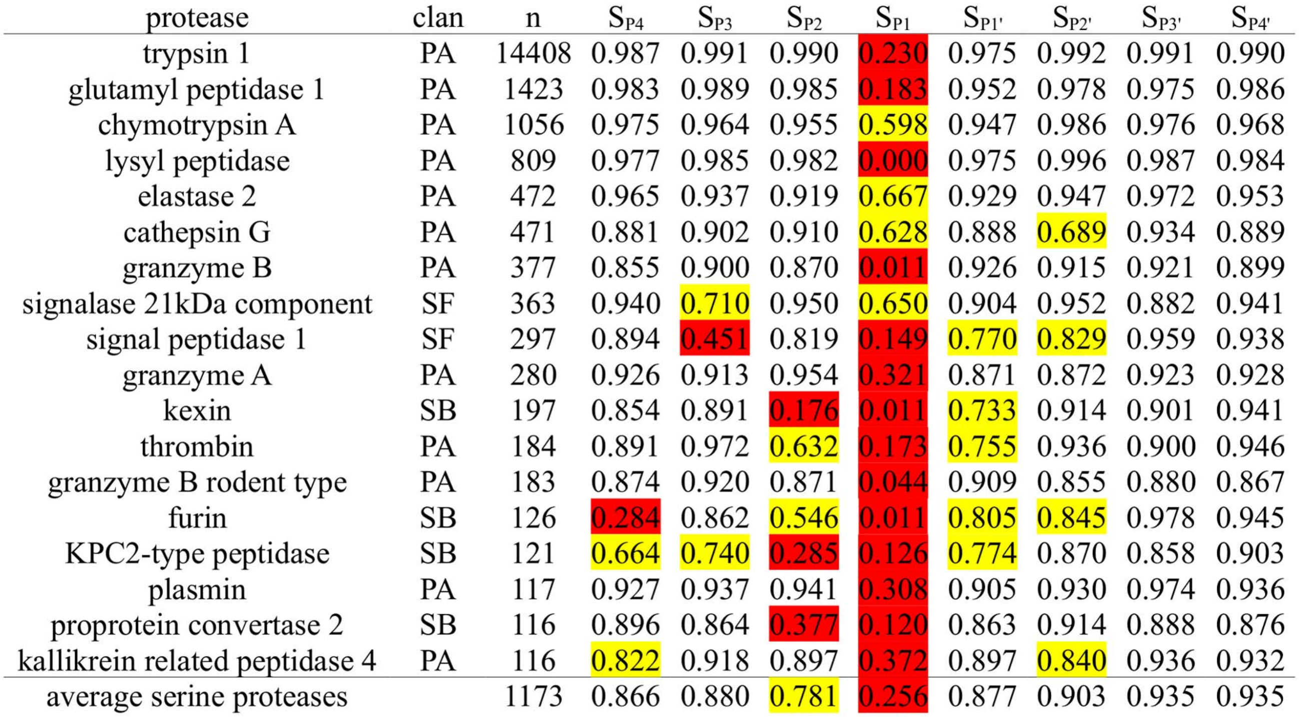
Formula 1: Calculation of subpocket-wise cleavage entropy S_i from subpocket-wise amino acid probabilities in known substrates $p_{a,i}$.

Calculation and Ranking According to Overall Cleavage Entropy

Subpocket-wise substrate specificity information is of high interest to compare individual subpockets of a single protease and individual specifity profiles between proteases. To facilitate analysis of different proteases as a whole, a summation of individual subpocket cleavage entropies yields quantitative overall cleavage entropy per protease (see Formula 2). This total cleavage entropy over eight substrate positions in the central binding site region (P4 to P4′) allows for ranking of proteases with respect to their whole substrate specificities. Entropy values range from 0 for a single conserved substrate to 8 for a random distribution of amino acids in cleaved substrates.

$$S_{Cleavage} = \sum_{i=1}^{8} S_i$$

Formula 2: Calculation of overall protease cleavage S_{Cleavage} entropy by summation of 8 subpocket-wise cleavage entropies S_i from P4-P4′ subpockets.



protease	clan	n	S_{P4}	S_{P3}	S_{P2}	S_{P1}	$S_{P1'}$	$S_{P2'}$	$S_{P3'}$	$S_{P4'}$
trypsin 1	PA	14408	0.987	0.991	0.990	0.230	0.975	0.992	0.991	0.990
glutamyl peptidase 1	PA	1423	0.983	0.989	0.985	0.183	0.952	0.978	0.975	0.986
chymotrypsin A	PA	1056	0.975	0.964	0.955	0.598	0.947	0.986	0.976	0.968
lysyl peptidase	PA	809	0.977	0.985	0.982	0.000	0.975	0.996	0.987	0.984
elastase 2	PA	472	0.965	0.937	0.919	0.667	0.929	0.947	0.972	0.953
cathepsin G	PA	471	0.881	0.902	0.910	0.628	0.888	0.689	0.934	0.889
granzyme B	PA	377	0.855	0.900	0.870	0.011	0.926	0.915	0.921	0.899
signalase 21kDa component	SF	363	0.940	0.710	0.950	0.650	0.904	0.952	0.882	0.941
signal peptidase 1	SF	297	0.894	0.451	0.819	0.149	0.770	0.829	0.959	0.938
granzyme A	PA	280	0.926	0.913	0.954	0.321	0.871	0.872	0.923	0.928
kexin	SB	197	0.854	0.891	0.176	0.011	0.733	0.914	0.901	0.941
thrombin	PA	184	0.891	0.972	0.632	0.173	0.755	0.936	0.900	0.946
granzyme B rodent type	PA	183	0.874	0.920	0.871	0.044	0.909	0.855	0.880	0.867
furin	SB	126	0.284	0.862	0.546	0.011	0.805	0.845	0.978	0.945
KPC2-type peptidase	SB	121	0.664	0.740	0.285	0.126	0.774	0.870	0.858	0.903
plasmin	PA	117	0.927	0.937	0.941	0.308	0.905	0.930	0.974	0.936
proprotein convertase 2	SB	116	0.896	0.864	0.377	0.120	0.863	0.914	0.888	0.876
kallikrein related peptidase 4	PA	116	0.822	0.918	0.897	0.372	0.897	0.840	0.936	0.932
average serine proteases		1173	0.866	0.880	0.781	0.256	0.877	0.903	0.935	0.935

Figure 1. Subpocket-wise Cleavage Entropies of Serine Proteases. Serine proteases and associated MEROPS clans sorted according to the number of known substrates n with their respective subsite-wise cleavage entropies S_i. Specific subpockets showing a cleavage entropy equal or less than an arbitrary cutoff of 0.85 are highlighted in yellow, values lower than 0.5 indicating stringent specificity in red.

doi:10.1371/journal.pcbi.1003007.g001

Cooperativity Effects in Substrate Readout: Pairwise Cleavage Entropy

Although cooperativity effects between subpockets were described for subtilisins [45] and reviewed by Ng et al. [46], available cleavage data only allows for a rough estimation of these correlation effects besides independent study of subpocket specificity. To cover inter-subpocket correlation effects in detail, data simply based on known substrates is too sparse. An extension from purely qualitative cleavage information to substrate-dependent quantitative binding affinity or kinetics measurements would be necessary. A suitable database containing diverse protease substrates is currently not known to the authors, but could also be of high interest to weight individual substrate contributions in order to refine the current implementation. A smaller set of fluorescence-based substrate turnover measurements for proteases was published by Harris et al. [22], but is restricted to variation of the P-region in substrates for eight proteases.

As only trypsin provides a sufficient data basis to study subpocket correlation effects with more than 14000 substrates listed in MEROPS, we performed an inter-subpocket correlation analysis only for this protease. The one-dimensional subpocket-wise cleavage entropy calculations presented above can directly be extended to a more-dimensional case yielding for two dimensions a pairwise cleavage entropy score depending on amino acids a and b at position i and j and their respective probabilities $p_{a,i}$, $p_{b,j}$.

$$S_{i,j} = -\sum_{a,b=1}^{400} \left(p_{a,i} * p_{b,j}\right) * {}^{400}\log\left(p_{a,i} * p_{b,j}\right)$$

Formula 3: Calculation of pairwise cleavage entropy $S_{i,j}$ from subpocket-wise amino acid pair probabilities in known substrates $p_{a,i}$, $p_{b,j}$.

This measure for inter-subpocket correlation effects yields as in the independent analysis (cleavage entropy) a score of 0 for a conserved single amino acid pair and a value of 1 for a distribution of amino acid pairs as expected by random chance from natural abundance [44]. To avoid artifacts from a lacking data basis we set a stringent cutoff of 10000 substrates in this two-dimensional analysis to allow for the same statistics as in the one-dimensional case (100 substrates).

Phylogenetic Analysis of Protease Clans

As part of the discussion, protease specificity is compared to evolutionary distance. Sequences downloaded from Uniprot [47] as indexed in the MEROPS database [31] were grouped into respective protease clans. Sequences of each clan were sorted according to total cleavage entropy and aligned by ClustalW using default settings [48]. Tools from the EMBOSS server [49] were used for phylogenetic tree construction: fprotdist using default settings to calculate protein distance matrices, fkitsch using default settings to construct phylogenetic trees using the Fitch-Margoliash method [50]. Phylogenetic trees were visualized using Interactive Tree of Life (ITOL) [51].

Visualization of Specificity Landscapes

Protein structure visualizations were created with PyMOL [52] based on the X-ray structures of trypsin and thrombin in complex with BIBR1109 (PDB: 1G32, 1G36) [53]. A subpocket definition derived from Bode et al. [54] was used for mapping of subpocket-wise cleavage entropies to the binding site region.

Results

Quantification of Subpocket-wise Cleavage Specificities

Entries with more than 100 annotated substrates in the MEROPS database represent 47 proteases comprise all major

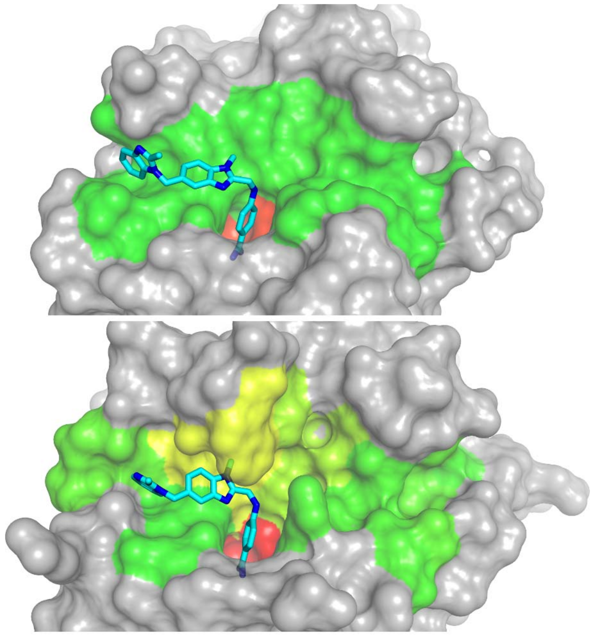


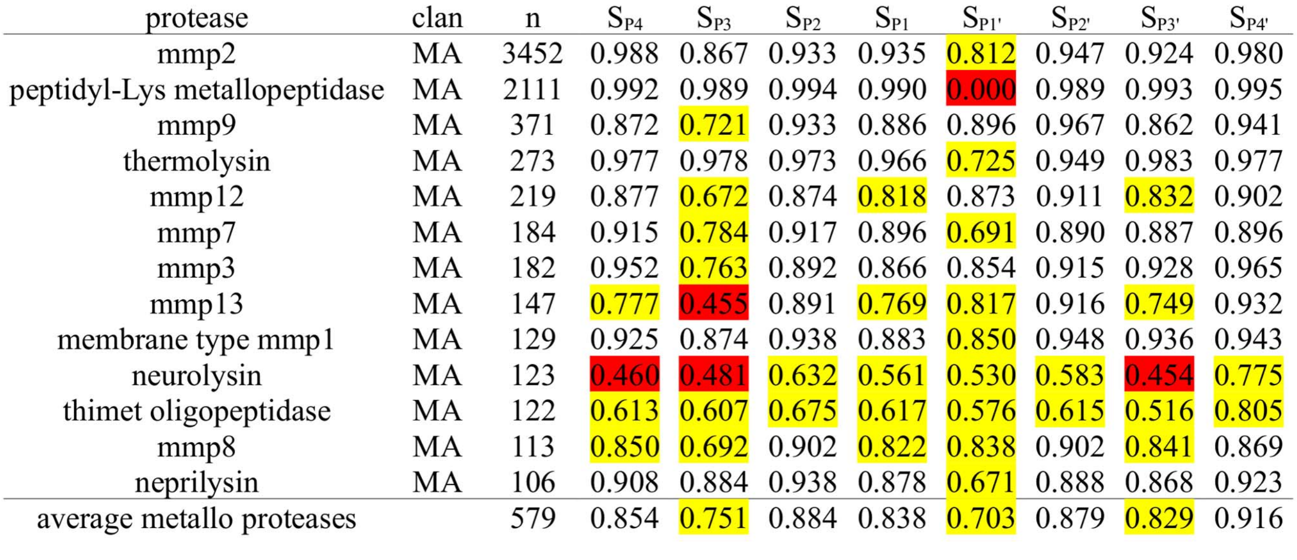
Figure 2. Specificity Landscapes of Trypsin and Thrombin. Subpocket-wise cleavage entropies mapped to the binding site region of trypsin (top) in a color spectrum of red (low, specific) over yellow to green (high, unspecific) highlight the central S1 pocket as only determinant of substrate specificity within the binding region S4-S4′ (left to right). By contrast, thrombin (bottom) binding the same small molecule inhibitor BIBR1109 [53] extends substrate recognition over further subpockets: the yellowish S1 and S1′ pockets in the specificity landscape of the binding site contribute to a more specific substrate readout.
doi:10.1371/journal.pcbi.1003007.g002

protease catalytic types. The three major protease catalytic types, serine, metallo and cysteine proteinases, covering more than 90% of known proteases [9], represent 40 entries or 85% of the test set. Calculated subpocket-wise cleavage entropies will be discussed by catalytic type to enable comparison of relative variation of binding specificity. Relative importance of subsites in determining cleavage specificity is highlighted by lowered entropy values providing specificity profiles for individual proteases.

Serine proteases show pronounced specificity at the P1 substrate site occupying the characteristic deep S1 pocket with an averaged cleavage entropy as low as $S_{P1} = 0.256$ (see Figure 1). The low P1 cleavage entropy value reflects widely accepted specificity rules for serine proteases solely based on P1-S1 interactions. A second hotspot for specific interactions of serine proteases is found in the P2-region with an average cleavage entropy of $S_{P2} = 0.781$, which is especially lowered for proprotein processing proteases kexin, furin and proprotein convertase 2 cleaving at paired basic residues [55]. Overall, serine proteases tend to bind conserved residues in P-region (average $S_{P4\text{-}P1} = 0.696$) rather than the P′-region (average $S_{P1'\text{-}P4'} = 0.912$) in accordance to findings of Page et al. for coagulation proteases as thrombin [56]. See Figure 2 for a detailed comparison of subpocket-wise cleavage entropies mapped to the three-dimensional structure of thrombin and trypsin.

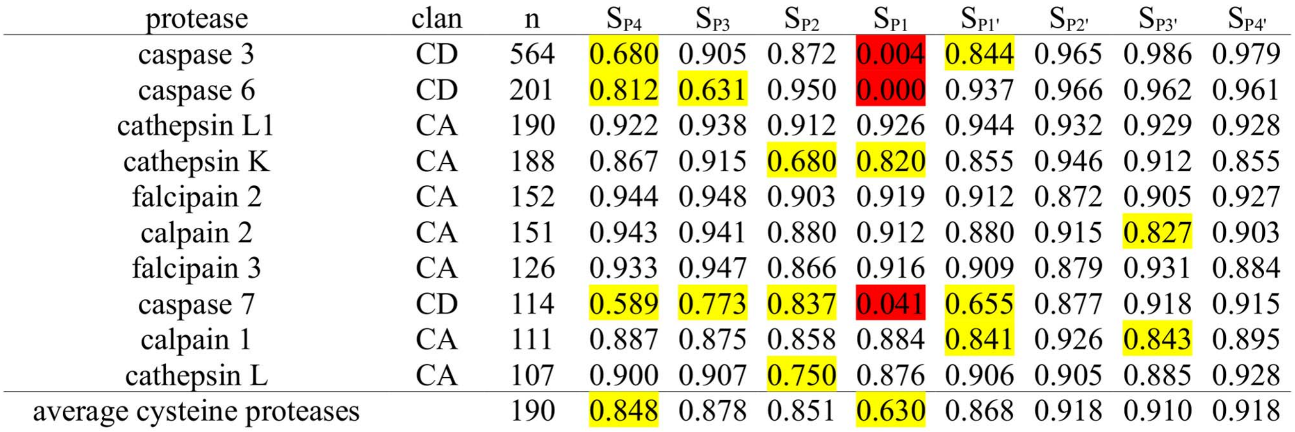
All serine proteases in the test set show pronounced specificity in the P1-region, including even so-called unspecific proteases as trypsin binding to highly conserved arginine and lysine residues at the P1 site. An extension of this specific reading frame in both directions of the substrate is observed for example for thrombin and furin, where the latter protease shows extraordinary specificity at the P4 site independent of other specific residues. These lowered entropy values reflect the proposed Arg-Xaa-Lys/Arg-Arg consensus in the P4-P1-region for furin substrates [57] and confirm general specificity rules for P4 specificity of the subtilisin-like clan of serine proteases [18].

Metallo proteases in general show less intense subpocket-wise specificity patterns than serine proteases. Their substrate readout is most pronounced in the P1′ position with an average cleavage entropy of 0.703 (see Figure 3) consistent with findings of Overall et al. for the substrate specificity of matrix metallo proteases [58]. Peptidyl-Lys metallo peptidase reads a perfectly conserved lysine residue at P1′ in all 2111 known substrates. However, P1′ is not the most specific subpocket in all metallo proteases. Further



protease	clan	n	S_{P4}	S_{P3}	S_{P2}	S_{P1}	$S_{P1'}$	$S_{P2'}$	$S_{P3'}$	$S_{P4'}$
mmp2	MA	3452	0.988	0.867	0.933	0.935	0.812	0.947	0.924	0.980
peptidyl-Lys metallopeptidase	MA	2111	0.992	0.989	0.994	0.990	0.000	0.989	0.993	0.995
mmp9	MA	371	0.872	0.721	0.933	0.886	0.896	0.967	0.862	0.941
thermolysin	MA	273	0.977	0.978	0.973	0.966	0.725	0.949	0.983	0.977
mmp12	MA	219	0.877	0.672	0.874	0.818	0.873	0.911	0.832	0.902
mmp7	MA	184	0.915	0.784	0.917	0.896	0.691	0.890	0.887	0.896
mmp3	MA	182	0.952	0.763	0.892	0.866	0.854	0.915	0.928	0.965
mmp13	MA	147	0.777	0.455	0.891	0.769	0.817	0.916	0.749	0.932
membrane type mmp1	MA	129	0.925	0.874	0.938	0.883	0.850	0.948	0.936	0.943
neurolysin	MA	123	0.460	0.481	0.632	0.561	0.530	0.583	0.454	0.775
thimet oligopeptidase	MA	122	0.613	0.607	0.675	0.617	0.576	0.615	0.516	0.805
mmp8	MA	113	0.850	0.692	0.902	0.822	0.838	0.902	0.841	0.869
neprilysin	MA	106	0.908	0.884	0.938	0.878	0.671	0.888	0.868	0.923
average metallo proteases		579	0.854	0.751	0.884	0.838	0.703	0.879	0.829	0.916

Figure 3. Subpocket-wise Cleavage Entropies of Metallo Proteases. Subpocket-wise cleavage entropies S_i of metallo proteases and associated MEROPS clans sorted by decreasing number of known substrates n. Specific pockets are highlighted in yellow and red according to their respective substrate promiscuity (yellow: $0.5 < S_i < 0.85$, red: $S_i < 0.5$).
doi:10.1371/journal.pcbi.1003007.g003



protease	clan	n	S_{P4}	S_{P3}	S_{P2}	S_{P1}	$S_{P1'}$	$S_{P2'}$	$S_{P3'}$	$S_{P4'}$
caspase 3	CD	564	0.680	0.905	0.872	0.004	0.844	0.965	0.986	0.979
caspase 6	CD	201	0.812	0.631	0.950	0.000	0.937	0.966	0.962	0.961
cathepsin L1	CA	190	0.922	0.938	0.912	0.926	0.944	0.932	0.929	0.928
cathepsin K	CA	188	0.867	0.915	0.680	0.820	0.855	0.946	0.912	0.855
falcipain 2	CA	152	0.944	0.948	0.903	0.919	0.912	0.872	0.905	0.927
calpain 2	CA	151	0.943	0.941	0.880	0.912	0.880	0.915	0.827	0.903
falcipain 3	CA	126	0.933	0.947	0.866	0.916	0.909	0.879	0.931	0.884
caspase 7	CD	114	0.589	0.773	0.837	0.041	0.655	0.877	0.918	0.915
calpain 1	CA	111	0.887	0.875	0.858	0.884	0.841	0.926	0.843	0.895
cathepsin L	CA	107	0.900	0.907	0.750	0.876	0.906	0.905	0.885	0.928
average cysteine proteases		190	0.848	0.878	0.851	0.630	0.868	0.918	0.910	0.918

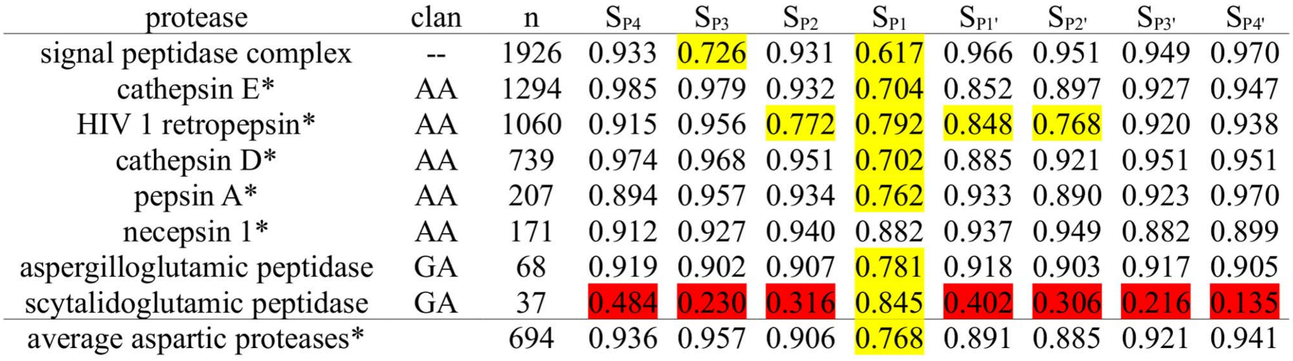
Figure 4. Subpocket-wise Cleavage Entropies of Cysteine Proteases. Cysteine proteases and associated MEROPS clans sorted according to the number of known substrates in MEROPS n. Subpocket-wise cleavage entropies S_i are color-coded to highlight specific pockets in yellow ($0.5<S_i<0.85$). Highly specific subpockets are shown in red ($S_i<0.5$).
doi:10.1371/journal.pcbi.1003007.g004

subpockets showing less pronounced substrate readout are located at P3 ($S_{P3}=0.751$) and P3′ ($S_{P3'}=0.829$) in analogy to computational predictions of Pirard [59]. Little substrate specificity is observed for other binding sites leading to an almost equivalent average substrate specificity over the whole P-and P′- region ($S_{P4\text{-}P1}=0.832$, $S_{P1'\text{-}P4'}=0.831$).

We find matrix metallo proteases (MMPs) to differ in their substrate specificity from other members of the metallo proteases. Cleavage entropy calculation highlights the P1′ position as major determinant of specificity in MMP-2, hence named "specificity pocket" [60], whereas other subsites show little substrate preferences. Additionally, a preference for proline at P3 has been observed [61,62], which is consistent with lowered cleavage entropy values at P3 found throughout the MMP family. MMP-13 shows particular preference for proline residues at P3 reducing cleavage entropy to 0.455. Strikingly, particular metallo proteases span substrate specificity over all covered subsites: The highly specific members thimet oligopeptidase and neurolysin show cleavage entropy values lower than 0.850 throughout all subpockets.

Cysteine proteases are characterized by cleavage entropies comparable to serine proteasaes rather than metallo proteases. P1 interactions dominate substrate specificity with a cleavage entropy of $S_{P1}=0.630$ similar to serine proteases (see Figure 4). Caspases account for the pronounced P1 interaction in this protease family as well as a smaller second specificity peak at P4 position ($S_{P4}=0.848$). The P-region exhibits most of cysteine protease' substrate specificites with average cleavage entropy $S_{P4\text{-}P1}=0.802$ compared to the P′-region $S_{P1'\text{-}P4'}=0.904$.

Caspases are shown to read conserved aspartate residues in P1 position with an extraordinarily high specificity ($P_1<0.05$), a characteristic not present in all other cysteine proteases. Subsite specificity of apoptosis signaling caspases [63] extends over larger areas of the P-region [64], especially pronounced in case of caspase 7 [29]. In contrast to caspases, calpains cleave broader substrate spectra whilst showing overlap with caspases in some regions of substrate space [65]. Traceable P3′ specificity is only observed for calpains amongst cysteine proteases. Broader distributions of substrates known for cathepsins [66] are quantitatively reflected by higher cleavage entropies. Cathepsin K's subtle substrate specificity at P1 and P1′ ($S_{P1}=0.680$, $S_{P1'}=0.820$) has been described by Schilling et al. [29]. Falcipains do not feature any particular subsite specificities, but tend to show complex and promiscuous specificity profiles. Simple counting of cleavage entries would have missed this unspecific behavior, as the



protease	clan	n	S_{P4}	S_{P3}	S_{P2}	S_{P1}	$S_{P1'}$	$S_{P2'}$	$S_{P3'}$	$S_{P4'}$
signal peptidase complex	--	1926	0.933	0.726	0.931	0.617	0.966	0.951	0.949	0.970
cathepsin E*	AA	1294	0.985	0.979	0.932	0.704	0.852	0.897	0.927	0.947
HIV 1 retropepsin*	AA	1060	0.915	0.956	0.772	0.792	0.848	0.768	0.920	0.938
cathepsin D*	AA	739	0.974	0.968	0.951	0.702	0.885	0.921	0.951	0.951
pepsin A*	AA	207	0.894	0.957	0.934	0.762	0.933	0.890	0.923	0.970
necepsin 1*	AA	171	0.912	0.927	0.940	0.882	0.937	0.949	0.882	0.899
aspergilloglutamic peptidase	GA	68	0.919	0.902	0.907	0.781	0.918	0.903	0.917	0.905
scytalidoglutamic peptidase	GA	37	0.484	0.230	0.316	0.845	0.402	0.306	0.216	0.135
average aspartic proteases*		694	0.936	0.957	0.906	0.768	0.891	0.885	0.921	0.941

Figure 5. Subpocket-wise Cleavage Entropies of Further Proteases. Further proteases in the test set and associated MEROPS clans not belonging to the catalytic types cysteine, serine or metallo proteases sorted according to decreasing number of known substrates n. Specific subpockets (subpocket cleavage entropy $0.5<S_i<0.85$) are shown in yellow, highly specific pockets ($S_i<0.85$) in red. Five aspartic proteases are marked with '*'.
doi:10.1371/journal.pcbi.1003007.g005

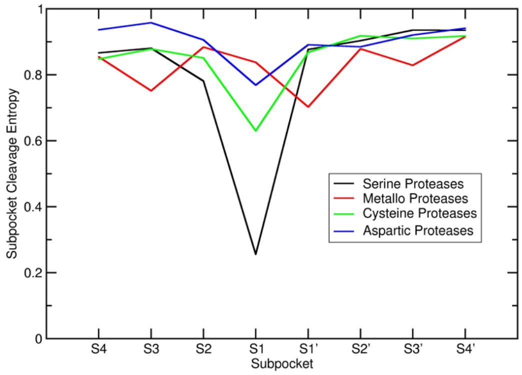


Figure 6. Subpocket-wise Cleavage Entropy Profiles for Protease Families. Subpocket-wise cleavage entropy profiles for protease catalytic classes reveal distinct substrate readout patterns for each of the protease groups. Serine proteases show most prominent subpocket specificity at the S1 site, whereas metallo proteases show specific binding behavior over a larger part of the binding pocket S4-S4′.
doi:10.1371/journal.pcbi.1003007.g006

number of available cleavage sites annotated in MEROPS is comparably low for falcipains.

Besides the three main classes of proteases, six further proteases with more than 100 cleavage patterns were found within MEROPS (see Figure 5): signal peptidase, containing a rare serine dyad at the active site [67], forming an active dimer complex in eukaryotes and hence indexed in MEROPS as complex peptidase, as well as five aspartic proteases. Two members of glutamic proteases showing distinct cleavage behavior were added to the sample to include this missing catalytic type, although less known cleaved peptides are indexed.

The signal peptidase complex is a membrane-bound protease involved in membrane translocation signaling [68]. Cleavage entropies $S_{P3} = 0.726$ and $S_{P1} = 0.617$ reflect the well-established specificity rules for signal peptidases focussing on positions P3 and P1 [69]. Distinct P1 specificity matches classical serine proteases involving a catalytic triad at the active site, whereas P3 readout is not a general characteristic of serine proteases.

All five aspartic proteases are found to depend mostly on P1 interactions with an average $S_{P1} = 0.768$. Other subpockets in P- and P′-region tend to exhibit likewise unspecific substrate binding ($S_{P4\text{-}P1} = 0.892$, $S_{P1'\text{-}P4'} = 0.909$). HIV retropepsin, a prominent target in drug design, shows distinct specificity at P2′ position with $S_{P2'} = 0.768$ supporting findings of Schilling et al. [29]. Furthermore, specific substrate readout of HIV retropepsin at positions P1 and P1′ was described in the literature [70] and is quantified with lowered cleavage entropies of $S_{P1} = 0.792$ and $S_{P1'} = 0.848$ respectively.

Aspergilloglutamic and scytalidoglutamic peptidase are added to the data set though sparse cleavage data to cover the group of glutamic peptidases represented by the members with highest number of annotated subtrates (68 and 37 respectively). Aspergilloglutamic and scytalidoglutamic peptidase provide two examples of variable cleavage profiles amongst the same protease class: Whereas the P1 position shows nearly identically lowered cleavage entropies, scytalidoglutamic peptidase reads substrate residues over the whole range of eight covered subpockets in contrast to aspergilloglutamic peptidase not showing pronounced substrate preferences at other subpockets than P1.

Summing up previous findings, average subpocket cleavage entropy profiles were calculated for protease catalytic types (see Figure 6). Serine proteases show distinct lowered cleavage entropy at their specific S1 site. Less pronounced S1 specificity is present for cysteine and aspartic proteases, whereas metallo proteases show subpocket cleavage entropy profiles including diverse cleavage entropy minima with the most specific substrate binding in the S1′ site.

Ranking of Proteases According to Overall Cleavage Specificity

Summation of subpocket-wise cleavage entropies yields a total estimate of protease specificity (see Figure 7). The additional information content of calculated total cleavage entropies compared to simple substrate counting is reflected by a squared linear correlation coefficient as low as $r^2 = 0.034$ over the core test set of 47 proteases. Likewise, qualitative ranking correlation is comparably low with a Spearman ranking correlation of $r = 0.334$ over 47 proteases. Taking into account the whole distribution of amino acids in known substrates rather than the plain number of known substrates, has the advantage to minimize the impact of large scale profiling of closely related substrates biasing the underlying data set towards non-specificity. A second bias of the selected set of investigated proteases is thereby inevitable: the selection of peptidases with more than 100 annotated cleavage sites in MEROPS favors well-studied as well as unspecific proteases. Hence, a putative perfectly specific protease cleaving only a single substrate and hence, cleavage entropy of zero, would not be covered in the presented test set.

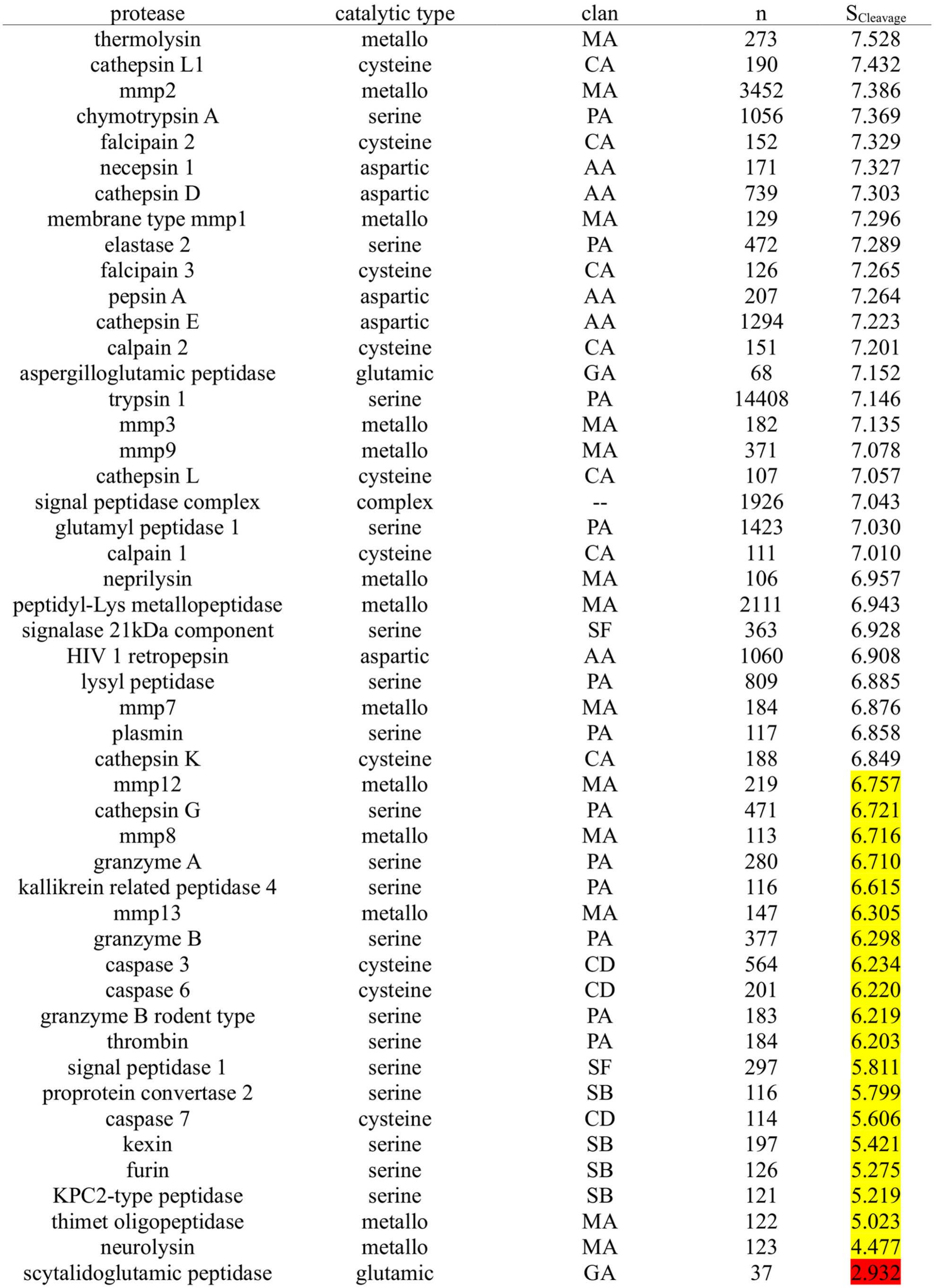
Proteases span a wide range of substrate specificites directly related to their biological roles. Ranking of the protease test set in respect to overall cleavage entropy $S_{Cleavage}$ thus yields a clear separation between unspecific digestive proteases and specific proteases involved in signaling pathways. The protease with highest observed cleavage entropy $S_{Cleavage} = 7.528$, thermolysin, is involved in bacterial nutrition by unspecificly degrading exogenous peptides [71]. The technical usage in protein sequencing [72] and peptide synthesis [73] is facilitated by this unspecific substrate recognition of thermolysin. On the other end of the test set's specificity spectrum, neurolysin is a primary example for a specific signaling protease with $S_{Cleavage} = 4.477$. The limited proteolysis of intracellular oligopeptides by neurolysin [74] assures proper regulation of cell signaling [75].

Cooperativity Effects Between Trypsin Subpockets

An exemplary analysis of inter-subpocket correlation was carried out based on over 14000 trypsin substrates listed in MEROPS (see Table S1). Only pairs including the specific P1 position show pronounced imbalances in two-dimensional distributions of substrate amino acid pairs reflected in lowered pairwise cleavage entropy scores. All other subpocket pairs show pairwise cleavage entropies in the range of 0.896 to 0.923 implying low correlation between subpocket readout. If at all a cooperative effect can be detected between P1′ and P2 in the underlying dataset for trypsin ($S_{P1',P2} = 0.896$).

Discussion

We proved cleavage entropy calculation as an intuitive approach to assess protease specificity quantitatively. In a first application of the presented score metric, we dissect the protease test set into groups of common cleavage machinery groups to elucidate potential descriptors of protease substrate specificity. This split yields four separate groups indicating distinct catalytic function: serine, metallo, cysteine and aspartic proteases (see Figure 8).



protease	catalytic type	clan	n	$S_{Cleavage}$
thermolysin	metallo	MA	273	7.528
cathepsin L1	cysteine	CA	190	7.432
mmp2	metallo	MA	3452	7.386
chymotrypsin A	serine	PA	1056	7.369
falcipain 2	cysteine	CA	152	7.329
necepsin 1	aspartic	AA	171	7.327
cathepsin D	aspartic	AA	739	7.303
membrane type mmp1	metallo	MA	129	7.296
elastase 2	serine	PA	472	7.289
falcipain 3	cysteine	CA	126	7.265
pepsin A	aspartic	AA	207	7.264
cathepsin E	aspartic	AA	1294	7.223
calpain 2	cysteine	CA	151	7.201
aspergilloglutamic peptidase	glutamic	GA	68	7.152
trypsin 1	serine	PA	14408	7.146
mmp3	metallo	MA	182	7.135
mmp9	metallo	MA	371	7.078
cathepsin L	cysteine	CA	107	7.057
signal peptidase complex	complex	--	1926	7.043
glutamyl peptidase 1	serine	PA	1423	7.030
calpain 1	cysteine	CA	111	7.010
neprilysin	metallo	MA	106	6.957
peptidyl-Lys metallopeptidase	metallo	MA	2111	6.943
signalase 21kDa component	serine	SF	363	6.928
HIV 1 retropepsin	aspartic	AA	1060	6.908
lysyl peptidase	serine	PA	809	6.885
mmp7	metallo	MA	184	6.876
plasmin	serine	PA	117	6.858
cathepsin K	cysteine	CA	188	6.849
mmp12	metallo	MA	219	6.757
cathepsin G	serine	PA	471	6.721
mmp8	metallo	MA	113	6.716
granzyme A	serine	PA	280	6.710
kallikrein related peptidase 4	serine	PA	116	6.615
mmp13	metallo	MA	147	6.305
granzyme B	serine	PA	377	6.298
caspase 3	cysteine	CD	564	6.234
caspase 6	cysteine	CD	201	6.220
granzyme B rodent type	serine	PA	183	6.219
thrombin	serine	PA	184	6.203
signal peptidase 1	serine	SF	297	5.811
proprotein convertase 2	serine	SB	116	5.799
caspase 7	cysteine	CD	114	5.606
kexin	serine	SB	197	5.421
furin	serine	SB	126	5.275
KPC2-type peptidase	serine	SB	121	5.219
thimet oligopeptidase	metallo	MA	122	5.023
neurolysin	metallo	MA	123	4.477
scytalidoglutamic peptidase	glutamic	GA	37	2.932

Figure 7. Total Cleavage Entropies of Investigated Proteases. Ranking of 49 proteases with respective MEROPS clan (including the added 2 glutamic proteases) in respect to their total cleavage entropy S_{Cleavage}. Specific proteases (S_{Cleavage}<6.8, corresponding to an average subpocket cleavage entropy S_i of 0.85 over eight investigated subpockets)are highlighted in yellow. No protease in the core test set of 47 proteases is found to be highly specific (S_{Cleavage}<4.0, reflecting an average S_i of lower than 0.5 over the whole binding site region of S_4-$S_{4'}$). Scytalidoglutamic peptidase present in the extended test set exhibits such strict substrate cleavage with a total cleavage entropy S_{Cleavage} of 2.932 owing to substrate recognition spreading over 7 highly specific subpockets (compare Figure 5).
doi:10.1371/journal.pcbi.1003007.g007

Diversity of Substrate Readout Within Proteases Sharing a Catalytic Mechanism

Strikingly, both extrema on the presented quantitative protease specificity scale for the core set of 47 proteases represent members of the metallo proteases (thermolysin and neurolysin respectively). This indicates that the catalytic cleavage machinery cannot be the major determinant of substrate specificity. Similarly, serine proteases including the prominent digestive enzymes trypsin, chymotrypsin, elastase as well as signaling peptidases kexin and furin show diverse substrate specificity. Solely the smaller sample of five aspartic proteases shows predominantly unspecific cleavage behavior with an average total cleavage entropy of $S_{\text{Cleavage}} = 7.205$ compared to an average of $S_{\text{Cleavage}} = 6.608$ for the other catalytic types. Other protease classes do not show significant differences in their substrate specificity (serine proteases: average $S_{\text{Cleavage}} = 6.433$, metallo proteases: average $S_{\text{Cleavage}} = 6.652$, cysteine proteases: average $S_{\text{Cleavage}} = 6.820$). All protease types except for aspartic proteases therefore include specific as well as unspecific members. Thus, our study underlines the broadly accepted finding that protease substrate specificity is determined by subpocket interactions of the protease rather than directly at the catalytic site.

Conserved Substrate Promiscuity of Proteases within Same Clan

As apparent from Figure 8, the catalytic mechanism, does not discriminate specific from unspecific function. Rather, evolutionary related sub-groups sharing common catalytic mechanisms, but differing in three-dimensional fold are found to be similar in substrate promiscuity (see Figure 9). These clans within a catalytic class are not present in the test set for metallo proteases or aspartic proteases. All 13 metallo proteases in the test set belong to the MEROPS clan MA and all 5 aspartic proteases to the clan AA. Cysteine proteases spread over two distinct clans: 7 members (cathepsins, calpains and falcipains) belong to the CA papain clan, 3 others to clan CD, caspases. Serine proteases span three clusters of homologue proteases: 12 members are part of the PA clan (chymotrypsin-like proteases), containing besides serine proteases also cysteine proteases, that are not covered within the test set. Two members of the clan SF share the signalase fold, whilst four others share a subtilisin fold and thus belong to MEROPS clan SB. Signal peptidase complex is not assigned to a particular MEROPS protease clan.

Surprisingly, subdivision into homologue clans allows to subdivide proteases sharing the same catalytic mechanism into specific and unspecific subgroups. Cysteine proteases are divided into a more specific clan CD (average $S_{\text{Cleavage}} = 6.020$) and a relatively unspecific clan CA (average $S_{\text{Cleavage}} = 7.163$). Only caspases, known to be highly specific signaling proteases [76], represent clan CD in our test set, whereas calpains showing complex substrate specificities [41] with average $S_{\text{Cleavage}} = 7.106$, cathepsins with average $S_{\text{Cleavage}} = 7.113$ or falcipains with $S_{\text{Cleavage}} = 7.297$ are contained in clan CA. Falcipains of malaria-causing *Plasmodium falciparum* are involved in cytoskeleton and hemoglobin degradation [77] requiring unspecific substrate binding.

The same subdivision into specific and unspecific folds works for serine proteases that comprise clans of high specificity (clan SB: average $S_{\text{Cleavage}} = 5.429$), intermediate specificity (clan SF: average $S_{\text{Cleavage}} = 6.370$) as well as less specific proteases (clan PA: average $S_{\text{Cleavage}} = 6.779$). Standard deviations of cleavage entropies calculated within clan members are low (see Figure 9), suggesting intrinsically encoded limits for specific/non-specific behavior within the three-dimensional fold of the respective clans. This finding could be attributed to an intrinsic presence or absence of preorganized subpockets allowing for specific enzyme-substrate interactions.

Thus, the whole structure of protease clans has to be considered to shed light on the molecular origins of general protease cleavage

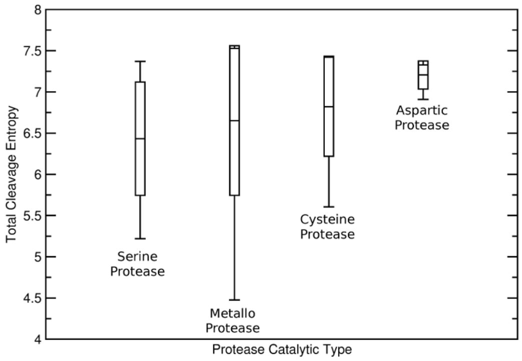


Figure 8. Total Cleavage Entropies of Protease Catalytic Types. Protease cleavage entropies indicate specific as well as unspecific members for each of the investigated protease catalytic machineries. As cleavage entropies (indicated by averages, maxima, minima and standard deviations) overlap between each of the types, the catalytic mechanism is found not to determine substrate specificity.
doi:10.1371/journal.pcbi.1003007.g008

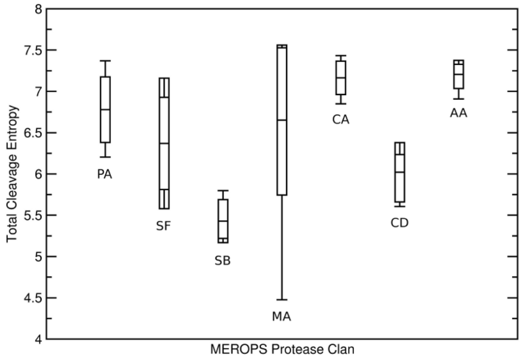


Figure 9. Total Cleavage Entropies of Protease Clans. Splitting of protease catalytic types into homologous protease clans allows to separate specific from unspecific members although they share a catalytic mechanism. Clan-wise total cleavage entropies are shown for MEROPS clans PA, SF, SB (all serine proteases), MA (metallo proteases), CA, CD (both cysteine proteases) and AA (aspartic proteases) with indicated averages, maxima, minima and standard deviations.
doi:10.1371/journal.pcbi.1003007.g009

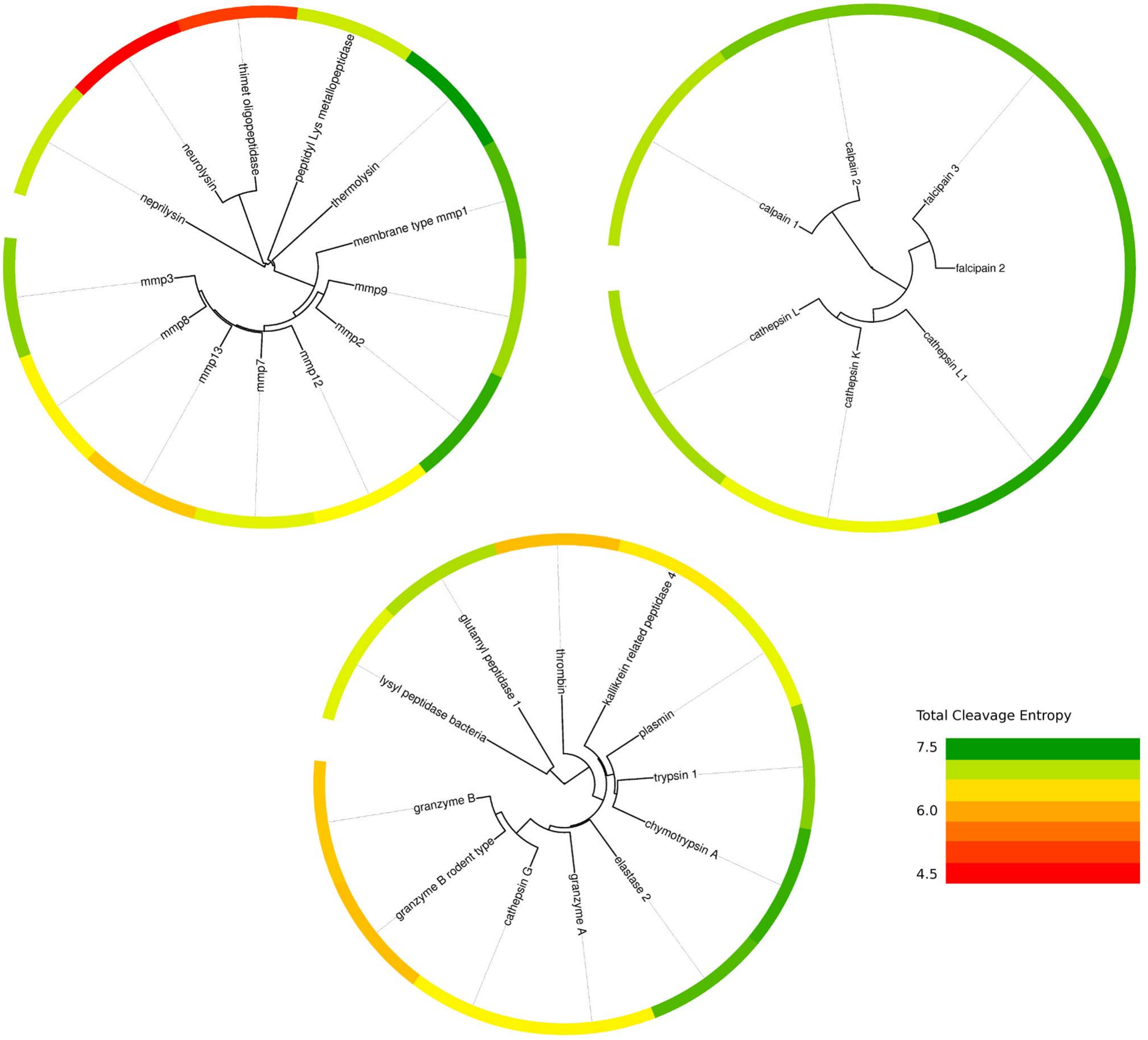


Figure 10. Phylogenetic trees of most prominent protease clans PA, CA and MA. Scattering of specific and unspecific behavior over respective evolutionary distances is apparent from the color-coded total cleavage entropy in all protease clans. Red fields indicate specific substrate recognition, whereas green fields mark unspecific proteases.

doi:10.1371/journal.pcbi.1003007.g010

spectra. Consistently, single mutations within specificity pockets of proteases are known to shift substrate spectra to other preferred substrates rather than to interchange specific and non-specific cleavage behavior. Nevertheless, a smooth interchange between specific and unspecific behavior including specialization and despecialization steps has been shown in case of granzymes [78], a class of serine proteases in clan PA.

Rapid Evolutionary Interchange of Specificity within Protease Clans

Further tracing the evolutionary development of protease specificity into particular protease clans arises the question, if evolutionary distance at sequence level is related to substrate specificity in these groups with conserved three-dimensional fold. Therefore, we performed a phylogenetic analysis for individual protease clans with more than five members contained in the test set (see Figure 10). MEROPS protease families are grouped in branches, confirming reasonability of presented phylogenetic trees. Whereas all members of clan PA belong to family S1, cysteine proteases spread over two distinct families: calpains are members of family C2 and are form a separate branch compared to all other proteases of the CA set that are part of the papain family C1. Metallo proteases belong to a wide-spread range of families: neprilysin is a singleton of family M13, neurolysin and thimet oligopeptidase of family M3 are nicely grouped in a separate branch. Two further singletons peptidyl-Lys metallo protease and thermolysin each form a separate tree branch for the families M35 and M4 respectively. All other members of clan MA are part of family M10, the matrix metallo peptidases, and are grouped into a broad branch separated from the other members.

Divergent evolution towards specific as well as unspecific members can be identified within all protease clans. Whereas a phylogenetic tree of metallo proteases of clan MA groups the highly specific members neurolysin and thimet oligopeptidase in a separate branch, indicating a close interplay between evolutionary distance and substrate specificity, this observation can not be extended to the whole set of proteases. The opposite holds even true in the MA clan for M10 family, where specific and unspecific members are grouped almost randomly compared to their evolutionary distance. The same complex behavior is found for cathepsins in clan CA: This branch includes the most specific member cathepsin L1 as well as the least specific member cathepsin K. Nevertheless, these members are grouped in closely related taxa indicating evolutionary proximity. Evolutionarily closely related proteases exhibit diverse substrate promiscuity in this protease group. Hence, protease evolution is capable of rapidly interchanging specific and non-specific substrate binding, implying a complicated relationship between protease sequence and substrate specificity.

The largest group of serine proteases of clan PA also groups specific and unspecific members in related taxa. E.g., cathepsin G and granzymes B of human and rodent origin exhibiting major different cleavage behavior are found as subbranch of closest evolutionary relation. Similarly, a branch including the rather specific signaling protease plasmin as well as the unspecific digestive enzymes trypsin 1 and chymotrypsin A, the most promiscuous members of this family, are grouped in close evolutionary proximity.

Implications of the Evolution of Protease Specificity

We therefore surmise that a detailed understanding of protease specificity is only in reach within an even smaller subset of homologue proteases, where changes in substrate specificity can be attributed to a limited set of amino acid mutations, and hence atom exchanges, in the binding region. We propose to join forces between computational and experimental groups to elucidate structural hot-spots crucial for binding specificity in particular protease folds. According to the observed small fluctuations in specificity within respective clans, a smaller set of homologous proteases should be suitable to allow such in-depth investigations.

The presented specificity metric "cleavage entropy" for proteases can be applied to map subpocket-wise specificity contributions based on experimental data to individual subpockets of proteases as well as to calculate an estimate of overall substrate specificity. Furthermore, the extension of subpocket-wise cleavage entropies to pairwise cleavage entropies facilitates the detection of subpocket cooperativities in proteases provided that a sufficient number of substrates for this two-dimensional analysis is known. Thereby, drug design targeting proteases will profit from a thorough understanding of specific interactions to achieve desired protease selectivity [79] for example in targeting matrix metallo proteases [80]. As parameters at the level of sequence [81], structure [18] and conformational flexibility [82] are known to influence protease specificity, a direct quantification of substrate promiscuity of proteases will help to distinguish individual contributions to this phenomenon [83] and thereby support structural biology, the rational design of protease specificity [84] and the emerging field of degradomics [85]. An extension of the information-theory based specificity mapping towards general protein-protein interfaces to assess specificity and hence druggability of the respective interface regions is envisaged.

A straight-forward interpretable specificity score generally applicable to all families of proteases was presented that confirms widely accepted rules of thumb for protease cleavage in a quantitative way. Calculated cleavage entropies purely based on amino acid frequencies in known substrates allow a straight-forward assessment of subpocket-wise substrate specificities. According to our specificity metric, the catalytic cleavage machinery and thus, protease class, does not discriminate specific and unspecific proteases. In contrast, homologue protease clans share intrinsic specific and non-specific properties suggesting that protease specificity is encoded directly in the shared three-dimensional protein fold. Within particular protease clans and folds, a small number of mutations can cause drastic alterations of substrate specificity. These subtle changes at sequence, structure and flexibility level, but heavily impacting substrate promiscuity, are thus of high interest for structural biology but challenging to predict.

Unlike classical rules-of-thumb for protease specificity, the quantification of subpocket-wise and overall substrate specificity provides a continuous metric for specificity rather than a 'yes'-or-'no' decision. The provided quantitative measure thus facilitates the comparison of the macromolecular descriptor "substrate specificity" with physicochemical, evolutionary and structural descriptors in protease recognition. Mapping of specificity to subpockets allows for intuitive visualization of structure-selectivity relationships in proteases and will thereby support the establishment of rules linking local protein structure and specificity.

Supporting Information

Table S1 Pairwise Cleavage Entropies of Trypsin. Interdependence in substrate readout of trypsin subpockets P4-P4′ is reflected quantitatively as pairwise cleavage entropies $S_{i,j}$. For comparison subpocket-wise cleavage entropies S_i are provided in the last row. Entropy values lower than 0.5 are highlighted in red, values between 0.5 and 0.85 in yellow. Besides readout of the P1 position, no pronounced cooperativity effect for trypsin can be observed.
(PDF)

Author Contributions

Conceived and designed the experiments: JEF SvG RGH MAM HGW KRL. Performed the experiments: JEF. Analyzed the data: JEF SvG RGH MAM GMS HGW KRL. Contributed reagents/materials/analysis tools: JEF SvG HGW KRL. Wrote the paper: JEF SvG RGH MAM GMS HGW KRL.

References

1. Hedstrom L (2002) Introduction: Proteases. Chem Rev 102: 4429–4429.
2. Turk B (2006) Targeting proteases: successes, failures and future prospects. Nat Rev Drug Discov 5: 785–799.
3. Puente XS, Sanchez LM, Overall CM, Lopez-Otin C (2003) Human and Mouse Proteases: A Comparative Genomic Approach. Nat Rev Genet 4: 544–558.
4. Davie EW, Fujikawa K, Kisiel W (1991) The Coagulation Cascade: Initiation, Maintenance, and Regulation. Biochemistry 43: 10363–10370.
5. Hengartner MO (2000) The biochemistry of apoptosis. Nature 407: 770–776.
6. Huber R, Bode W (1978) Structural Basis of the Activation and Action of Trypsin. Acc Chem Res 11: 114–122.
7. Schechter I, Berger A (1967) On the Size of the Active Site in Proteases: I. Papain. Biochem Biophys Res Commun 2: 157–162.
8. Tyndall JDA, Nall T, Fairlie DP (2005) Proteases Universally Recognize Beta Strand In Their Active Sites. Chem Rev 105: 973–999.
9. Madala PK, Tyndall JDA, Nall T, Fairlie DP (2010) Update 1 of: Proteases Universally Recognize Beta Strands In Their Active Sites. Chem Rev 110: PR1–PR31.
10. Hedstrom L (2002) Serine Protease Mechanism and Specificity. Chem Rev 102: 4501–4523.
11. Steitz TA, Henderson R, Blow DM (1969) Structure of crystalline alpha-chymotrypsin. 3. Crystallographic studies of substrates and inhibitors bound to the active site of alpha-chymotrypsin. J Mol Biol 2: 337–348.
12. Hedstrom L, Szilagyi L, Rutter WJ (1992) Converting trypsin to chymotrypsin – the role of surface loops. Science 5049: 1249–1253.

13. Perona JJ, Craik CS (1997) Evolutionary Divergence of Substrate Specificity within the Chymotrypsin-like Serine Protease Fold. J Biol Chem 48: 29987–29990.
14. Ma W, Tang C, Lai L (2005) Specificity of Trypsin and Chymotrypsin: Loop-Motion-Controlled Dynamic Correlation as a Determinant. Biophys J 89: 1183–1193.
15. Schellenberger V, Braune K, Hofmann HJ, Jakubke HD (1991) The specificity of chymotrypsin. Eur J Biochem 199: 623–636.
16. Bode W, Meyer E, Powers JC (1989) Human Leukocyte and Porcine Pancreatic Elastase: X-ray Crystal Structures, Mechanism, Substrate Specificity, and Mechanism-Based Inhibitors. Biochemistry 5: 1951–1963.
17. Lu D, Fütterer K, Korolev S, Zheng X, Tan K, et al. (1999) Crystal Structure of Enteropeptidase light Chain Complexed with an Analog of the Trypsinogen Activation Peptide. J Mol Biol 292: 361–373.
18. Perona JJ, Craik CS (1995) Structural basis of substrate specificity in the serine proteases. Protein Sci 4: 337–360.
19. Verspurten J, Gevaert K, Declercq W, Vandenabeele P (2009) SitePredicting the cleavage of proteinase substrates. Trends Biochem Sci 7: 319–323.
20. Diamond SL (2007) Methods for mapping protease specificity. Curr Opin Chem Biol 11: 46–51.
21. Poreba M, Drag M (2010) Current strategies for probing substrate specificity of proteases. Curr Med Chem 17: 3968–3995.
22. Harris JL, Backes BJ, Leonetti F, Mahrus S, Ellman JA, et al. (2000) Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. Proc Natl Acad Sci U S A 14:7754–7759.
23. Kleifeld O, Doucet A, Auf dem Keller U, Prudova A, Schilling O, et al. (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. Nat Biotechnol 3: 281–291.
24. Timmer JC, Enoksson M, Wildfang E, Zhu W, Igarashi Y, et al. (2007) Profiling constitutive proteolytic events in vivo. Biochem J 407: 41–48.
25. Matthews DJ, Wells JA (1993) Substrate Phage – Selection of Protease Substrates by Monovalent Phage Display. Science 5111: 1113–1117.
26. Boulware KT, Daugherty PS (2006) Protease specificity determination by using cellular libraries of peptide substrates (CliPS). Proc Nat Acad Sci U S A 20: 7583–7588.
27. Salisbury CM, Maly DJ, Ellman JA (2002) Peptide Microarrays for the Determination of Protease Substrate Specificity. J Am Chem Soc 124: 14868–14870.
28. Gosalia DN, Salisbury CM, Maly DJ, Ellman JA, Diamond SL (2005) Profiling serine protease substrate specificity with solution phase fluorogenic peptide microarrays. Proteomics 5: 1292–1298.
29. Schilling O, Overall CM (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nat Biotechnol 6: 685–694.
30. Schilling O, Auf dem Keller U, Overall CM (2011) Factor Xa subsite mapping by proteome-derived peptide libraries improved using WebPICS, a resource for proteomic identification of cleavage sites. Biol Chem 392: 1031–1037.
31. Rawlings ND, Barrett AJ, Bateman A (2010) MEROPS: the peptidase database. Nucleic Acids Res 38: D227–233. Database accession 11.10.2011
32. Rawlings ND, Barrett AJ, Bateman A (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 40: D343–350.
33. Newell NE (2011) Cascade detection for the extraction of localized sequence features; specificity results for HIV-1 protease and structure-function results for the Schellman loop. Bioinformatics 27: 3415–3422.
34. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 20: 6097–6100.
35. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K (2009) Improved visualization of protein consensus sequences by iceLogo. Nat Methods 11: 786–787.
36. Shannon CE (1948) A Mathematical Theory of Communication. Bell System Technical J 3: 379–423.
37. Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 4: 620–630.
38. Schneider TD, Stormo GD, Gold L (1986) Information Content of Binding Sites on Nucleotide Sequences. J Mol Biol 188: 415–431.
39. Nath A, Zientek MA, Burke BJ, Jiang Y, Atkins WM (2010) Quantifying and predicting the Promiscuity and Isoform Specificity of Small-Molecule Cytochrome P450 Inhibitors. Drug Metab Dispos 38: 2195–2203.
40. Nath A, Atkins WM (2008) A Quantitative Index of Substrate Promiscuity. Biochemistry 47: 157–166.
41. DuVerle DA, Mamitsuka H (2012) A review of statistical methods for prediction of proteolytic cleavage. Brief Bioinform 3: 337–349.
42. Igarashi Y, Eroshkin A, Gramatikova S, Gramatikoff K, Zhang Y, et al. (2007) CutDB: a proteolytic event database. Nucleic Acids Res 35: D546–D549.
43. Igarashi Y, Heureux E, Doctor KS, Talwar P, Gramatikova S, et al. (2009) PMAP: databases for analyzing proteolytic events and pathways. Nucleic Acids Res 37: D611–D618.
44. McCaldon P, Argos P (1988) Oligopeptide Biases in Protein Sequences and Their Use in Predicting Protein Coding Regions in Nucleotide-Sequences. Proteins 2: 99–122.
45. Gron H, Breddam K (1992) Interdependency of the Binding Subsites in Subtilisin. Biochemistry 31: 8967–8971.
46. Ng NM, Pike RN, Boyd SE (2009) Subsite cooperativity in protease specificity. Biol Chem 390: 401–407.
47. The Uniprot Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acid Res 40: D71–D75.
48. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 21: 2947–2948.
49. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 6: 276–277.
50. Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 3760: 279–284.
51. Letunic I, Bork P (2011) Interactive Tree of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 39 (Suppl 2): W475–W478.
52. DeLano WL (2008) The Pymol Molecular Graphics System. Available: http://pymol.org
53. Nar H, Bauer M, Schmid A, Stassen JM, Wienen W, et al. (2001) Structural Basis for Inhibition Promiscuity of Dual Specific Thrombin and Factor Xa Blood Coagulation Inhibitors. Structure 9: 29–37.
54. Bode W, Turk D, Karshikov A (1992) The refined 1.9-A X-ray crystal structure of D-Phe-Pro-Arg chloromethylketone-inhibited human a-thrombin: Structure analysis, overall structure, electrostatic properties, detailed active-site geometry, and structure-function relationships. Protein Sci 1: 426–471.
55. Fuller RS, Brake A, Thorner J (1989) Yeast prohormone processing enzyme (KEX2 gene product) is a Ca^{2+}-dependent serine protease. Proc Natl Acad Sci U S A 86: 1434–1438.
56. Page MJ, Macgillivray RTA, Di Cera E (2005) Determinants of specificity in coagulation proteases. J Thromb Haemost 3: 2401–2408.
57. Henrich S, Cameron A, Bourenkov GP, Kiefersauer R, Huber R, Lindberg I, et al. (2003) The crystal structure of the preprotein processing proteinase furin explains its stringent specificity. Nat Struct Biol 7: 520–526.
58. Overall CM (2002) Molecular Determinants of Metalloproteinase Substrate Specificity. Mol Biotechnol 22: 51–86.
59. Pirard B (2007) Insight into the structural determinants for selective inhibition of matrix metalloproteinases. Drug Discov Today 15–16: 640–646.
60. Visse R, Nagase H (2003) Matrix Metalloproteinases and Tissue Inhibitors of Metalloproteinases: Structure, Function, and Biochemistry. Circ Res 92: 827–839.
61. Turk BE, Huang LL, Piro ET, Cantlay LC (2001) Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. Nat Biotechnol 19: 661–667.
62. Auf dem Keller U, Prudova A, Gioia M, Butler GS, Overall CM (2010) A Statistics-based Platform for Quantitative N-terminome Analysis and Identification of Protease Cleavage Products. Mol Cell Proteomics 5: 912–927.
63. Thornberry NA, Lazebnik Y (1998) Caspases: Enemies Within. Science 281: 1312–1316.
64. Thornberry NA, Rano TA, Peterson EP, Rasper DM, Timkey T, et al. (1997) A Combinatorial Approach Defines Specifities of Members of the Caspase Family and Granzyme B. J Biol Chem 29: 17907–17911.
65. Wang KKW (2000) Calpain and caspase: can you tell the difference? Trends Neurosci 2: 20–26.
66. Choe Y, Leonetti F, Greenbaum DC, Lecaille F, Bogyo M, et al. (2006) Substrate Profiling of Cysteine Proteases Using a Combinatorial Peptide Library Identifies Functionally Unique Specificities. J Biol Chem 18: 12824–12832.
67. Paetzel M, Karla A, Strynadka NCJ, Dalbey RE (2002) Chem Rev 102: 4549–4579.
68. Choo KH, Tong JC, Ranganathan S (2008) Modeling Escherichia coli signal peptidase complex with bound substrate: determinants in the mature peptide influencing signal peptide cleavage. BMC Bioinformatics 9(Suppl I): S15.
69. Fikes JD, Barkocy-Gallagher GA, Klapper DG, Bassford PJ (1990) Maturation of Escherichia coli Maltose-binding Protein by Signal Peptidase I in Vivo. J Biol Chem 6: 3417–3423.
70. Dunn BM, Rao M (2004) Human immunodeficiency virus 1 retropepsin. In: Barrett AJ, Rawlings ND, Woessner JF, editors. Handbook of Proteolytic Enzymes, 2 ed. pp.144–154.
71. Van Den Burg B, Eijsink V (2004) Thermolysin and related *Bacillus* metallopeptidases. In: Barrett AJ, Rawlings ND, Woessner JF, editors. Handbook of Proteolytic Enzymes, 2 ed. pp. 374–387.
72. Ambler RP, Meadway RJ (1968) The Use of Thermolysin in Amino Acid Sequence Determination. Biochem J 108: 893–895.
73. Trusek-Holownia A (2003) Synthesis of ZAlaPheOMe, the precursor of bitter dipeptide in the two-phase ethyl acetate/water system catalysed by thermolysin. J Biotechnol 102: 153–163.
74. Checler F, Vincent JP, Kitabgi P (1986) Purification and Characterization of a Novel Neurotensin-degrading Peptidase from Rat Brain Synaptic Membranes. J Biol Chem 24: 11274–11281.
75. Cunha FM, Bertl DA, Ferreira ZS, Klitzke CF, Markus RP, et al. (2008) Intracellular Peptides as Natural Regulators of Cell Signalling. J Biol Chem 36: 24448–24459.
76. Talanian RV, Quinlan C, Trautz S, Hackett MC, Mankovich JA, et al. (1997) Substrate Specificities of Caspase Family Proteases. J Biol Chem 15: 9677–9682.
77. Blackman MJ (2008) Malarial proteases and host cell egress: an 'emerging' cascade. Cell Microbiol 10: 1925–1934.
78. Wouters MA, Liu K, Riek P, Husain A (2010) A Despecialization Step Underlying Evolution of a Family of Serine Proteases. Mol Cell 12: 343–354.
79. Drag M, Salvesen GS (2010) Emerging principles in protease-based drug discovery. Nat Rev Drug Discov 9: 690–701.

80. Devel L, Czarny B, Beau F, Georgiadis D, Stura E, et al. (2010) Third generation of matrix metalloprotease inhibitors: gain in selectivity by targeting the depth of the S_1' cavity. Biochimie 92: 1501–1508.
81. Neurath H (1984) Evolution of Proteolytic Enzymes. Science 4647: 350–357.
82. Wallnoefer HG, Lingott T, Gutierrez JM, Merfort I, Liedl KR (2010) Backbone Flexibility Controls the Activity and Specificity of a Protein-Protein Interface: Specificity in Snake Venom Metalloproteases. J Am Chem Soc 132: 10330–10337.
83. Babtie A, Tokuriki N, Hollfelder F (2010) What makes an enzyme promiscuous? Curr Opin Chem Biol 2: 200–207.
84. Di Cera E (2008) Engineering protease specificity made simple, but not simpler. Nat Chem Biol 5: 270–271.
85. Lopez-Otin C, Overall CM (2002) Protease degradomics: a new challenge for proteomics. Nat Rev Mol Cell Biol 3: 509–519.