Health Biotechnology and Biopharma (2019), 3(3): 50-58

Original Research Article

Importance of the catalytic site neighbouring residues on the functionality of the enzymes

Masoumeh Alinaghi [1], Samira Beyramysoltan [1], Soroush Sardari [1,*]

[1]Drug Design and Bioinformatics Unit, Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran

***Corresponding author:**. Soroush Sardari, Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran. Email: ssardari@hotmail.com*

DOI: 10.22034/HBB.2019.22
Received: November 10, 2019; **Accepted:** December 15, 2019

ABSTRACT

Catalytic residues are highly conserved among the proteins, amino acid residues which are sequentially neighbor to the catalytic sites, some extent contain similar structural information. Collected data from the Protein Data Bank (PDB) database contain one aspartic protease group (n=28) and two serine protease groups (n=37 and n=7) specified by catalytic triads. Analysis of the molecular structures of amino acids by Sparse Linear Discriminant Analysis (SLDA) and t-test revealed that the neighbor residues contain information on the functional differentiation of the enzymes, suggesting that the sequentially close residues to the catalytic triads are more preserved than the rest of amino acid residues. Moreover, similar structural properties in the vicinity of all three catalytic sites can be observed by correlation heat maps. These results can hopefully facilitate a better understanding of the protein function from structure and lead to novel protein design.

***Keywords*:** Catalytic residues, active sites, sequence conservation, enzyme function

INTRODUCTION

Enzymes, which are required by every biological process, can function as a proper catalyst even if other parts of the enzyme are mutated [1]. According to the literature, the catalytic residues are highly preserved in comparison with other amino acid residues [2]. Accordingly, due to their greater degree of conservation, protein catalytic residues are expected to be distinct from the rest of the

protein sequence which is not involved in the interaction. However, the extent of validity of this argument can be a question [3,4]. In the literature, active site residues are mostly used for investigation of proteins' functionality [5-8], while amino acid residues in the vicinity of active sites (neighbor amino acid residues) might also be of importance. Accordingly, we hypothesized that even though the proteins' hot spots and catalytic residues are well preserved in the revolution, but the neighbor residues cannot be apart from this evolution and they might still contain related information. In this letter, it is aimed to investigate if the neighbor residues contain preserved information on the functionality of proteins and to which degree and distance, they might be important. Thus, linear combination of the whole molecular structure of amino acids on shape, size, and symmetry and atom distribution was employed to investigate the information content of the neighbour amino acids about function-based differentiation of the enzymes.

MATERIALS AND METHODS

Data collection

An aspartic protease group and two serine protease groups (serine protease 1 and 2 with different catalytic triad in each group), containing 28, 37 and 7 enzymes respectively, were extracted from Protein Data Bank (PDB) [9-10], while the catalytic information obtained from the Peptidase Database MEROPS (release 11.0) [11] (Table S1). All selected enzymes have three catalytic sites, CS1, CS2 and CS3 by CS1 being close to N-terminus and consequently four amino acid sequences (AAS1: the amino acid sequence before CS1, AAS2: the amino acid sequence between CS1 and CS2, AAS3: the amino acid sequence between CS2 and CS3, AAS4: the amino acid sequence after CS3). Enzymes within each group have the same catalytic triads; aspartic proteases contain the catalytic triad of aspartate, tyrosine and aspartate, and serine proteases 1 have the catalytic triad of histidine, aspartate and serine, while catalytic triad of serine proteases 2 is serine, lysine and tyrosine.

Data analysis

Matlab (version R2016b, MathWorks Inc., United States) was used for data analysis and visualization. For evaluating the discriminative property of amino acids, the first descriptor of VSW [12] was used. VSW is derived from the principal component analysis (PCA) of a matrix of 99 weighted holistic invariant molecular indices of amino acids containing the whole molecular structure on shape, size, symmetry and atom distribution. Therefore, every amino acid is specified by the first descriptor of VSW,

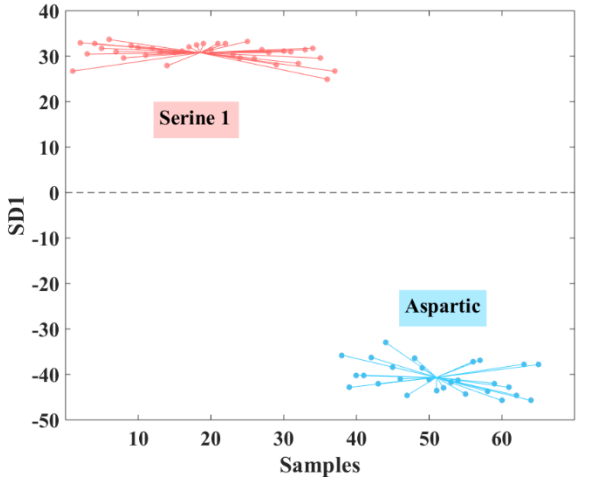
which is the linear combination of the structural properties of amino acids. Each enzyme was described by a 1×n vector which n is the whole number of the amino acid residues in the enzyme. However, the catalytic triad was excluded from the analysis to specifically investigate the effect of the neighbor amino acid residues. In order to equalize the length of all enzyme vectors, number of amino acids in an enzyme with minimum length was considered. For the amino acids with a longer length, the farthest amino acids from the catalytic triad were removed from the analysis; i.e. the amino acids close to N-terminus and C-terminus in AAS1 and AAS4 respectively, as well as the middle amino acids in AAS2 and AAS3. The serine 2 enzymes have two amino acid residues in AAS2 and therefore, AAS2 was removed from related analysis. The variation in the amino acid profile of enzymes was explored by PCA analysis on mean-centered and pareto-scaled data. For a supervised classification, sparse linear discriminant analysis (SLDA) [13-14] was applied on data to provide the discrimination pattern of the enzymes as well as selecting the discriminative features (amino acids), while the sparseness constraint was applied to simultaneously perform variable selection. The data was normalized prior to analysis and SLDA was performed between every two groups of amino acids; between aspartic and serine 1 (Data 1), aspartic and serine 2 (Data 2), and serine 1 and serine 2 (Data 3). Moreover, a two-sample t-test analysis was employed to evaluate the significance level of differences in structural properties of amino acid in each position. The h=1 indicates the rejection of the null hypothesis with 5 % significance level, h=0 otherwise indicates a failure to reject the null hypothesis. Moreover, correlations between the structural properties of amino acid residues in different groups were calculated and visualized by heat map plots. For this visualization, Pearson's correlation coefficients and *p*-values for the correlation between variables were evaluated.

RESULTS

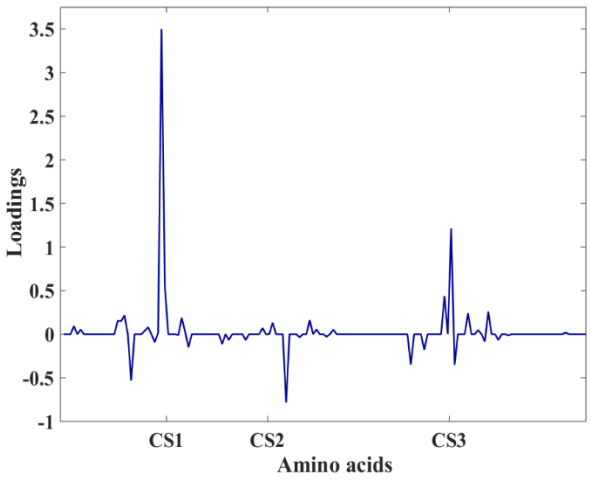
In order to investigate the importance of amino acids for discrimination of the enzymes with different catalytic triad, molecular structural descriptor of amino acids in each position was investigated. A PCA score plot reveals a strong separation between enzyme groups while the aspartic and serine groups are clustered along the first principal component (PC1) and the serine 1 and 2 are grouped along the PC2. (Fig. S1). The first two PCs explained 9% of the total variability in the data. Furthermore, the SLDA results evaluated the classification of

each two groups of the enzymes (Fig. 1). Elucidation of the loading plots showed that the clear separation of the enzyme groups on the first SLDA discriminate vectors (SD1) of each pair-wise analysis could be ascribed mainly by a higher weight for neighbor residues and zero loading for residues located far from the catalytic triad. In Fig. 1-c,f,i, the t-test results for each position in the sequence are presented with a filled circle if it is significantly different between the enzyme groups (h=1, p<0.05). Calculated pair-wise t-test for every position of amino acid residue in the sequence revealed significant differences for neighbor residues. Nevertheless, inspecting farther amino acids from catalytic triad, relatively fewer numbers of significant positions can be observed.

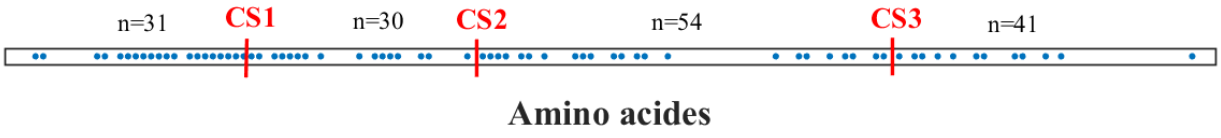
(a)



(b)

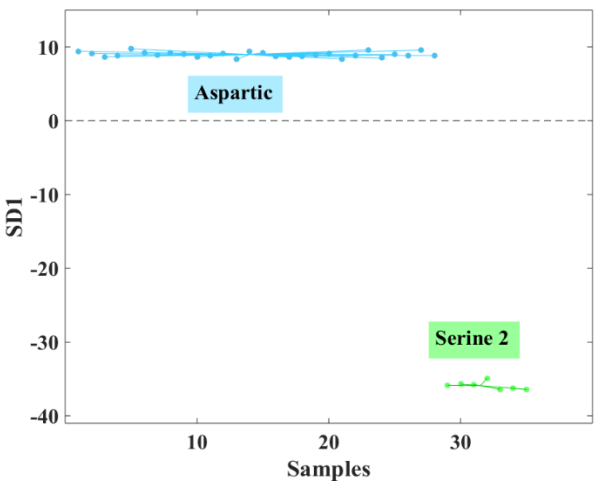


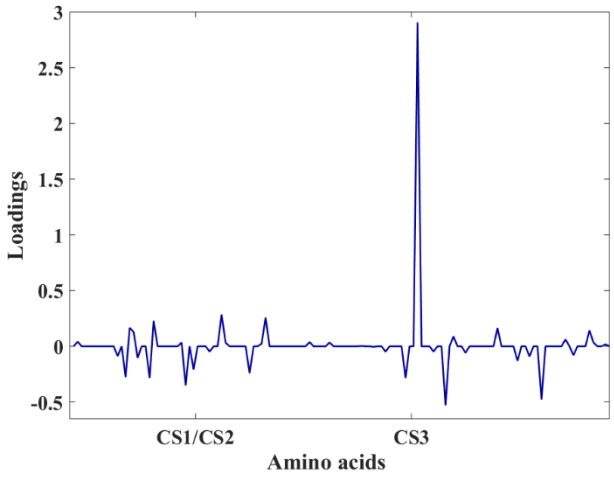
(c)



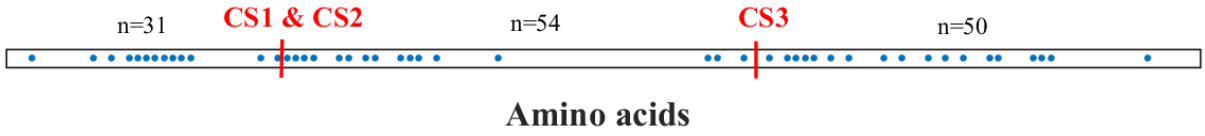
Amino acides

(d) (e)



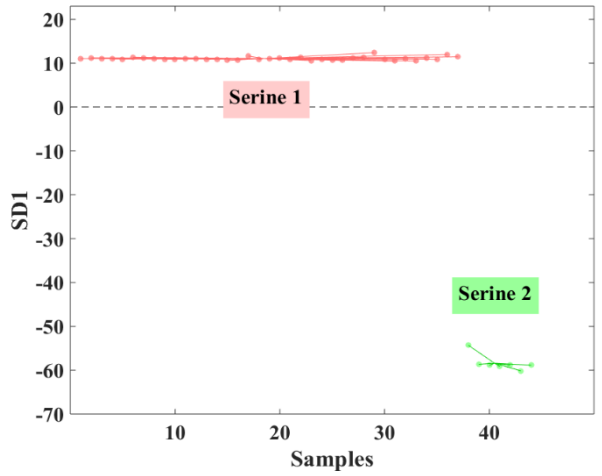


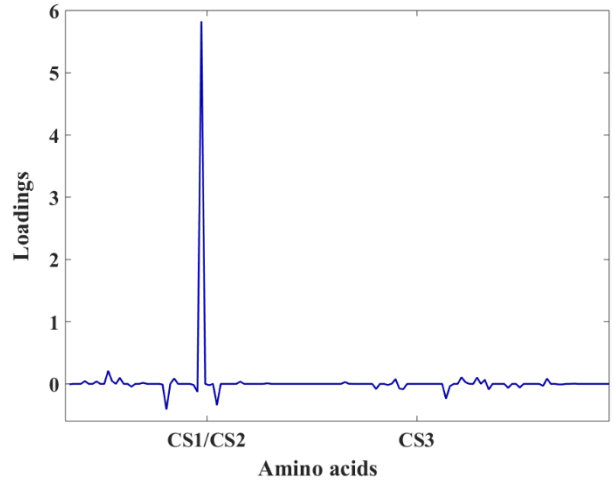
(f)



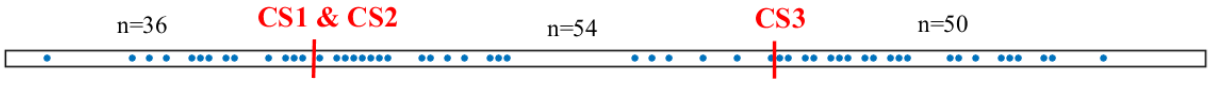
Amino acids

(g) (h)





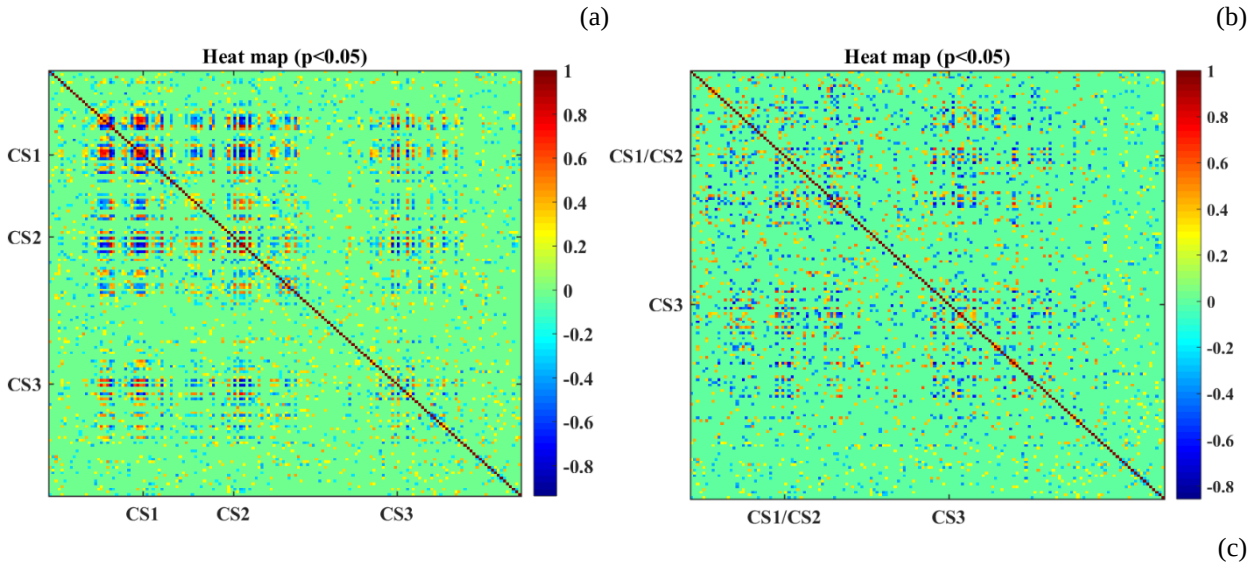
(i)



Amino acids

Figure 1. The visualization plot of SLDA and T-test results. (a), (d) and (g) represents the SLDA discriminate vectors for the data 1, 2 and 3 respectively. (b), (e) and (h) represents the SLDA loadings for data 1, 2 and 3 respectively. (c), (f) and (i) illustrate the pair-wise t-test results of amino acid residues in each sequence position for data 1, 2 and 3 respectively, while significantly different results (h=1, p<0.05) are specified by filled circles. Data 1 contain the structural properties of amino acid residues of aspartic protease and serine protease 1, data 2 contains the structural properties of amino acid residues of aspartic protease and serine protease 2 and data 3 contains the structural properties of amino acid residues of serine protease 1 and 2. CS1, CS2 and CS3 are the catalytic triad in each enzyme group: aspartic proteases contain the catalytic triad of aspartate (CS1), tyrosine (CS2) and aspartate (CS3), and serine proteases 1 have the catalytic triad of histidine (CS1), aspartate (CS2) and serine (CS3), while catalytic triad of serine proteases 2 is serine (CS1), lysine (CS2) and tyrosine (CS3).

To further explore the impact of neighbor residues and identify the relationship among them, correlation analysis was carried out on the data, resulted in some patterns of correlating residues in the heat map (Fig. 2). The revealed pattern is in agreement with previous results while illustrating the correlation between the different catalytic triad's neighbors.



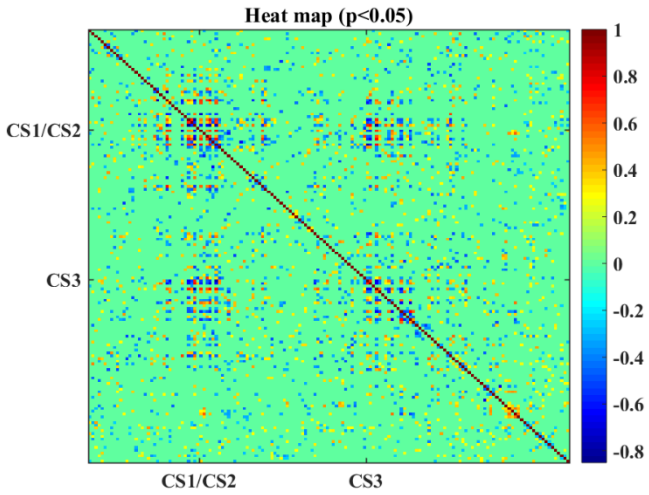


Figure 2. The correlation heat map of amino acid residues (a) first data (S1A) (b) second data (S2A) and (c) third data (SS). The correlation coefficient values are displayed as a color-coded map. The rows and columns of the heat map indicate the neighbor residues of the catalytic triads, i.e. CS1, CS2 and CS3. Data 1 contains the structural properties of amino acid residues in aspartic protease and serine protease 1, data 2 contains the structural properties of amino acid residues in aspartic protease and serine protease 2 and data 3 contains the structural properties of amino acid residues in serine protease 1 and 2.

DISCUSSION

In the present study, we investigated the prominence of neighbor amino acid residues which are not directly involved in the enzymatic reactions. The PCA grouping revealed that the amino acid sequence contains information about the functionality of the enzyme and catalytic triad (Fig. S1), the SLDA and t-test analysis illustrated that the neighbor residues are more responsible for the separation of the enzymes with different catalytic triads (Fig 1). The residues positioned far from catalytic triad have less information about the functional discrimination of the enzymes.

Identifying the site of enzymatic reaction in proteins is a key for deciphering its functional mechanisms. According to the literature, the catalytic triads are the most preserved spots in the protein during the evolution and mutation; therefore, the enzymes' functionality are mostly described by their catalytic triads. On the other hand, analysis of the whole amino acids residues reveals that there is still some information on the whole sequence even though there are less preserved. Interestingly, our analysis illustrates that

the sequentially close residues to the catalytic triad can hold higher weight on differentiating of the enzymes. The conservation of the residues gradually falls as the distance from the catalytic residues enhances. This observation suggests that the neighbor residues might be structurally closer to the catalytic triads. The heat map pattern of amino acid residues also presents the correlating information in the vicinity of the catalytic residues (Fig 2), not only the neighbors of the same catalytic position but also between different sites, suggesting the similar environment of the all catalytic residues within an enzyme group.

CONCLUSION

Furthermore, regardless of the enzyme group, the results suggest that the preserved neighborhood of each catalytic site can be defined by a limited number of amino acid residues, nearly around 20 amino acids. An approximate border for neighbor importance cannot be defined for pair-wise comparison of the serine 1 and aspartic in AAS2 as there are totally 30 residues between CS1 and CS2. However, a robust statement on the number of substantial neighbor needs further investigation of different enzyme groups to be approved and generalized. In conclusion, the purpose of this letter was to emphasize on the neighbor amino acid residues by displaying their importance on the differentiation of the enzymes. Further investigation can be implemented by evaluating effect of proteins' length on the number of preserved neighbors as well as investigating its tertiary structures.

SUPPLEMENTARY MATERIAL

Supplementary Data Table S_1, Figure S_1, S_2

REFERENCES

[**1].** Ringe, D.; Petsko, G. A., How enzymes work. *SCIENCE-NEW YORK THEN WASHINGTON*- 2008, *320* (5882), 1428.

[2]. Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M., Analysis of catalytic residues in enzyme active sites. *J Mol Biol*, 2002, *324* (1): 105-21.

[3]. Szilágyi, A.; Grimm, V.; Arakaki, A. K.; Skolnick, J., Prediction of physical protein–protein interactions. *Phys Biol*, 2005, *2* (2), S1.

[4]. Caffrey, D. R.; Somaroo, S.; Hughes, J. D.; Mintseris, J.; Huang, E. S., Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 2004, *13* (1): 190-202.

[5]. Ofran, Y.; Rost, B., Analysing six types of protein–protein interfaces. *J Mol Biol*, 2003, *325* (2): 377-87.

[6]. Ofran, Y.; Rost, B., Predicted protein–protein interaction sites from local sequence information. *FEBS lett*, 2003, *544* (1-3): 236-39.

[7]. Zhou, H. X.; Shan, Y., Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Struc, Func Bioinf*, 2001, *44* (3): 336-43.

[8]. Chung, J. L.; Wang, W.; Bourne, P. E., Exploiting sequence and structure homologs to identify protein–protein binding sites. *Proteins: Struc, Func Bioinf*, 2006, *62* (3): 630-40.

[9]. Bank, P. D., HM Berman, J. Westbrook, Z. Feng, G. Gilliland, TN Bhat, H. Weissig, IN Shindyalov, PE Bourne. *Nucleic Acids Res* 2000, *28*: 235.

[10]. Berman, H.; Henrick, K.; Nakamura, H., Announcing the worldwide protein data bank. *Nat Struc Mol Biol*, 2003, *10* (12): 980-80.

[11]. Rawlings, N. D.; Waller, M.; Barrett, A. J.; Bateman, A., MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids Res* 2014, *42* (D1): 503-509.

[12]. Tong, J.; Liu, S.; Zhou, P.; Wu, B.; Li, Z., A novel descriptor of amino acids and its application in peptide QSAR. *J Theo Biol* 2008, *253* (1): 90-97.

[13]. Clemmensen, L.; Hastie, T.; Witten, D.; Ersbøll, B., Sparse discriminant analysis. *Technometrics* 2011, *53* (4): 406-13.

[14]. Sjöstrand, K.; Clemmensen, L. H.; Larsen, R.; Einarsson, G.; Ersbøll, B. K., Spasm: A matlab toolbox for sparse statistical modeling. *J Stat Softw* 2018, *84* (10).

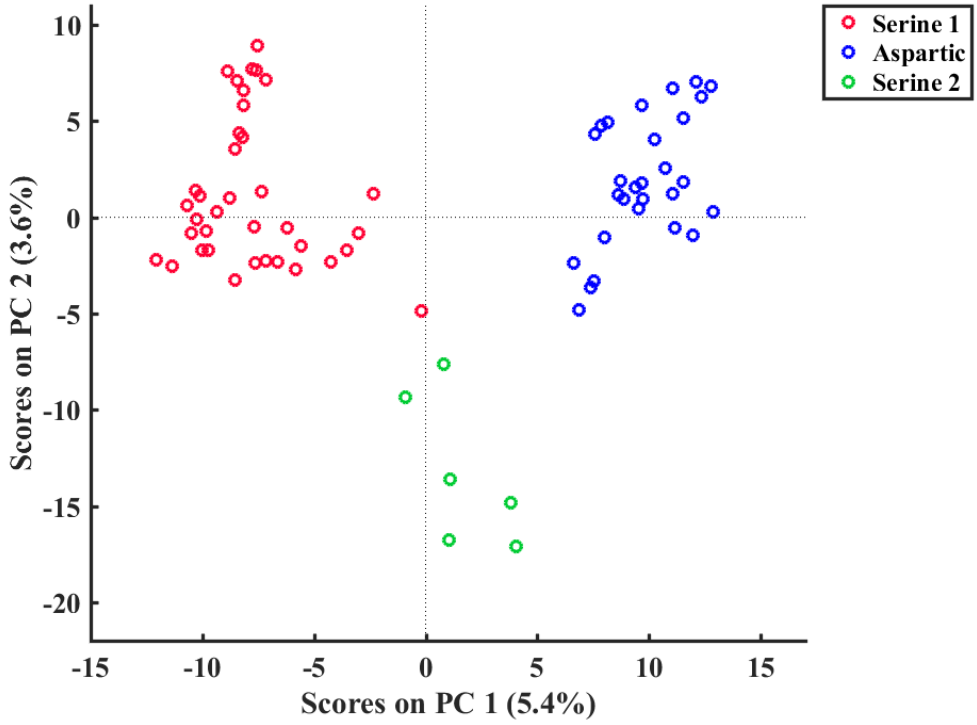


Figure. S1. PCA score plot of the data containing the physico-chemical properties of amino acid sequences of aspartic protease, serine protease 1 and 2.

Table S1. List of enzymes

	EC number	Molecule	Organism	Catalytic triad
Aspartic protease	EC 3.4.23.1	Pepsin 3a	*Homo sapiens*	Asp, Tyr, Asp
	EC 3.4.23.3	Gastricsin	*Homo sapiens*	Asp, Tyr, Asp
	EC 3.4.23.46	Memapsin 2 (beta-secretase)	*Homo sapiens*	Asp, Tyr, Asp
	EC 3.4.23.4	Prochymosin a/b precursor	*Bos taurus*	Asp, Tyr, Asp
	EC 3.4.23.15	Renin	*Homo sapiens*	Asp, Tyr, Asp
	EC 3.4.23.15	Renin	*Mus musculus*	Asp, Tyr, Asp
	EC 3.4.23.20	Penicillopepsin	*Penicillium janthinellum*	Asp, Tyr, Asp
	EC 3.4.23.21	Rhizopuspepsin	*Rhizopus microsporus*	Asp, Tyr, Asp
	EC 3.4.23.23	Pepsin	*Rhizomucor pusillus*	Asp, Tyr, Asp
	EC 3.4.23.24	Candidapepsin-1	*Candida albicans*	Asp, Tyr, Asp
	EC 3.4.23.18	Aspergillopepsin	*Aspergillus saitoi*	Asp, Tyr, Asp

	EC 3.4.23.22	Endothiapepsin	*Cryphonectria parasitica*	Asp, Tyr, Asp
	EC 3.4.23.25	Proteinase a	*Saccharomyces cerevisiae*	Asp, Tyr, Asp
	EC 3.4.23.29	Polyporopepsin	*Polyporus tulipiferae*	Asp, Tyr, Asp
	EC 3.4.23.40	Prophytepsin	*Hordeum vulgare*	Asp, Tyr, Asp
	EC 3.4.23.38 and 39	Plasmepsin	*Plasmodium vivax*	Asp, Tyr, Asp
	EC 3.4.23.38	Plasmepsin-1	*Plasmodium falciparum*	Asp, Tyr, Asp
	EC 3.4.23.39	Plasmepsin ii	*Plasmodium falciparum*	Asp, Tyr, Asp
	EC 3.4.23.18	Aspartic protease	*Trichoderma reesei*	Asp, Tyr, Asp
	EC 3.4.23.24	Aspartic proteinase	*Candida tropicalis*	Asp, Tyr, Asp
	EC 3.4.23.24	Sapp1p-secreted aspartic protease 1	*Candida parapsilosis*	Asp, Tyr, Asp
	EC 3.4.23.45	Beta-secretase 2	*Homo sapiens*	Asp, Tyr, Asp
	EC 3.4.23.-	Pepsin	*Gadus morhua*	Asp, Tyr, Asp

	EC 3.4.23.-	Plasmepsin IV	*Plasmodium falciparum*	Asp, Tyr, Asp
	EC 3.4.23.24	Aspartic proteinase (sap2 gene product)	*Candida albicans*	Asp, Tyr, Asp
	EC 3.4.23.24	Candidapepsin-3	*Candida albicans*	Asp, Tyr, Asp
	EC 3.4.23.24	Candidapepsin-5	*Candida albicans*	Asp, Tyr, Asp
	EC 3.4.23.18	Aspartic proteinase	*Aspergillus oryzae*	Asp, Tyr, Asp
Serine protease 1	EC 3.4.21.-	Kallikrein-5	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.79	Granzyme B	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.-	Transmembrane protease, serine 11E	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.-	Venom serine proteinase	*Deinagkistrodon acutus*	His, Asp, Ser
	EC 3.4.21.4	Trypsin	*Pontastacus leptodactylus*	His, Asp, Ser
	EC 3.4.21.79	Rat granzyme b	*Rattus norvegicus*	His, Asp, Ser
	EC 3.4.21.4	Trypsin	*Streptomyces griseus*	His, Asp, Ser

	EC 3.4.21.49	Hypoderma lineatum Collagenase	*Hypoderma lineatum*	His, Asp, Ser
	EC 3.4.21.4	Trypsin	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.37	Human LEUCOCYTE elastase	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.20	Cathepsin G	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.76	PR3 (myeloblastin)	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.78	Granzyme A	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.-	Granzyme C	*Mus musculus*	His, Asp, Ser
	EC 3.4.21.-	Granzyme M	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.-	Rat mast cell protease ii	*Rattus norvegicus*	His, Asp, Ser
	EC 3.4.21.59	Tryptase alpha-1	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21	Pro-granzyme K	*Homo sapiens*	His, Asp, Ser
	EC	Chymase 2	*Mesocricetus auratus*	His, Asp, Ser
	EC 3.4.21.4	Trypsin	*Bos taurus*	His, Asp, Ser
	EC 3.4.17.1	Procarboxypeptidase a	*Bos taurus*	His, Asp, Ser
	EC 3.4.21.-	Prostasin	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.77	Prostate specific antigen	*Homo sapiens*	His, Asp, Ser

	EC	Nerve growth factor	*Mus musculus*	His, Asp, Ser
	EC 3.4.21.35	Kallikrein-1e2	*Equus caballus*	His, Asp, Ser
	EC 3.4.21.119	Glandular kallikrein-13	*Mus musculus*	His, Asp, Ser
	EC 3.4.21.-	Plasminogen activator	*Trimeresurus stejnegeri*	His, Asp, Ser
	EC 3.4.21.46	Complement factor d	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.41	C1r complement serine protease	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.42	Ccomplement c1s component	*Homo sapiens*	His, Asp, Ser
	EC	Complement C2	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.47	Complement factor b	*Homo sapiens*	His, Asp, Ser
	EC 3.4.21.-	Pro-phenoloxidase activating enzyme-I	*Holotrichia diomphalia*	His, Asp, Ser
	EC 3.4.21.22	Factor ixa	*Sus scrofa*	His, Asp, Ser
	EC 3.4.21.35	Nerve growth factor	*Mus musculus*	His, Asp, Ser
	EC	Chymotrypsin-like serine protease	*Equine arteritis virus*	His, Asp, Ser
	EC	Serine protease	*Sesbania mosaic virus*	His, Asp, Ser

Serine protease 2	EC 3.4.16.4	D-alanyl-d-alanine carboxypeptidase transpeptidase	*Streptomyces lividans*	Ser, Lys, Tyr
	EC 3.4.11.19	D-aminopeptidase	*Ochrobactrum anthropi*	Ser, Lys, Tyr
	EC	Alkaline D-peptidase	*Bacillus cereus*	Ser, Lys, Tyr
	EC 3.5.2.6	Beta-lactamase	*Enterobacter cloacae*	Ser, Lys, Tyr
	EC	Pbp related beta-lactamase	*Pyrococcus abyssi*	Ser, Lys, Tyr
	EC	D-amino acid amidase	*Ochrobactrum anthropi*	Ser, Lys, Tyr
	EC	Pbp related beta-lactamase	*Pyrococcus abyssi*	Ser, Lys, Tyr

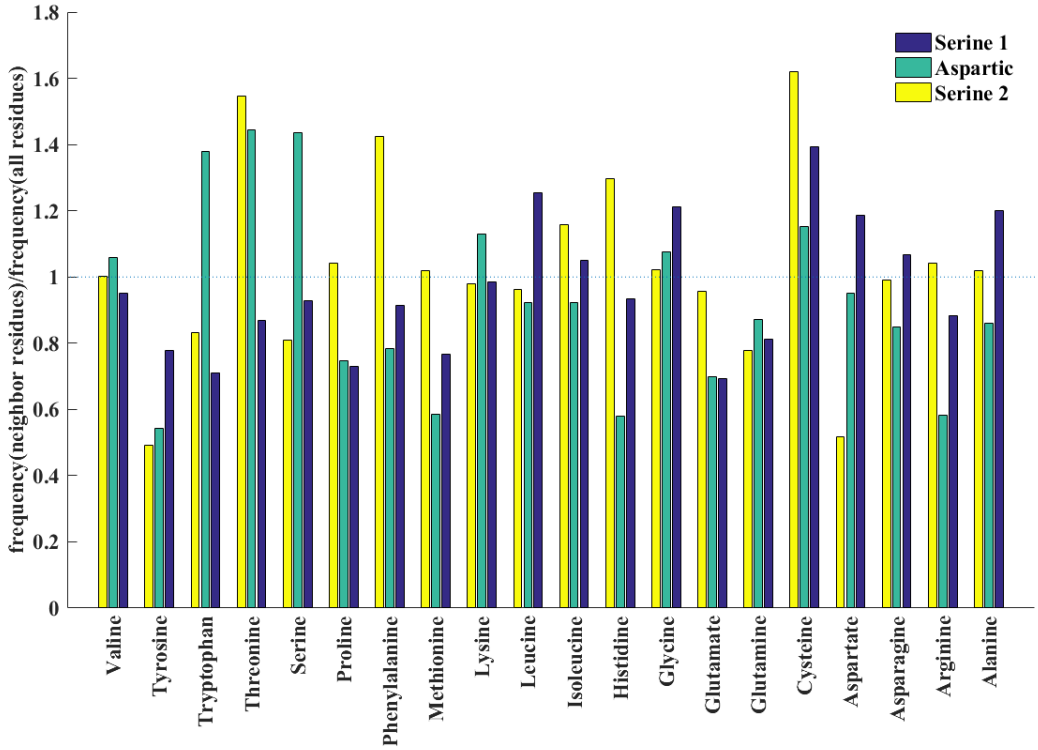


Figure. S2. Frequency plot of the data of amino acid sequences of aspartic protease, serine protease 1 and 2.