Mathematical Biosciences and Engineering

MBE, 18(1): 837–850.
DOI: 10.3934/mbe.2021044
Received: 10 October 2020
Accepted: 21 December 2020
Published: 25 December 2020

http://www.aimspress.com/journal/MBE

Research article

Insights into protease sequence similarities by comparing substrate sequences and phylogenetic dynamics

Enfeng Qi[1,2,*], Can Fu[2], Ying Zhai[1] and Jianghui Dong[2,*]

[1] School of Mathematics and Statistics, Guangxi Normal University, Guilin 541000, China
[2] College of Biotechnology, Guilin Medical University, Guilin 541004, China

* **Correspondence:** Email: djh1028@126.com or ef521@126.com; Tel: +867733680230; Fax: +867735891498.

Abstract: Based on substrate sequences, we proposed a novel method for comparing sequence similarities among 68 proteases compiled from the MEROPS online database. The rank vector was defined based on the frequencies of amino acids at each site of the substrate, aiming to eliminate the different order variances of magnitude between proteases. Without any assumption on homology, a protease specificity tree is constructed with a striking clustering of proteases from different evolutionary origins and catalytic types. Compared with other methods, almost all the homologous proteases are clustered in small branches in our phylogenetic tree, and the proteases belonging to the same catalytic type are also clustered together, which may reflect the genetic relationship among the proteases. Meanwhile, certain proteases clustered together may play a similar role in key pathways categorized using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Consequently, this method can provide new insights into the shared similarities among proteases. This may inspire the design and development of targeted drugs that can specifically regulate protease activity.

Keywords: protease phylogeny; substrate sequences; phylogenetic tree; KEGG, MEROPS; homology

1. Introduction

Proteases play important roles in “life and death” processes, acting as a biological switches to activate or deactivate hydrolysis reactions on the peptide bonds of their substrates [1,2], that in turn mediate important biological functions, including cell proliferation [3], migration [4–6],

differentiation [7], and apoptosis [8], among others. Protease activity is commonly associated with a myriad of standard, yet complex physiological functions, that when dysregulated can result in abnormal pathophysiology leading to a number of diseases [9–12]. A considerable number of known protease inhibitors have been evaluated as potential targets against disease, including dipeptidyl-peptidases, which have been studied in type 2 diabetes, rein, and angiotensin-converting enzyme [13,14] for regulating blood pressure, among others.

Profiling the specificity of cleavage sites in substrates is an important step in characterizing the biochemical properties of proteases. Uncovering the substrate specificity of these proteases is central to understanding the important role of proteases in biological processes, including metabolic activity, as well as in determining actionable targets for the design of specific inhibitory agents. In particular, the specificity is determined by interactions between the substrate and its proteases. Peptide residues of substrates around the cleavage bond are indexed towards the N-terminus as P_1, P_2, P_3, P_4, …, P_n, and residues towards the C-terminus are indexed as P_1', P_2', P_3' P_4', …, P_n'. In addition, the binding pockets in the protease are named S_n–S_n' according to indices of the residues occupying the substrates [15].

Notably, with the application of high-throughput techniques in proteomics, an increasing number of cleavage sites have been identified for protease substrates [16–19] and several curated datasets have been established for detailed annotation of proteases, substrates, and pathways, including the MEROPS online database for peptidases [20], as well as CutDB [21], PMAP [22], the Degradome database [23], and CASBAH [24] for caspases. Besides the experimental methods used, multiple studies have focused on the qualitative interpretations that are visualized using sequence logo [25], weblogo [26], icelogo [27], and heat maps [28]. Some methods of quantitative analysis have been designed to measure the specificity of several proteases [29–32]. Additionally, a number of methods have been developed for predicting substrate cleavage sites for several proteases [33,34].

In this study, we aimed to develop a method that could be used to compare the similarities among 68 proteases, in order to build a phylogenetic tree for protease specificity based on substrate sequences, which may better highlight the genetic relationships among proteases.

2. Materials and methods

2.1. Extraction and processing of protease-substrate data

The protease dataset generated consisted of a total of 68 endopeptidases, covering four types of enzyme catalytic machinery that included metallo proteases, as well as aspartic, cysteine, and serine threonine proteases (Table S1 and Additional file 1). All the known substrate sequences were downloaded from the MEROPS online database (https://www.ebi.ac.uk/merops/; Wellcome Sanger Institute, Cambridgeshire, UK). The proteases were selected primarily by at least 100 annotated substrate sequences, with the signal peptidase complex removed from the dataset.

2.2. Algorithm steps

This method was improved from Fuchs et al. [30] as follows:

Step 1: First, substrate sequences of less than two amino acids over the whole binding site were removed. Next, in order to avoid a false positive, we deleted sequences that had less than two amino acids. The remainder sequences with more than two amino acids were pair-wisely aligned and the

redundant sequences were removed by a greedy algorithm [35], displaying a similarity equal to or greater than 0.875.

Step 2: For each protease, a matrix of substrate sequences spreading from S_4–S_4' was generated, containing entries with the frequencies of the 20 natural amino acids. A vector with a 20 dimension at each position was extracted, where each element corresponded to the frequencies of the amino acids. Then, from the non-prime terminus to the prime terminus, eight vectors were obtained, respectively, and denoted as $p_i = (n_{Ala}, n_{Arg}, \ldots, n_{Val})$, $(i = 1, \ldots, 8)$.

Step 3: In order to compare whole binding regions, a vector p with a 160-dimension resulted by combining the vector pi at each position for each protease, that is $p = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$.

Step 4: Next, we sorted all the 160 elements in vector p by descending order. If there were multiple elements having the same frequencies, we ranked them by alphabetical order and then adjusted the ranks by averaging them with the same values. Therefore, a rank vector r was obtained for each protease.

Step 5: The similarities between proteases were calculated using Eq 1 [36], where X and Y represent the different proteases and are respectively a 160-dimensional rank vector composed of 20 components at 8 sites, xi and yi are respectively the component of the rank vector X and Y, in which the amino acid information at each site can be reflected. In addition, n is 160, which is the dimension of the rank vector. The symmetric similarity matrix, S, was generated by a complete comparison of all proteases in the dataset.

$$S_{X,Y} = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2 - 1)} \tag{1}$$

Step 6: A distance matrix D was generated by subtracting all the elements in similarity matrix S from Eq 1. The elements in the distance matrix were regarded as the differences of the pairwise proteases and the entry of matrix D was defined as follows:

$$d_{ij} = 1 - S_{ij} \tag{2}$$

2.3. Phylogenetic tree and heatmap construction

Molecular evolutionary genetics analysis (MEGA) 7.0.14 software (Institute of Molecular Evolutionary Genetics; University Park, PA) and the web-based tool Interactive Tree of Life (iTOL; http://itol.embl.de/) [37] were used to construct a phylogenetic tree. This phylogenetic tree enabled the visualization of the differences between proteases based on substrate sequences. Heml 1.0 (http://hemi.biocuckoo.org/down.php) was used to construct heatmaps and cluster heatmaps.

3. Results

3.1. Substrate amino acid frequency of eight binding sites

The substrate amino acid frequencies of the proprotein convertase subtilisin/kexin type 9 (PCSK) family of serine proteases at eight binding sites are shown in Figure 1. Among these frequencies, it

can be observed that amino acid changes at each site of the proteases belonging to this family are very similar. Especially, the P_4 and P_1 site in all the proteases of the PCSK family show specificity for arginine (Arg), whereas the amino acids appearing at the P_2 sire are mainly lysine (Lys) and arginine (Arg) [38]. Although, there are more amino acid species at each site of the prime terminal, the trend of amino acid changes is still consistent, indicating that this family of serine proteases shows very similar specificity in substrate recognition.

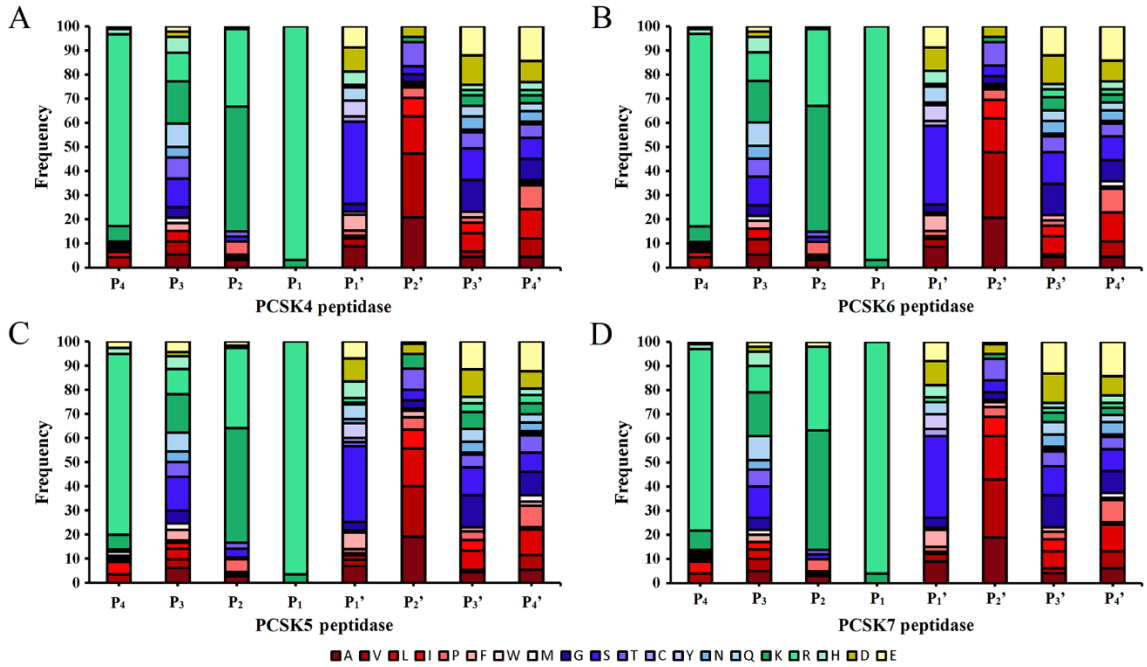

Figure 1. Amino acid frequency maps of serine proteases over their substrate sites.

Analysis of the amino acid frequency map of certain metalloproteases showed that four proteases in the metalloprotease family presented a large number of amino acids at each site (Figure 2). Although, we found that there was no preference on the specific amino acids needed at all of the eight sites, the trend underlying amino acid changes among these proteases appeared exactly similar.

As observed in the amino acid frequency map of certain cysteine proteases (Figure 3), proteases in the caspase family exhibited specificity for aspartic acid (Asp) at P_1 site [39]. In particular, both caspase 3 and caspase 7, demonstrated a very high frequency of Asp at the P_4 site [39], and both proteases appeared involved in several pathways depicted in the KEGG database, especially both proteases were found to play a significant role in signaling pathways mediating apoptosis. Importantly, these two proteases were found to be associated with debilitating diseases, such as Alzheimer's disease, non-alcoholic fatty liver disease, pertussis (i.e., whooping cough), among others. It can also be observed in the amino acid frequency map that the proportion of polar-neutral amino acids at the P_1' site of caspase 1, caspase 3, and caspase 7, has a close proximity (Figure 3).

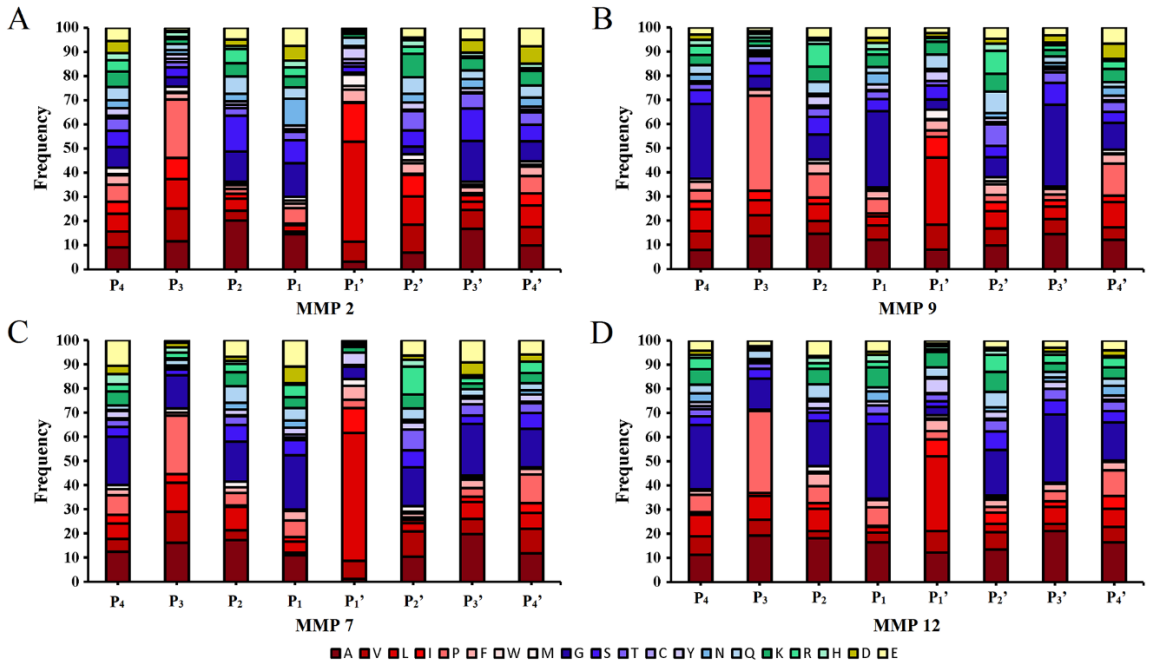

Figure 2. Amino acid frequency maps of metalloproteases over their substrate sites.

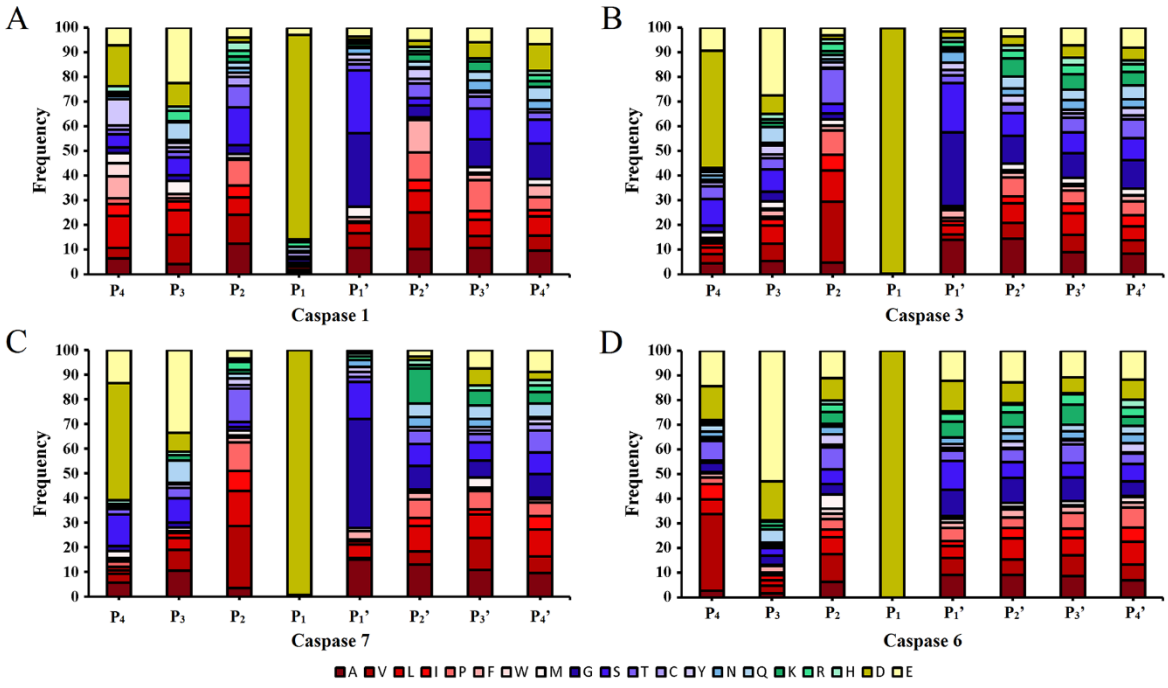

Figure 3. Amino acid frequency maps of cysteine proteases over their substrate sites.

In addition, as it can be seen in the amino acid frequencies of four aspartic proteases (Figure 4), there are no specially preferred amino acids at almost every site, and only cathepsin E and cathepsin D, present more leucine (Leu) and phenylalanine (Phe) residues at the P_1 site [40]. Although, there is no

preference of specific amino acids at other sites, the changes in the amino acid trend among these proteases, is highly consistent. Meanwhile, both proteases can activate precursors of biologically active proteins in the prelysosomal compartment.

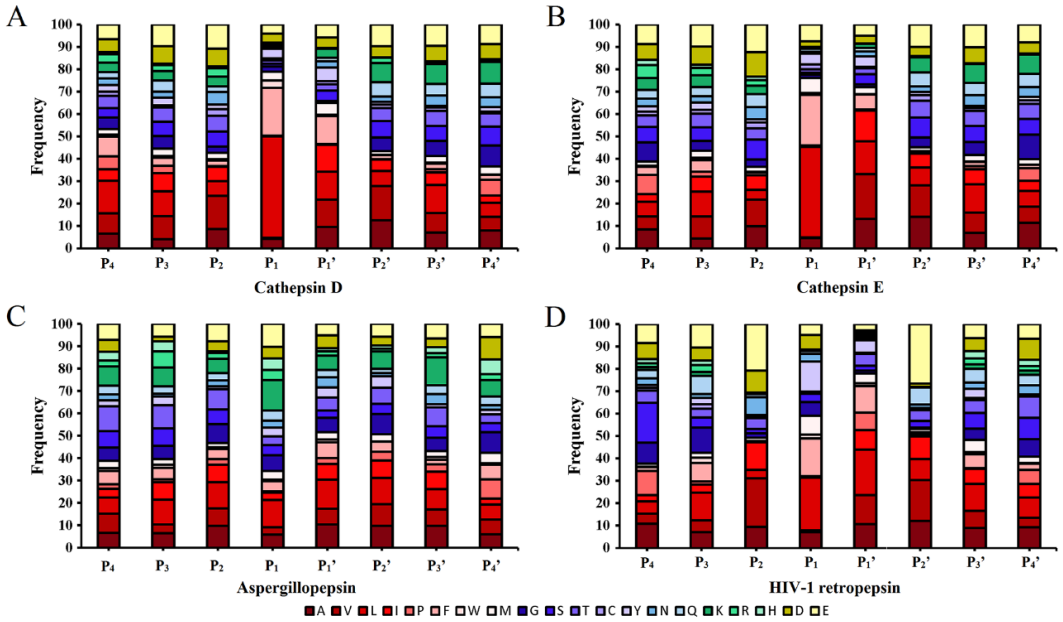

Figure 4. Amino acid frequency maps of aspartic proteases over their substrate sites.

3.2. The distance between proteases

Using our quantitative method, the distance values obtained were according to the similarities between the different proteases in our dataset. Figure 5 is an example of the distance values found for the four proteases analyzed, which were calculated from the protease substrates found in the MEROPS database covering the eight sites flanking the cleavage bond.

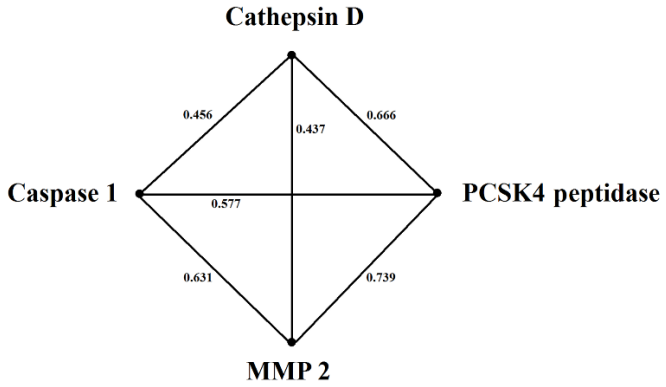

Figure 5. Exemplary distance values depicting four proteases. The dots in the graph represent proteases and the values above the edge, connecting the two dots, represent the distance value between the corresponding two proteases.

The distances between the proteases are shown in the Additional file 1. Regarding serine proteases, those belonging to the PCSK family of proteases (i.e., PCSK2, PCSK4, PCSK6, PCSK5, and PCSK7) behave very similarly when recognizing substrates at each binding site, and therefore the distance value between them is minimal, of not more than 0.05. Several proteases in the caspase family, also show very similar preference for substrates. Consequently, the distance values found between them were also minimal, especially for caspase 3 and caspase 7, as they are very similar in how they bind to amino acids at multiple sites, and therefore have the lowest distance value 0.078 in the caspase family. Moreover, analyses on the matrix metalloprotease (MMP) family showed some similarities among these proteases in recognizing broken substrates, therefore the distance value found between them was relatively low, and the minimum distance between MMP9 and MMP12 was 0.114. Aspartic proteases have almost no specifically recognized amino acids at all the eight sites evaluated, therefore the distance value between the two proteases in the family was larger, when compared to those found between proteases in other catalytic types.

3.3. The phylogenetic tree and heatmap of proteases

The evolutionary relationships between these proteases are reflected in the phylogenetic tree generated using the distance matrix. The phylogenetic tree generated based on the substrate sequences P_4–P_4' of the substrates analyzed. Figure 6 showed that almost all homologous proteases were clustered in small branches, with the proteases belonging to the same catalytic type also clustered together.

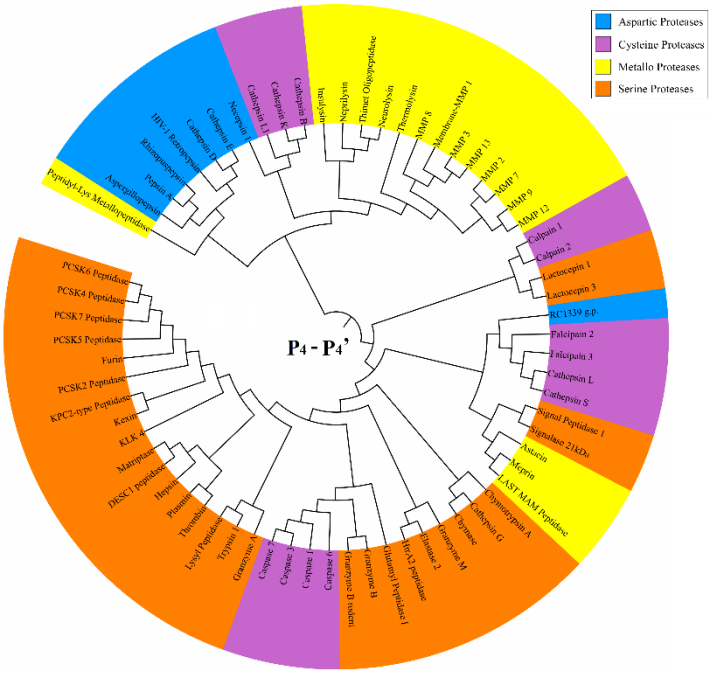

Figure 6. Schematic depicting the phylogenetic tree of proteases over the P_4–P_4' substrate sequences. Four colors distinguish the analyzed proteases according to their catalytic type: serine proteases are colored in orange, aspartic proteases in blue, cysteine proteases in purple, and metalloproteases in yellow.

Among them, the serine proteases included in our dataset were mainly concentrated in a large branch of the phylogenetic tree, containing the PCSK and S08 subfamily of proteases, such as kexin and furin, among others. Although, signal peptidase 1 and signalase 21kDa were not clustered together with other serine proteases on the concentrated phylogenic tree branch, another two proteases were found clustered on a small branch, mainly because these two proteases are signal peptidases that cleave remnant signal peptides located on the plasma membrane. In addition, while lactocepin 1 and lactocepin 3 were not clustered in the major branches of the serine protease in the phylogenetic tree, the two homologous proteases were indeed found clustered on a small branch. Further protease analysis based on the KEGG pathway database, linked to the MEROPS database, showed that both cathepsin G and chymase proteases, clustered in the same branch, were involved in regulating the renin-angiotensin system. Both proteases are responsible for the initiation of the chain reaction in the system, turning it on in order to maintain cardiovascular development and blood pressure regulation. In addition, the serine proteases thrombin and plasmin were also found clustered together, and according to KEGG pathway database, both can activate complement proteins to produce the corresponding enzymatic activity in the complement immune system, triggering a cascade of proteolytic reactions.

Analysis on MMPs showed that most of these proteases concentrate on a more concentrated branch. Among them, eight proteases belonging to the MMP family in the dataset were grouped in this branch. Interestingly, thimet oligopeptidase and neurolysin operate in a similar manner while processing the degradation of the substrate, mainly by cleaving the oligopeptide. Therefore, two proteases are clustered on a small branch. In addition, the peptidyl-Lys metallopeptidase showed a unique preference for Lys at the P_1' site. However, other MMPs analyzed in our dataset, did not have such specific preferences. Therefore, it was not clustered in the concentrated branch of MMPs.

In the protease dataset, the cysteine proteases were scattered across several small branches. Among them, four proteases belonging to the caspase family, including caspase 1, caspase 3, caspase 6, and caspase 7, were found clustered together in a minor branch. It is obvious from the phylogenetic tree generated that the homologous cathepsin L1, cathepsin B, and cathepsin K clustered on a small branch. Meanwhile, calpain 1 and calpain 2 were also found clustered on a small branch, and importantly, both proteases were found to be key regulators of apoptotic signals and Alzheimer's disease pathogenesis as depicted in the KEGG pathway database.

Regarding aspartic proteases, only two of them, necepsin 1 and RC1339g.p., were found scattered in other branches. The remaining six aspartic proteases anlyzed in our dataset were all derived from the same branch and clustered together. Although, there are no assumptions on the amino acid sequence information of the proteases during the constructing process of the phylogenetic tree, these proteases were still clustered by substrate sequence specificities.

To verify reliability of the results of the phylogenetic tree, we constructed a heatmap and performed clustering analysis (Figure 7). Most of the branches in the cluster were almost consistent with the results of the phylogenetic tree. For example, the clustering results of the 14 proteases in the largest branch of serine proteases completely coincided with the results observed in the phylogenetic tree, thereby proving the reliability of these results.

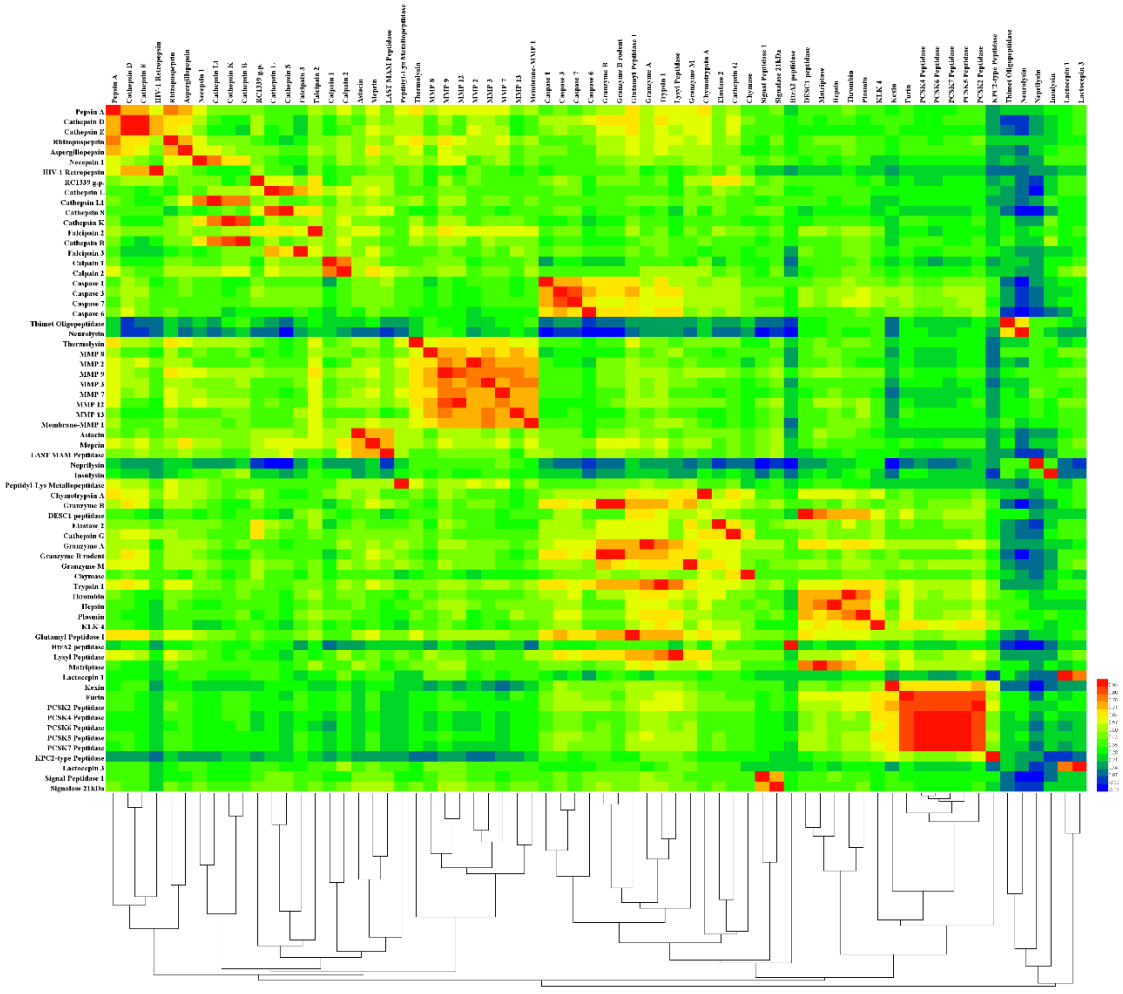

Figure 7. Heatmap and clustering analysis of proteases over the P_4–P_4' substrate sequences. Color bar from red to green represents the order of similarity from high to low.

In addition, in order to create a phylogenetic tree of proteases based on substrate sequences P_4–P_4', we applied a similar approach to the single site P_1, aimed at generating an evolutionary tree of the protease datasets at this site (Figure 8). For this phylogenetic tree, the protease dataset was roughly divided into three branches. The first major branch was mainly composed of serine proteases together with the PCSK family, kexin, and furin, among others. The amino acids found at the P_1 site in the substrate sequences of these proteases analyzed were mainly Arg residues. The homologous protease cathepsin L1, cathepsin B, and cathepsin K were still found clustered on a small branch of the phylogenetic tree. The second branch was mainly composed of the MMP family of proteases and the caspase family of cysteine proteases. Among them, granzyme B and granzyme rodent, were found still clustered with the caspase family in the phylogenetic tree. Importantly, these proteases are known to degrade substrate sequences whose amino acid at the P_1 site consists mainly of Asp residues. The homologous signal peptidase signal peptidase 1 and signalase 21kDa, were found clustered in a minor branch of the phylogenetic tree. The third branch was found dominated by aspartic proteases. Moreover, the homologous proteases, such as lactopepin 3 and lactopepin 1, were

also clustered in a minor branch of the phylogenetic tree. By studying the pathways generated from the KEGG database, both insulysin and neprilysin were found to be associated with Alzheimer's disease pathogenesis. For instance, insulysin has been shown to reduce the levels of Alzheimer's disease-associated proteins and reduce the occurrence of Alzheimer's disease, while neprilysin is known to be of great significance in the treatment and prevention of Alzheimer's disease [41].

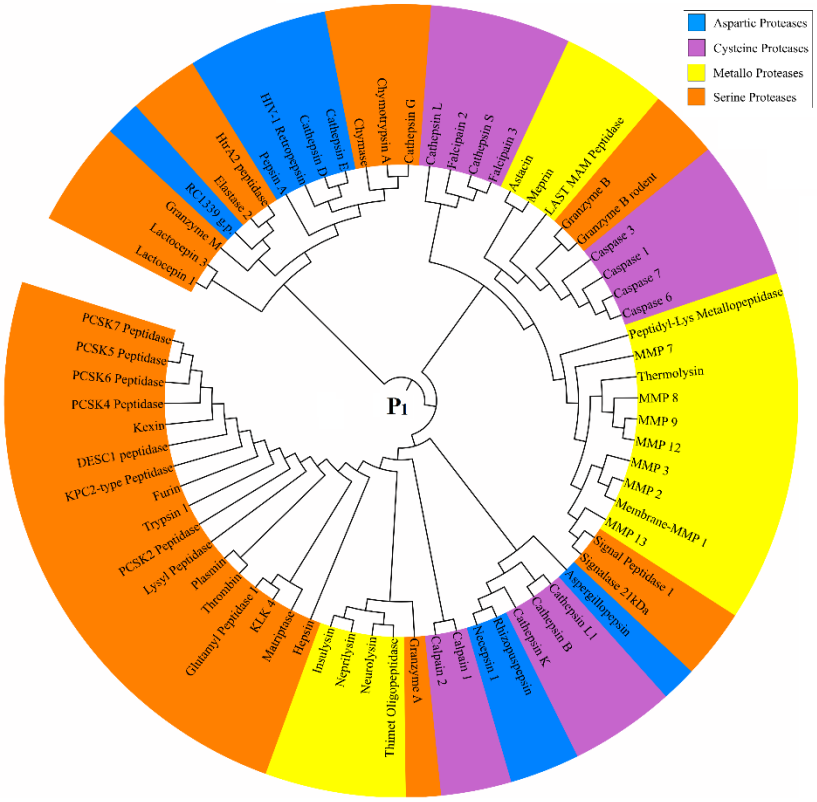

Figure 8. Schematic depicting the phylogenetic tree of proteases over the P_1 substrate sequences. Proteases are colored according to their catalytic type: serine proteases are colored in orange, aspartic proteases in blue, cysteine proteases in purple, and metalloproteases in yellow.

4. Discussions

In this work, we presented a new method to capture the similarities between specific proteases based on substrate sequence analysis. Our method does not rely on amino acid sequences or the enzymatic structure of proteases, and is not restricted to homologous proteases and catalytic types. Therefore, our proposed method is easier to operate and improves clustering accuracy when evaluating the evolutionary tree of proteases. In short, it can be utilized to analyze the similarities between more extensive protease data, which can provide new insights into the study of protease function similarities.

First, the method used in this study is computationally easier to operate and reflect the evolutionary tree of proteases. By visualizing the amino acid frequencies of the substrates at eight

different binding sites, the similarities of each are directly probed. The calculation we used to determine protease similarities was based on substrate vectors containing amino acid frequencies at each sub pocket of their binding sites. Moreover, the distance matrix was constructed according to the similarity matrix. Consequently, evolutionary trees were constructed to form comparisons between the different proteases. When compared to the calculation used to determine cleavage entropy [29], our method is computationally easier to operate. In addition, when compared to the visualization method based on a single cleavage site [25], our method of calculating similarities between the proteases can better reflect their evolutionary relationship and conserved functions.

Second, the phylogenetic tree we constructed was clustered in a more efficient manner. For instance, almost all of the four types of catalytic proteases in this tree were found clustered together according to their respective catalytic type. Among them, metalloproteases, as well as serine and aspartic proteases, were found to have formed a certain scale of the three kinds of proteases, while some homologous cysteine proteases were found clustered in a small branch, despite lack of concentrated branches in the phylogenetic tree. According to Fuchs et al. [30], all the serine proteases in their phylogenic tree, were found divided into several small branches, clustered together with other catalytic types of proteases. In contrast, most of the proteases belonging to the catalytic type of serine in Figure 6, were found concentrated on a larger branch. Although, some serine proteases did not seem to be concentrated on this branch, homologous proteases, such as lactocepin 3 and lactopepin 1, signal peptidase 1, and signalase 21kDa, were found clustered together in a small branch.

Lastly, our method appears more effective in determining these similarities. Particularly, there are various proteases that show specificity at the P_1 site. Similar analyses were performed and our phylogenetic tree of proteases was finally constructed based on the single binding site, S_1 Figure 8, which clusters better than the phylogenic tree established by Fuchs et al. [30], Interestingly, both proteases belonging to metalloprotease and aspartic protease families, displayed few branches. As the proteases in the MMP family display several degrees of similarity at the P_1 site Figure 2, the distance value determined between the MMP family was relatively low and most of the metalloprotease family of proteins were found clustered together Figure 8. Regarding aspartic proteases, HIV-1 retropepsin and pepsin A, displayed more Leu and Phe residues at the P_1 site, showing more similarities and observed clustered together in our phylogenic tree. Interestingly, the fact that some proteases clustered together were also found to play similar roles in a number of biological pathways curated in the KEGG database [42], as well as the similarities captured by the substrates, has an immense potential to guide the development of therapeutic agents that can be specifically designed to target key proteases.

5. Conclusions

In conclusion, we expect that the similarities found between substrate sequences and their respective protease, can promote the recognition of protease ligands, which are already widely used in polypharmacology predictions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62061012),

Research Fundamental Capacity Improvement Project for Middle Age and Youth Teachers of Guangxi Universities (2019KY0078 and 2019KY0517), Guangxi Science and Technology Base and Talent Special Project (2018AD19238), and Guangxi Youth Science Foundation Project (2019JJB110017).

Conflict of Interest

The authors have no competing interests to declare.

References

1. N. D. Rawlings, F. R. Morton, C. Y. Kok, J. Kong, A. J. Barrett, MEROPS: The peptidase database, *Nucleic Acids Res.*, **36** (2008), 320–325.
2. B. Turk, Targeting proteases: Successes, failures and future prospects, *Nat. Rev. Drug Discovery*, **5** (2006), 785–799.
3. M. Egeblad, Z. Werb, New functions for the matrix metalloproteinases in cancer progression, *Nat. Rev. Cancer*, **2** (2002), 161–174.
4. K. Nabeshima, T. Inoue, Y. Shimao, T. Sameshima, Matrix metalloproteinases in tumor invasion: Role for cell migration, *Pathol. Int.*, **52** (2002), 255–264.
5. A. C. Newby, Matrix metalloproteinases regulate migration, proliferation, and death of vascular smooth muscle cells by degrading matrix and non-matrix substrates, *Cardiovascul. Res.*, **69** (2006), 614–624.
6. R. Palmisano, Y. Itoh, Analysis of MMP-dependent cell migration and invasion, *Methods Molecul. Biol.*, **622** (2010), 379–392.
7. A. Page-McCaw, A. J. Ewald, Z. Werb, Matrix metalloproteinases and the regulation of tissue remodelling, *Nat. Rev. Molecul. Cell Biol.*, **8** (2007), 221–233.
8. O. Julien, J. A. Wells, Caspases and their substrates, *Cell Death Diff.*, **24** (2017), 1380–1389.
9. X. L. Li, P. Wang, Y. Xie, Protease nexin-1 protects against Alzheimer's disease by regulating the sonic hedgehog signaling pathway, *Int. J. Neurosci.*, (2020), 1–10.
10. M. A. Slack, S. M. Gordon, Protease activity in vascular disease, *Arterioscler. Thromb. Vascul. Biol.*, **39** (2019), 210–218.
11. C. Tomuschat, A. M. O'Donnell, D. Coyle, P. Puri, Increased protease activated receptors in the colon of patients with Hirschsprung's disease, *J. Pediatr. Surg.*, **55** (2020), 1488–1494.
12. L. J. Visser, G. N. Medina, H. H. Rabouw, R. J. de Groot, M. A. Langereis, T. de Los Santos, et al., Foot-and-mouth disease virus leader protease cleaves G3BP1 and G3BP2 and inhibits stress granule formation, *J. Virol.*, **93** (2019), 922–918.
13. K. Ożegowska, J. Bartkowiak-Wieczorek, A. Bogacz, A. Seremak-Mrozikiewicz, A. J. Duleba, L. Pawelczyk, Relationship between adipocytokines and angiotensin converting enzyme gene insertion/deletion polymorphism in lean women with and without polycystic ovary syndrome, *Gynecol. Endocrinology .: Off. J. Int. Soc. Gynecol. Endocrinol.*, **36** (2020), 496–500.
14. X. S. Ren, Y. Tong, Y. Qiu, C. Ye, N. Wu, X.Q. Xiong, et al., MiR155-5p in adventitial fibroblasts-derived extracellular vesicles inhibits vascular smooth muscle cell proliferation via suppressing angiotensin-converting enzyme expression, *J. Extracell. Vesicles*, **9** (2020), 1698795.

15. I. Schechter, A. Berger, On the size of the active site in proteases. I. Papain. 1967, *Biochem. Biophys. Res. Commun.*, **425** (2012), 497–502.
16. P. Van Damme, A. Staes, S. Bronsoms, K. Helsens, N. Colaert, E. Timmerman, et al., Complementary positional proteomics for screening substrates of endo and exoproteases, *Nat. Methods*, **7** (2010), 512–515.
17. O. Schilling, O. Barré, P. F. Huesgen, C. M. Overall, Proteome-wide analysis of protein carboxy termini: C terminomics, *Nat. Methods*, **7** (2010), 508–511.
18. P. Van Damme, S. Maurer-Stroh, K. Plasman, J. Van Durme, N. Colaert, E. Timmerman, et al., Analysis of protein processing by N-terminal proteomics reveals novel species-specific substrate determinants of granzyme B orthologs, *Mol. Cell. Proteomics : MCP*, **8** (2009), 258–272.
19. S. Mahrus, J. C. Trinidad, D. T. Barkan, A. Sali, A. L. Burlingame, J. A. Wells, Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini, *Cell*, **134** (2008), 866–876.
20. N. D. Rawlings, A. J. Barrett, R. Finn, Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors, *Nucleic Acids Res.*, **44** (2016), 343–350.
21. Y. Igarashi, A. Eroshkin, S. Gramatikova, K. Gramatikoff, Y. Zhang, J. W. Smith, et al., CutDB: A proteolytic event database, *Nucleic Acids Res.*, **35** (2007), 546–549.
22. Y. Igarashi, E. Heureux, K. S. Doctor, P. Talwar, S. Gramatikova, K. Gramatikoff, et al., PMAP: Databases for analyzing proteolytic events and pathways, *Nucleic Acids Res.*, **37** (2009), 611–618.
23. V. Quesada, G. R. Ordóñez, L. M. Sánchez, X. S. Puente, C. López-Otín, The Degradome database: Mammalian proteases and diseases of proteolysis, *Nucleic Acids Res.*, **37** (2009), 239–243.
24. A. U. Lüthi, S. J. Martin, The CASBAH: A searchable database of caspase substrates, *Cell Death Differ.*, **14** (2007), 641–650.
25. K. K. Dey, D. Y. Xie, M. Stephens, A new sequence logo plot to highlight enrichment and depletion, *Bmc Bioinf.*, **19** (2018), 1–9.
26. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator, *Genome Res.*, **14** (2004), 1188–1190.
27. N. Colaert, K. Helsens, L. Martens, J. L. Vandekerckhove, K. Gevaert, Improved visualization of protein consensus sequences by iceLogo, *Nat. Methods*, **6** (2009), 786–787.
28. M. M. Dix, G.M. Simon, B. F. Cravatt, Global mapping of the topography and magnitude of proteolytic events in apoptosis, *Cell*, **134** (2008), 679–691.
29. J. E. Fuchs, S. von Grafenstein, R. G. Huber, M. A. Margreiter, G. M. Spitzer, H. G. Wallnoefer, et al., Cleavage entropy as quantitative measure of protease specificity, *PLoS Comput. Biol.*, **9** (2013), 1003007.
30. J. E. Fuchs, S. von Grafenstein, R. G. Huber, C. Kramer, K. R. Liedl, Substrate-driven mapping of the degradome by comparison of sequence logos, *PLoS Comput. Biol.*, **9** (2013), 1003353.
31. E. Qi, D. Wang, Y. Li, G. Li, Z. Su, Revealing favorable and unfavorable residues in cooperative positions in protease cleavage sites, *Biochem. Biophys. Res. Commun.*, **519** (2019), 714–720.
32. E. F. Qi, D. Y. Wang, B. Gao, Y. Li, G. J. Li, Block-based characterization of protease specificity from substrate sequence profile, *Bmc Bioinf.*, **18** (2017), 438.
33. J. Song, H. Tan, A. J. Perry, T. Akutsu, G. I. Webb, J. C. Whisstock, et al., PROSPER: An integrated feature-based tool for predicting protease substrate cleavage sites, *PloS one*, **7** (2012), 50300.

34. J. Verspurten, K. Gevaert, W. Declercq, P. Vandenabeele, SitePredicting the cleavage of proteinase substrates, *Trends Biochem. Sci.*, **34** (2009), 319–323.
35. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, A greedy algorithm for aligning DNA sequences, *J. Comput. Biol.: J. Comput. Mol. Cell Biol.*, **7** (2000), 203–214.
36. C. Spearman, The proof and measurement of association between two things, *Am. J. Psychol.*, **100** (1987), 441–471.
37. I. Letunic, P. Bork, Interactive Tree Of Life v2: Online annotation and display of phylogenetic trees made easy, *Nucleic Acids Res.*, **39** (2011), 475–478.
38. N. M. Ng, R. N. Pike, S. E. Boyd, Subsite cooperativity in protease specificity, *Biol. Chem.*, **390** (2009), 401–407.
39. H. R. Stennicke, M. RENATUS, M. MELDAL, G. S. SALVESEN, Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1, 3, 6, 7 and 8, *Biochem. J.*, **350** (2000), 563–568.
40. Y. Choe, F. Leonetti, D. C. Greenbaum, F. Lecaille, M. Bogyo, D. Brömme, et al., Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities, *J. Biol. Chem.*, **281** (2006), 12824–12832.
41. S. Elamouri, H. Zhu, J. Yu, R. A. Marr, I. M. Verma, M. S. Kindy, Neprilysin: An enzyme candidate to slow the progression of Alzheimer's disease, *Am. J. Pathol.*, **172** (2008), 1342–1354.
42. M. Eguiluz, F. Kulcheski, R. Margis, F. Guzman, De novo assembly of vriesea carinata leaf transcriptome to identify candidate cysteine-proteases, *Gene*, **691** (2019), 96–105.

Supplementary: Additional file 1. Distance.xlsx file containing distance values between proteases.

Table S1. The indices of 68 proteases in MEROPS database.