Qi et al. BMC Bioinformatics (2017) 18:438
DOI 10.1186/s12859-017-1851-1

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Block-based characterization of protease specificity from substrate sequence profile

Enfeng Qi[1], Dongyu Wang[2], Bo Gao[1], Yang Li[1] and Guojun Li[1*]

Abstract

Background: The mechanism of action of proteases has been widely studied based on substrate specificity. Prior research has been focused on the amino acids at a single amino acid site, but rarely on combinations of amino acids around the cleavage bond.

Results: We propose a novel block-based approach to reveal the potential combinations of amino acids which may regulate the action of proteases. Using the entropies of eight blocks centered at a cleavage bond, we created a distance matrix for 61 proteases to compare their specificities. After quantitative analysis, we discovered a number of prominent blocks, each of which consists of successive amino acids near a cleavage bond, intuitively characterizing the site cooperation of the substrate sequences.

Conclusion: This approach will help in the discovery of specific substrate sequences which may bridge between proteases and cleavage substrate as more substrate information becomes available.

Keywords: Protease, Block, Entropy, Site cooperation

Background

Proteases are a category of enzymes capable of hydrolyzing peptide bonds and irreversibly modifying functions of substrate proteins. These hydrolyzations and modifications are essential for cell growth and differentiation [1, 2]. Recognition of the target substrate of a protease depends partly on the complementation between the protease active site and the sequence surrounding the scissile bond in the substrate. Proteases have pockets that accommodate substrate residues. Substrate sequences that bind the pockets are indexed by P_4, P_3, P_2, P_1, P_1', P_2', P_3', P_4' in order from N-terminal to C-terminal following the convention of Schechter and Berger [3].

Some proteases show strict specificities on the cleavage sequences of the substrates. For example, trypsin 1 requires Lys and Arg at the P_1 site [4], and granzyme B shows strict specificity for Asp at the P_1 site [5]. The specificity of protease has been widely used not only in identifying the biologically relevant substrates, but also in applying protease to site-specific proteolysis [6, 7]. Proteases participate in various disease processes, exhibiting a potentially huge future application in the design of new drug targets for enzyme [8, 9] and protease inhibitors [10]. Although all the proteases function in hydrolyzing peptide bonds, almost all are linked to a particular cleavage pattern [11].

The MEROPS database is a manually curated information resource for peptidases [12]. According to MEROPS, more than 10,000 known substrates are profiled for some proteases [13], so it is necessary to develop an approach to map the abundant substrate-sequence information to specificities of proteases to highlight the enzymatic preferences, especially for specific catalytic types [14]. Integrating features of substrate sequences characteristics, PoPS [15] and PROSPER [16] are proposed to predict protease substrate cleavage sites. A well-designed approach of identifying the specificity of the protease will contribute to a better method of predicting the substrate cleavage site.

Previous analyses of protease cleavage data, such as visualized sequence logos [17], iceLogo [18], heat maps [19] and several techniques [20–22], have been focused on qualitative interpretation. Using LC-MS/MS sequencing [23], a simple and rapid multiplex substrate profiling method was presented to demonstrate the substrate specificity. Further measures include using fluorogenic substrates [4], specific labeling techniques of N-terminal

* Correspondence: guojunsdu@gmail.com
[1]School of Mathematics, Shandong University, Jinan 250100, China
Full list of author information is available at the end of the article

[24, 25], and proteome-derived peptide libraries [26–30]. Fuchs [31] developed a method to quantify protease specificity and rank proteases with the cleavage entropy of a single position. Several quantitative measures were developed [32–35], in which the specificities of proteases were shown by the occurrences of amino acids at the binding sites. As mentioned by Schilling and Overall [19] in profiling the specificity of the MMP2, the preferred amino acid residues at different site may cooperate in the hydrolysis process, therefore, it is critically important to elucidate the hydrolyzation process by the closely cooperative relationship of successive positions on the substrate sequences.

In this study, we designed a novel approach to present the protease specificity based on blocks which are composed by successive amino acids from the substrate sequence. The essential difference between our approach and previous ones lies in that we characterize the specificity of proteases based on successive amino acids rather than a single binding site. This new approach could more reliably identify protease specificity by considering cooperation among the successive sites of the substrate peptides during the hydrolyzation process.

Methods

Data extraction

The dataset is composed of 61 proteases for analysis as described by Fuchs [35]. The cleavage information from all experimental sources is obtained from the MEROPS database [12] and is updated according to MEROPS 10.0. This study focuses on the protease specificity directly on the active sites, ignoring differences in allosteric sites and exosite interactions. Among the data, signal peptidase complex (XS26.001) has been deleted from the dataset since the complex contains two peptidases, and it is not possible to assign a particular cleavage to one activity due to not a single component.

Greedy algorithm for filtering the data

First, the substrate sequence with less than two amino acids is filtered out. Then all of substrate sequences left primarily are aligned pairwise. Starting from sequences with the maximum number of similar amino acids, we remove redundant sequences by greedy algorithm [36] to make sure that there is no pair of sequences whose similarity is greater than or equal to 0.875. Therefore, there are at least two different amino acid residues between any two remaining substrate sequences.

Construction of blocks

The indices of the residues in the substrate sequence are centered on the cleavage bond and extended to both sides incrementally, namely P_4, P_3, P_2, P_1, P_1', P_2', P_3', P_4'.We define a set of eight blocks, denoted by B = (B_4, B_3, B_2, B_1, B_1', B_2', B_3', B_4'), where B_1 (B_1') is a vector of amino acids occurred in respective substrate sequences at P_1 (P_1'); B_2 (B_2') is a vector of two successive amino acids occurred in respective substrate sequences at P_2, P_1 (P_1', P_2'); B_3 (B_3') is a vector of three successive amino acids occurred in respective substrate sequences at P_3, P_2, P_1 (P_1', P_2', P_3'); B_4 (B_4') is a vector of four successive amino acids occurred in respective substrate sequences at P_4, P_3, P_2, P_1 (P_1', P_2', P_3', P_4'). The construction of blocks is shown in Fig. 1.

Calculation of entropy

Information entropy was firstly proposed by Shannon [37]. The block-based entropy information of the substrate reflects the specific or broad property of the protease. The randomness of the block-based substrate information is given by the entropy:

$$\mathrm{E}_k(\text{or } \mathrm{E}_k^{'}) = -\sum p_i \log_2 p_i \quad (k = 1,\ 2,\ 3,\ 4) \qquad (1)$$

where p_i is the frequency of a component in block B_k (B_k'). Consequently, we can get the entropy of the block B as E = (E_4, E_3, E_2, E_1, E_1', E_2', E_3', E_4').

Calculation of distance matrix

A distance matrix is created by pairwise comparison of all 61 proteases' cleavage bonds. The distance between two proteases is calculated by the Euclidean distance of the total entropies calculated as follows:

$$d(P,Q) = \sqrt{\sum_{i=1}^{4}\left[\mathrm{E}_i(P) - \mathrm{E}_i(Q)\right]^2 + \sum_{i=1}^{4}\left[\mathrm{E}_i^{'}(P) - \mathrm{E}_i^{'}(Q)\right]^2} \qquad (2)$$

Where E_i (P) and E_i'(P) are the entropies of blocks B_i and B_i' of protease P respectively. This yields a symmetric distance matrix. The elements on the diagonal are 0, which is the distance of identical proteases.

Principal components analysis

All the eight blocks for each of 61 proteases are used for principal components analysis (PCA). Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy is computed as 0.733 which indicates that the sample size is sufficient for the application of PCA. The PCA is performed in SPSS 19.0 (SPSS Inc., Chicago, IL, USA) with the correlation method and Varimax with Kaiser Normalization as the rotation method.

Fisher's exact test

Fisher's exact test [38] is used in calculating the p-value of combinations. We simulated the substrates according to the frequencies of the amino acids, and repeated 1000 times for the prominent combinations in each block. As

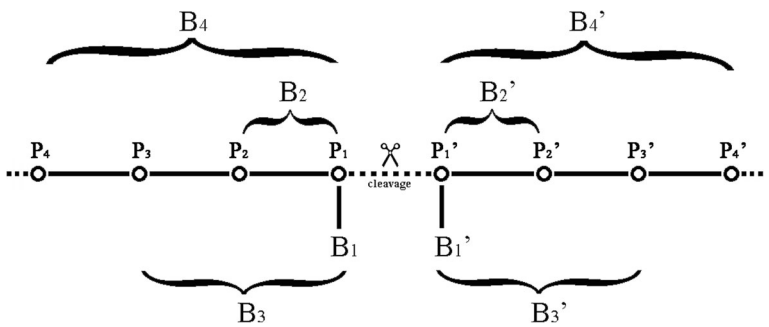

Fig. 1 A schematic diagram of construction of different blocks. The blocks of successive amino acids are denoted from N-terminal to C-terminal, so that block B_1 represents the P_1 site, block B_1' represents the P_1' site, block B_2 represents the successive sites of P_2 and P_1, and block B_2' represents the successive sites of P_1' and P_2' and so on. For example, block B_2 LeuLys implies Leu at the site P_2 and Lys at the site P_1, and block B_2' PheArg implies Phe at the site P_1' and Arg at the site P_2'. Other blocks may be deduced similarly

the false positive would be a waste of time, the Bonferroni correction [39] is used for the p-value threshold by $p < 0.05/N$, where N is the number of different kinds of combinations. For one combination occurring in the input substrate sequences, we consider the number of sequences containing this combination in both experiment and background sequences. We make the null hypothesis that there's no difference between proportions of sequences containing this combination in experiment and background sequences. The combination with significance level $p < 0.05/N$, occurring more than half of 1000 times, is regarded as a prominent combination. If a combination is significant, then the null hypothesis is rejected. The data might look like Table 1. The probability of obtaining any such set of values if given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\quad b!\quad c!\quad d!\quad n!} \tag{3}$$

where $\binom{n}{k}$ is the binomial coefficient and the symbol ! indicates the factorial operator. The software package of methods can be obtained in Additional file 1.

Table 1 2 × 2 contingency table for Fisher's exact test

	Yes	No	Row total
Experiment data	a	b	$a + b$
Background data	c	d	$c + d$
Column total	$a + c$	$b + d$	$a + b + c + d$

Creation of sequence profile

To depict the substrate preferences at different sites, the data of substrate sequences after removing the redundancy is submitted to Weblogo [17, 40] to generate sequence profiles of substrate cleavage site.

Results

Distance character of 61 proteases

The entropies of eight blocks B_4, B_3, B_2, B_1, B_1', B_2', B_3', B_4' are calculated and denoted as E_4, E_3, E_2, E_1, E_1', E_2', E_3', E_4' correspondingly (Additional file 2: Table S1). There are three blocks with entropy 0, including, caspase 6 with $E_1 = 0$ implying the unique amino acid Asp at site P_1; peptidyl-Lys metallopeptidase with $E_1' = 0$ implying the unique amino acid Lys at site P_1'; lysyl peptidase with $E_1 = 0$ implying the unique amino acid Lys at site P_1.

We obtained a distance matrix (Additional file 2: Table S2) by calculating the distances of the entropies of eight blocks between 61 proteases. We found obvious distinctions between proteases. The maximum entry 16.630 in the matrix is the distance between the proteases neurolysin and trypsin 1, with their corresponding entropies being shown in Fig. 2a. From that, all of the block entropies except B_1 of trypsin 1 are higher than the corresponding block entropies of neurolysin. The fundamental difference between neurolysin and trypsin 1 is their different activities. Where neurolysin is an oligopeptidase unable to cleave proteins [41], trypsin 1 is an endopeptidase [4]. Another factor is the great gap in the numbers of distinct substrate sequences between neurolysin (45) and trypsin 1 (9014). Excluding the diagonal entries, the minimum entry 0.125 in the matrix is the distance between the proteases PCSK4 and PCSK6, with their corresponding entropies being shown in Fig. 2b. This is due to a large amount of similar blocks between them.

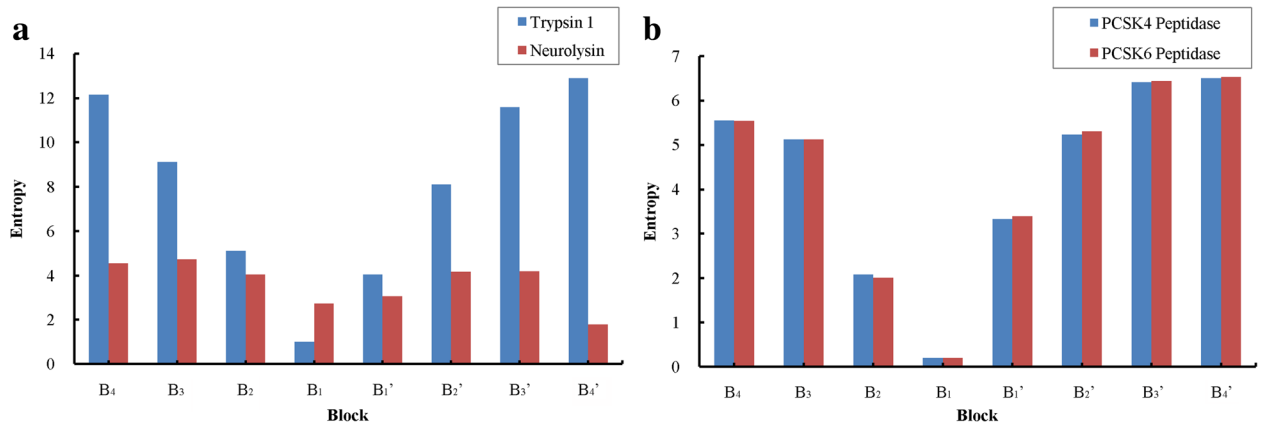

Fig. 2 Comparisons of eight entropies of proteases with maximum and minimum distance. **a** Entropy distributions of eight blocks from proteases neurolysin and trypsin 1 with the maximum distance in the distance matrix. **b** Entropy distributions of eight blocks from proteases PCSK4 and PCSK6 with the minimum distance in the distance matrix

Principal components analysis

The entropies of eight blocks reveal the complexity of different combination types. In order to mine the blocks which play the crucial role in the specificity recognition of substrate sequence, we used principal components analysis.

The distribution of eight different eigenvalues is shown in a scree plot (Additional file 2: Figure S1.). Three principal components (PC1: the first principal component; PC2: the second principal component; PC3: the third principal component) are obtained according to the principle of eigenvalues more than 1. Among the three principal components, PC1, PC2 and PC3 contribute 57.938%, 23.284% and 15.960% to the total variance respectively, and the cumulative contribution of three principal components is 97.183% (Additional file 2: Table S3). Thus, the three principal components may represent the main features in the recognition of substrate specificities of different proteases.

PC1 shows a strongly positive correlation with E_4, E_3, E_2', E_3', E_4' demonstrated by the principal components load matrix (Additional file 2: Table S4). The lower the entropies are, the more prominent blocks there will be at the corresponding binding sites. As E_1, E_2 and E_1' possess a weak correlation with the composition of PC1, the corresponding B_1, B_2 and B_1' are most likely to have the prominent blocks. PC2 correlates with E_1 and E_2 (Additional file 2: Table S4). From the scatter plot of PC2 versus PC1 in Fig. 3, note that PC2 separates metallo proteases from serine proteases approximately. Almost all of the proteases from metallo and aspartic proteases are above the zero of the vertical axis implying a positive correlation with PC2.

Block-based sequence profile

Our algorithm has uncovered a number of prominent blocks in different proteases. The proportions of prominent combinations in the substrate at each block are presented by different shades of green in the heat map (Fig. 4), indicating that a large number of significant combinations can be analyzed with this approach. For each block, the proportions of proteases possessing a prominent block in 61 proteases is demonstrated in Fig. 5, the ratios at B_2, and B_2' are higher than those at B_4, B_3, B_3' and B_4', implying that the amino acids close to the cleavage bond cooperate more preferably than those far away.

There are a few prominent blocks from prime side. For instance, except for strict specificity for Lys at the P_1' site, peptidyl-Lys metallopeptidase has block B_2' with LysGlu = 179 from 1869 substrates, and signal peptidase

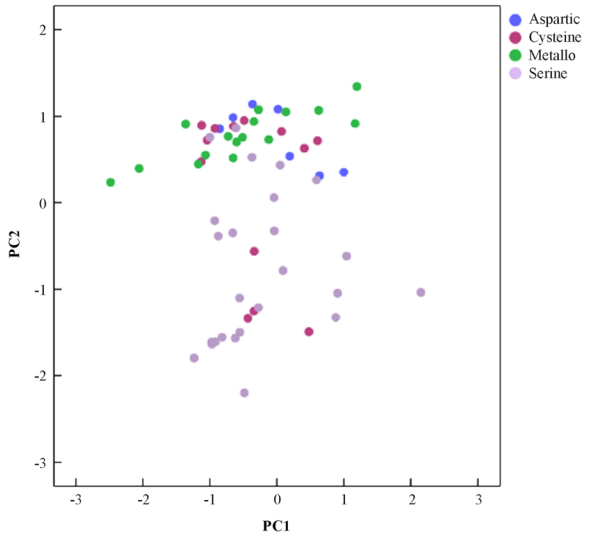

Fig. 3 Scatter plot of principal component analysis from PC1versus PC2. The selected data is grouped into four types according to the MEROPS database, including aspartic, cysteine, metallo and serine. Coloring according to catalytic types, aspartic protease: blue; cysteine protease: red; metallo protease: green; serine protease: pink

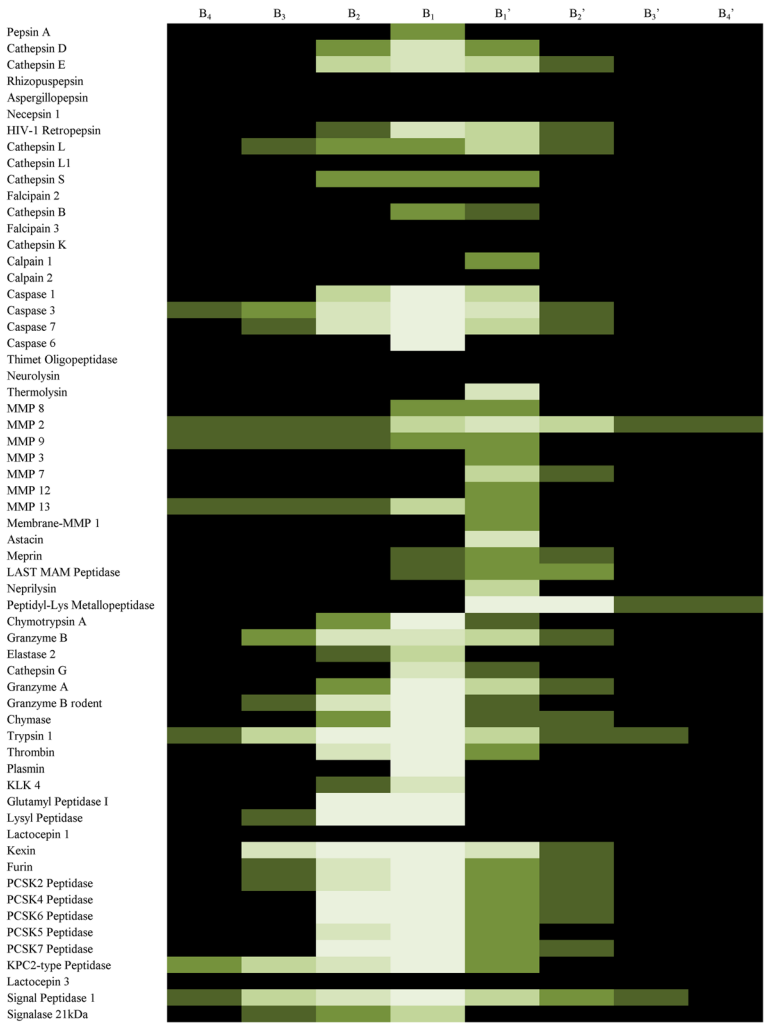

Fig. 4 Heat map of prominent combinations in each block. Five shades are shown ranging from darkest green (less 20% of substrates) to brightest green (greater 80% of substrates), and black background indicates no prominent combination in the block

1 has block B_3' with AlaGluAla = 19 from 297 substrates (The number behind the equal sign represents the amount of combination of amino acids in the corresponding block).

Meanwhile, a few blocks from non-prime side show the specificity. For example, kexin has block B_2 LysArg = 147 from 171 substrates. With caspase 3 having 571 substrates, besides the prominent block B_3 GluValAsp = 43, we still find the prominent block B_4 AspGluValAsp = 19.

Some proteases show the specificity at site P_1, and the prominent B_2 blocks (Table 2) are apparent in the sequence logos shown in Fig. 6a. For example, B_2 block ValAsp in caspase 3, and B_2 block LysArg in kexin, furin and PCSK6 peptidase. However, in the sequence logos shown in Fig. 6b, there are two or more amino acids in the binding sites P_1 and P_2, which indicates the preference rather than the strict specificity. As listed in Table 2, the top three amino acids of HIV-1 retropepsin at sites P_2 and P_1 are Val, Glu, Ile and Leu, Phe, Tyr, respectively, yet the prominent block B_2 with the highest number of combination are GluLeu = 35. For MMP2, the top three amino acids at sites P_2 and P_1 are Ala, Ser, Gly and Ala, Gly, Asn, respectively, and the prominent block B_2 with the highest number of combination is AlaAla = 100. For MMP 9, amino acids on the top at sites P_2 and P_1 are all unpolar amino acids such as Ala, Gly, Pro and Gly, Ala, Pro respectively, yet the top one Gly at the site P_1 has the preference of Pro at the site P_2 forming the prominent block B_2 ProGly = 25, and Pro at the site P_2 shows no preference of the top amino acids at the site P_1 except Gly in the formed block B_2.

Some blocks B_2' show the similar combination property as in blocks B_2. For example, the top three amino acids of HIV-1 retropepsin at sites P_1' and P_2' are Leu, Val, Phe and Glu, Val, Ala, respectively, yet the

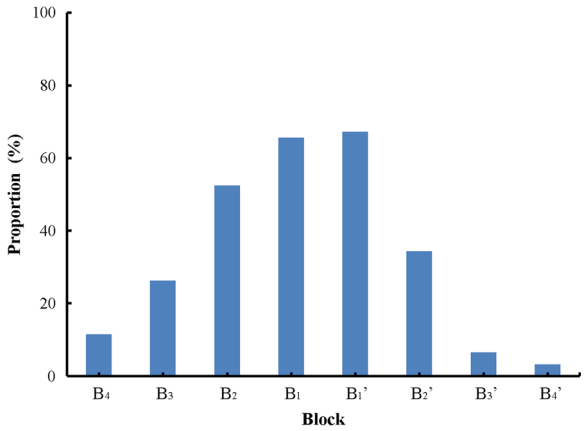

Fig. 5 The proportion of proteases with prominent blocks. The horizontal axis shows eight blocks B_4, B_3, B_2, B_1, B_1', B_2', B_3' and B_4'. The vertical axis shows the proportions of proteases with significant block in 61 proteases. The proportions from B_4 to B_4' are 11.475%, 26.230%, 52.459%, 65.574%, 67.213%, 34.426%, 6.557% and 3.279% respectively

prominent block B_2' with the highest number of combination are LeuAla = 33. For LAST_MAM peptidase, amino acids on the top at sites P_1' and P_2' are Asp, Ala, Glu and Pro, Ala, Glu respectively, yet the top one Asp at the site P_1' shows no preference of the top amino acid Pro at the site P_2', and the prominent block B_2' with the highest number of combination is AlaPro = 44 from 429 substrates. For the proteases which cleavage sites possess two or more preferred residues, the prominent combinations in the blocks reflect the cooperation of the residues in one position with other positions, characterizing the specificity of proteases detailedly.

Discussion

Some specificities of certain proteases have been determined, such as trypsin 1 [4], caspase 3 [42], kexin [43], furin [44] and so on. However, by focusing on single positions and not taking into consideration the interaction of adjacent amino acids, the study of substrate specificity is too limited.

Table 2 The top prominent B_2 blocks of proteases listed in Fig. 6

Protease	P_2	P_1	Block B_2
(a) The top prominent B_2 blocks of proteases listed in Fig. 6a			
Caspase 3	Val	Asp	ValAsp
Kexin	Lys	Arg	LysArg
Furin	Lys, Arg	Arg	LysArg, ArgArg
PCSK6 peptidase	Lys, Arg	Arg	LysArg, ArgArg
(b) The top prominent B_2 blocks of proteases listed in Fig. 6b			
HIV-1 Retropepsin	Val, Glu, Ile	Leu, Phe,Tyr	GluLeu, ValLeu
MMP2	Ala, Ser, Gly	Ala, Gly, Asn	AlaAla, SerGly
MMP9	Ala, Gly, Pro	Gly, Ala, Pro	ProGly, AlaAla

Taking the cooperation of amino acids into account, we propose a quantitative method to characterize substrates specificity of different proteases. By calculating entropies of different blocks, some distinctions of substrates between different proteases can be conceived (Fig. 2a). The principal component analysis gives evidence on the existence of blocks which play the crucial role in the specificity recognition of substrate sequence, and most of them are block B_2, B_1 and B_1'. This is confirmed by the statistical analysis showing the ratios at B_2, B_1, and B_1' are higher than those in other blocks (Fig. 5). With Fisher's exact test, a number of prominent blocks of different proteases have been discovered. For example, blocks B_2 in kexin and furin are consistent with the previous discovery that both of the proteases cleave after dibasic residues [45]. Other block B_2, e.g. GluLeu in HIV-1 retropepsin, AlaAla in MMP2 and ProGly in MMP 9, are more likely to reflect the preferences and the cooperation of the successive amino acids in the substrate sequences which could not be found previously.

Cathepsin B is an endopeptidase and as an exopeptidase acts as a peptidyl-dipeptidase, releasing a dipeptide from the C-terminus of a protein or peptide. As no distinction is made in MEROPS between cleavages resulting from either activity, a view of the endopeptidase activity would be clear if the substrates of the exopeptidase activity were filtered out.

From the specificity matrix in MEROPS and the heat map [19], the preference of the protease is shown by the amino acids at one single binding site. However, it will not show the combinations of amino acids if proteases show multiple preferences at each binding site. Our method indicates interactions of different compositions of successive amino acids which can't be obtained previously. For example, MMP9 has preferences for Ala, Gly and Pro at the site P_2, Gly, Ala and Pro at the site P_1 from the specificity matrix, yet the combination is clear using our method, such as ProGly, AlaAla in block B_2. Whether a prominent combination exists in a block is obviously presented in the heat map of prominent combinations in each block (Fig. 4). These findings of specific blocks will shed light on future experiments and further investigation of proteolytic specificity.

Although in this study we only focused on the specificity of selected proteases, the method would be applicable to other proteases for mining the specificity pattern of substrates. In conclusion, we can obtain more substrate specificity patterns by site cooperation as more and more substrates data becomes available. Further investigations of the substrate specificity will be important to reveal the hydrolyzation mechanism of proteases.

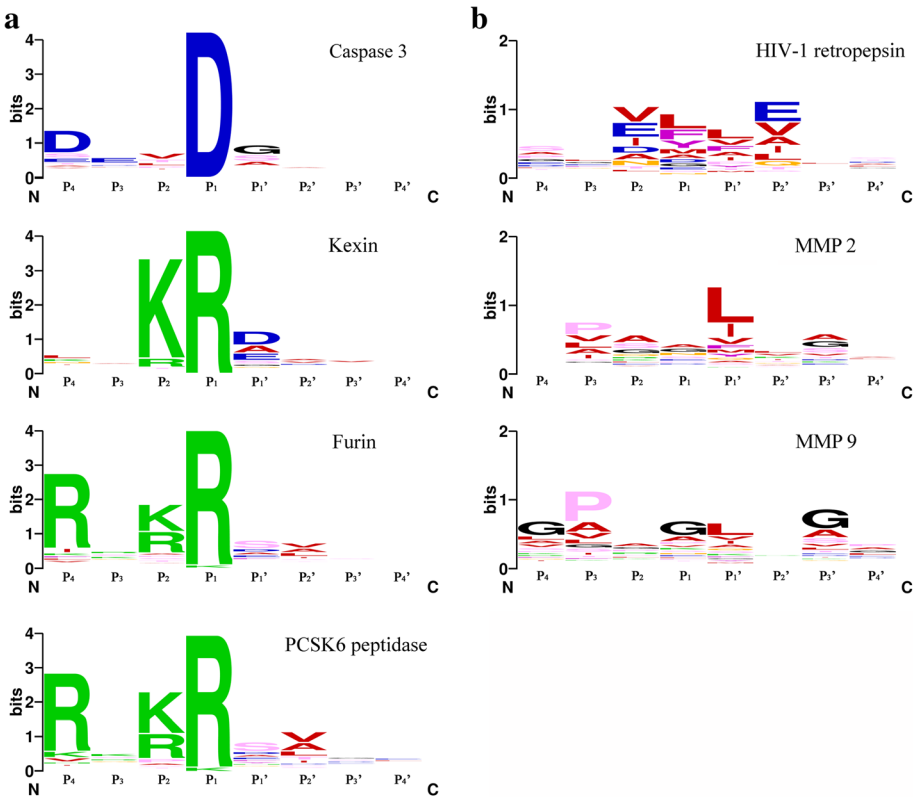

Fig. 6 Cleavage site sequence logos and the prominent B_2 blocks of proteases. **a** The sequence logos of caspase 3, kexin, furin and PCSK6 peptidase, which have obvious specificity at the site P_1. **b** The sequence logos of HIV-1 retropepsin, MMP 2 and MMP 9, which have multiple preferences at sites P_1 and P_2

Conclusions

Generally, the design of experiments and the description of the specificity of the protease are based on the assumption that the process of binding amino acid residue to the corresponding subsite is independent. However, it is not exactly true and the binding of amino acid residues at one site can more or less influence the binding at other subsites. It is essential to take the site cooperation into consideration for understanding fully the active site.

Our approach provides a new framework for dealing with the specificity pattern of substrates of the proteases. The combinations of site cooperation in the substrates offer a new sight in mining the specificity of the protease. We successfully find the significant blocks B_2 in kexin and furin which are consistent with the previous discovery that both of the proteases cleave after dibasic residues. Other significant combinations found by the new approach could be more reliable to capture the activity of the active site. In principle, this method is useful for the further research relying on the substrate dataset, such as the identification of the novel substrate and the design of the inhibitor for the protease.

Additional files

Additional file 1: Software package. A .tar.gz file that contains Perl and C++ scripts and an example to illustrate our approach. The package also includes a manual file (txt) for the instruction of the software. (GZ 11 kb)

Additional file 2: Supplementary Information. A .pdf file including Supplementary Tables and Figures. (PDF 65 kb)

Acknowledgements

The authors thank the editorial staff for their help in editing this manuscript and thank the anonymous reviewers for their suggestions and comments to improve the manuscript.

Funding

This work was supported by National Science Foundation (NSF Grant No.1553680), and National Science Foundation of China (NSFC Grant No. 61432010, 61,272,016 and 31,571,354).

Availability of data and materials

The datasets used in this research are available at http://www.merops.ac.uk.

Authors' contributions

E.Q and G.L conceived and designed the approach. E.Q, B.G and Y.L implemented the software. D.W and Y.L performed the data analysis. E.Q wrote the manuscript. G.L contributed to revise the manuscript. All author approved the final version of this manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's note

Author details
[1]School of Mathematics, Shandong University, Jinan 250100, China. [2]The State Key Laboratory of Microbial Technology, Shandong University, Jinan 250100, China.

Received: 15 June 2017 Accepted: 26 September 2017
Published online: 03 October 2017

References

1. Turk B, Turk D, Turk V. Protease signalling: the cutting edge. EMBO J. 2012; 31(7):1630–43.
2. López-Otín C, Bond JS. Proteases: multifunctional enzymes in life and disease. J Biol Chem. 2008;283(45):30433–7.
3. Schechter I, Berger A. On the size of the active site in proteases. I Papain Biochem Bioph Res Co. 1967;27(2):157–62.
4. Harris JL, Backes BJ, Leonetti F, Mahrus S, Ellman JA, Craik CS. Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. Proc Natl Acad Sci U S A. 2000;97(14):7754–9.
5. Waugh SM, Harris JL, Fletterick R, Craik CS. The Structure of the Pro-Apoptotic Protease Granzyme B Reveals the Molecular Determinants of its Specificity. Nat Struct Biol. 2000;7(9):762–5.
6. Denning DW, Anderson MJ, Turner G, Latgé JP, Bennett JW. Sequencing the Aspergillus fumigatus genome. Lancet Infect Dis. 2002;2(4):251–3.
7. López-Otín C, Overall CM. Protease degradomics: a new challenge for proteomics. Nat Rev Mol Cell Bio. 2002;3(7):509–19.
8. Turk B. Targeting proteases: successes, failures and future prospects. Nat Rev Drug Discov. 2006;5(9):785–99.
9. Lopez-Otin C, Matrisian LM. Emerging roles of proteases in tumour suppression. Nat Rev Cancer. 2007;7(10):800–8.
10. Liu H, Shi X, Guo D, Zhao Z, Yimin. Feature Selection Combined with Neural Network Structure Optimization for HIV-1 Protease Cleavage Site Prediction. Biomed Res Int. 2015;2015:263586.
11. Hedstrom L.Introduction: proteases. 2002;102(12):4429.
12. Rawlings ND, Barrett AJ, Finn R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res Nucleic Acids Res. 2016;44(D1):D343–50.
13. Rawlings ND. Peptidase specificity from the substrate cleavage collection in the MEROPS database and a tool to measure cleavage site conservation. Biochimie. 2016;122:5–30.
14. Song J, Tan H, Boyd SE, Shen H, Mahmood K, Webb GI, Akutsu T, Whisstock JC, Pike RN. Bioinformatic approaches for predicting substrates of proteases. J Bioinforma Comput Biol. 2011;9(1):149–78.
15. Boyd SE, Pike RN, Rudy GB, Whisstock JC, Garcia de la Banda M. PoPS: a computational tool for modeling and predicting protease specificity. J Bioinforma Comput Biol. 2005;3(3):551–85.
16. Song J, Tan H, Perry AJ, Akutsu T, Webb GI, Whisstock JC, Pike RNPROSPER. an integrated feature-based tool for predicting protease substrate cleavage sites. PLoS One. 2012;7(11):e50300.
17. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–100.
18. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. Improved visualization of protein consensus sequences by iceLogo. Nat Methods. 2009;6(11):786–7.
19. Schilling O, Overall CM. database-searchable peptide libraries for identifying protease cleavage sites. Nat Biotechnol. 2008;26(6):685–94.
20. Poreba M, Drag M. Current strategies for probing substrate specificity of proteases. Curr Med Chem. 2010;17(33):3968–95.
21. Huesgen PF, Overall CM. N- and C-terminal degradomics: new approaches to reveal biological roles for plant proteases from substrate identification. Physiol Plant. 2012;145(1):5–17.
22. Van Damme P, Staes A, Bronsoms S, Helsens K, Colaert N, Timmerman E, Aviles FX, Vandekerckhove J, Gevaert K. Complementary positional proteomics for screening substrates of endo- and exoproteases. Nat Methods. 2010;7(7):512–5.
23. O'Donoghue AJ, Eroy-Reveles AA, Knudsen GM, Ingram J, Zhou M, Statnekov JB, Greninger AL, Hostetter DR, Qu G, Maltby DA, Anderson MO, Derisi JL, McKerrow JH, Burlingame AL, Craik CS. Nat Methods 2012;9(11):1095–100.
24. 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(5):866–76.
25. Kleifeld O, Doucet A, Prudova A, Auf dem Keller U, Gioia M, Kizhakkedathu JN, Overall CM. Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. Nat Protoc. 2011;6(10):1578–611.
26. Boulware KT, Daugherty PS. Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). Proc Natl Acad Sci U S A. 2006;103(20):7583–8.
27. Turk BE, Huang LL, Piro ET, Cantley LC. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. Nat Biotechnol. 2001;19(7):661–7.
28. Schilling O, Huesgen PF, Barré O, Auf dem Keller U, Overall CM. Characterization of the prime and non-prime active site specificities of proteases by proteome-derived peptide libraries and tandem mass spectrometry. Nat Protoc. 2011;6(1):111–20.
29. Wang C, Ye M, Bian Y, Liu F, Cheng K, Dong M, Dong J, Zou H. Determination of CK2 specificity and substrates by proteome-derived peptide libraries. J Proteome Res. 2013;12(8):3813–21.
30. Tucher J, Linke D, Koudelka T, Cassidy L, Tredup C, Wichert R, Pietrzik C, Becker-Pauly C, Tholey A. LC-MS based cleavage site profiling of the proteases ADAM10 and ADAM17 using proteome-derived peptide libraries. J Proteome Res. 2014;13(4):2205–14.
31. Fuchs JE, von Grafenstein S, Huber RG, Margreiter MA, Spitzer GM, Wallnoefer HG, Liedl KR. Cleavage entropy as quantitative measure of protease specificity. PLoS Comput Biol. 2013;9(4):e1003007.
32. Julien O, Zhuang M, Wiita AP, O'Donoghue AJ, Knudsen GM, Craik CS, Wells JA. Quantitative MS-based enzymology of caspases reveals distinct protein substrate specificities, hierarchies, and cellular roles. Proc Natl Acad Sci U S A. 2016;113(14):E2001–10.
33. Schauperl M, Fuchs JE, Waldner BJ, Huber RG, Kramer C, Liedl KR. Characterizing protease specificity: how many substrates do we need? PLoS One. 2015;10(11): e0142658.
34. Liu J, Duan X, Sun J, Yin Y, Li G, Wang L, Liu B. Bi-factor analysis based on noise-reduction (BIFANR): a new algorithm for detecting coevolving amino acid sites in proteins. PLoS One. 2013;8(11):e79764.
35. Fuchs JE, von Grafenstein S, Huber RG, Kramer C, Liedl KR. Substrate-driven mapping of the degradome by comparison of sequence logos. PLoS Comput Biol. 2013;9(11):e1003353.
36. Zhang Z, Schwartz S, Wagner L, Miller WA. greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1–2):203–14.
37. Shannon CEA. mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423.
38. Fisher RA. On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. J R Stat Soc. 1922;85(1):87–94.
39. Miller RG. Simultaneous statistical inference. 2nd ed. New York: Springer; 1981.
40. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.
41. Oliveira V, Campos M, Melo RL, Ferro ES, Camargo AC, Juliano MA, Juliano L. Substrate specificity characterization of recombinant metallo oligo-peptidases thimet oligopeptidase and neurolysin. Biochemistry. 2001;40(14):4417–25.
42. Demon D, Van Damme P, Vanden Berghe T, Deceuninck A, Van Durme J, Verspurten J, Helsens K, Impens F, Wejda M, Schymkowitz J, Rousseau F, Madder A, Vandekerckhove J, Declercq W, Gevaert K, Vandenabeele P. Proteome-wide substrate analysis indicates substrate exclusion as a mechanism to generate caspase-7 versus caspase-3 specificity. Mol Cell Proteomics. 2009;8(12):2700–14.
43. Bader O, Krauke Y, Hube B. Processing of predicted substrates of fungal Kex2 proteinases from Candida albicans, C. glabrata, Saccharomyces cerevisiae and Pichia pastoris. BMC Microbiol. 2008;8:116.

44. Remacle AG, Shiryaev SA, ES O, Cieplak P, Srinivasan A, Wei G, Liddington RC, Ratnikov BI, Parent A, Desjardins R, Day R, Smith JW, Lebl M, Strongin AY. Substrate cleavage analysis of furin and related proprotein convertases, A comparative study. J Biol Chem. 2008;283(30):20897–906.
45. Page MJ, Di Cera E. Serine peptidases: classification, structure and function. Cell Mol Life Sci. 2008;65(7–8):1220–36.