國立交通大學

生物資訊研究所

碩 士 論 文

搜尋蛋白質中相關聯殘基所構成的網絡

Search the Network of Coupling Residues in Proteins

研 究 生：簡思樸

指導教授：黃鎮剛　教授

中 華 民 國 九 十 五 年 七 月

搜尋蛋白質中相關聯殘基所構成的網絡

Search the Network of Coupling Residues in Proteins

研 究 生：簡思樸　　Student：Szu-Pu Chien

指導教授：黃鎮剛　　Advisor：Jenn-Kang Hwang

國 立 交 通 大 學

生 物 資 訊 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Bioinformatics

College of Biological Science and Technology

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Bioinformatics

July 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年七月

搜尋蛋白質中相關聯殘基所構成的網絡

學生：簡思樸　　指導教授：黃鎮剛

國立交通大學生物資訊研究所碩士班

中 文 摘 要

同源蛋白的多重序列排比帶有演化紀錄，從這些紀錄中可以分析出許多有用的資訊。這類分析中最常見的即是保守性分析；一個排比的位置如果總是只能觀察到一種胺基酸殘基，我們就稱它是保守性十分強的位置。保守性強的位置通常是蛋白質中行使功能的區域，這是因為功能的行使會限制住該區域可能產生的突變。

然而，有些位置雖然保守性不強，卻仍然是蛋白質行使功能的重要區域。胺基酸之間的關聯性解釋了這種現象。許多研究者表示，要了解蛋白質之中胺基酸間的關聯性，可從對兩個位置作關聯突變分析得知。

在這份研究中，我們利用保守性熵和關聯性熵來預測相關聯的殘基並進行比較。我們的研究初步證實經校正的關聯性熵是有效的預測關聯突變的指數。我們同時也提出一種新的空間搜尋方法來找出關聯殘基所組成的網絡和它的意義。

Search the Network of Coupling Residues in Proteins

Student：Szu-Pu Chien　　Advisor：Jenn-Kang Hwang

Institute of Bioinformatics
National Chiao Tung University

Abstract

Multiple sequence alignments (MSA) carry evolutionary records. By analyzing these records, one may get useful biological information. A common analysis method is conservation analysis. If there is no record in mutation at one MSA position, we say that position tends to be conserved. Conserved positions are always functional important sites since functionality imposes constraints on these sites from mutation.

Although some positions show little conservation, they are still functionally important. The coupling residues account this observation. Many investigators have proposed that analyzing correlated mutation in two positions can effectively predict the coupling degree between these two residues.

Thus, in this study, we introduce conservation entropy and coupling entropy to detect coupling between two sites and also compare the effect between two. Our study generally confirmed the effect of normalized coupling entropy to be a good indicator in detecting correlated mutation. We also introduce a network searching strategy to verify the identity of coupling residues network

致　　謝

終於到了這最後寫致謝的時刻~ 說真的，從很早以前我就一直在想論文的致謝要寫什麼內容，只是沒想到真的到了要寫的此時此刻，竟一時間不知該如何寫起。

首先，我真的很慶幸來到交大，來到生資所，來到黃鎮剛老師的實驗室。這一連串偶然與偶然的組合，現在回想起來更覺得不可思議。在所有的偶然中，能成為黃鎮剛老師實驗室研究生的這個偶然是最難能可貴的。在我所有要感謝的人當中我最感謝的就是老師。雖然我經常惹老師生氣，但這兩年來他仍然是不厭其煩地指導我這做事總是慢半拍的學生；從老師身上我學到了許多作研究該有的態度，和一位科學家應該抱有的精神。我想這些東西是在論文上面無法看到、但是會一輩子永遠記得的寶貴學習經驗。

再來，當然要感謝朝夕相處的實驗室夥伴，沒有你們，我的碩士生活大概就要是黑白的了。景盛總是定時來關心我的進度，很和藹可親的提出一些想法和建議；施董大概是學長輩裡我最好的朋友，我不會忘記那些冒著大雨一起去買宵夜的日子，以及那些很白痴的搞笑連結；玉菁總是會說很多八卦的笑話；草霸、水蛭、蔚倫也不時給我許多鼓勵，草霸跟水蛭更是實驗室裡的開心果。鐵熊跟梁涵堃學長平時雖不常見到，但梁學長常常會分享他的研究心得，而鐵熊則是十分搞笑。少偉跟勇欣雖然我總是沒什麼機會深交，但偶爾的虛寒問暖還是令人感到很窩心。當然更要感激跟我並肩作戰的同學們：它口一直是我分享零食的好鄰居、書瑋則是解決問題的好幫手、盧慧則是常常傳一些充滿文化氣息的消息給我。最後，要感謝學弟在最後的一段時間慷慨地提供住所給我，身為學長好像也沒能怎麼幫助你們，就祝你們將來也口試順利吧! 也特別感激施董和草霸在最後關頭指導我們的口試演說。

還有太多太多的感謝可以說的；其他生資所的老師們、隔壁實驗室的同學們…。雖然好像沒有直接關係，但若沒有與你們有這些每一日相處和指教的一點一滴，我的碩士生活是不可能有什麼成果的~

再一次感謝你們所有人!

CONTENTS

FIGURE CAPTIONS

1. Introduction

Multiple sequence alignment (MSA) of homologous proteins demonstrates evolutionary records. Changes in these records are made by the mutation of amino acids. Each position of MSA shows different degree of conservation. Conserved positions are often functional important positions, since functionality would impose constraints on mutation. The degree of conservation at each amino acid site is similar to the inverse of the site's rate of evolution; slowly evolving sites are evolutionarily conserved, while rapidly evolving sites are variable[1].

However, mutation studies have shown that many non-conserved positions may also be functionally important. How, then, do these non-conserved positions change during evolution without eliminating the activity of the protein? In the laboratory experiment, the detrimental effects of one mutation may be suppressed by a compensating mutation at another position, that is, a second-site suppressor mutation, and it is expected that during evolution the effects of mutations are often counterbalanced in a similar way. Thus, this concept forms the basis that non-conserved positions in functional important sites may have some coupling relationships between each other.

But what is the actual meaning of "coupling"? Many investigators proposed that these coupling residues, while not necessarily nearby in the primary sequence, form three –dimensional contacts at an above-random frequency[2]. Some investigators reasoned that, as previously mentioned, coupling residues may comprise functional sites and play an important role in functional interactions[3,4]. There are others suggested another explanation of this coupling relationship that since some coupling residues do not interact in close distance but in long-range interactions, they could be responsible for the energetic process such as allosteric communication in signal transduction proteins[5,6]. For example, ligand binding at an externally accessible site in G protein–coupled receptors (GPCRs) reliably triggers structural changes at distant cytoplasmic domains that mediate interaction with heterotrimeric G proteins[7,8]. It is also been proposed that coupling residues in distal functional sites are connected by a small subset of coupling residues which forms a sparse

physically connected networks that serves like a bridge in the protein tertiary structure[6].

To verify the true meaning of coupling residues, an important step is to globally map coupling relationships between amino acid residues in protein structure. A common technique for understanding coupling residues is the double mutant cycle analysis, in which the energetic independence of two residue positions is established if the sum of the free energy changes of two independent mutations is equal to the free energy change of the double mutant[9-13]. Even a modestly sized protein, however, has millions of possible double mutations that render experimental laboratory study of double mutant cycles impractical. Thus, an alternative approach is suggested by relating evolutionary information in MSA to coupling residues in protein by detection of correlated mutations. If every time a given residue in a position of MSA changes, there is a corresponding change in another position of the MSA, then the two corresponding residue positions may be coupled and under selective pressure.

Many researchers have reasoned that coupling positions in sequence alignments must be correlated mutate and thus identifiable. In 1994, Valencia *et al.*[14] proposed that analyzing correlated mutation can be used to convert sequence correlation patterns into three-dimensional structure information. Others proposed that this analysis can be used to determine functionally important residues[15]. Ranganathan and colleagues[5,6] proposed a statistical coupling analysis (SCA) algorithm and claimed it could find "pathways of energetic connectivity" that "have emerged early in the evolution of the protein folds and, much like the atomic structure, are fundamentally conserved features of the domain families[5]". Oliviera *et al.*[16] (2002) developed a qualitative method, Correlated Mutation Analysis (CMA), to identify coupling residues based on the variability, entropy and correlation of the residues. Singer *et al.*[2] (2002), in a variation of CMA using empirically derived likelihood scores, predicted residues to be in contact with a 16% accuracy for a variety of protein families. Tiller and Liu[17] (2003) and Dunn *et al.*[3,4] (2005) used mutual interdependence and mutual information to identify positions that coupled with each other, but were not coupled significantly with other positions in the alignment.

In our work, we adopted entropy calculations to measure the couplings between

positions. We applied both conservation entropy and coupling entropy to map global sequence of target proteins to verify what the coupled residues stand for, and see if the information in MSA can be used to predict the residue contact in three-dimensional structure, functional important sites, or energetic connectivity in proteins, as previously researchers proposed. Toward that end, we examined a few published double mutant cycle data sets and compared our results to previous works. In addition, we proposed a novel "space-searching" method which includes structural information that has yet been used in previous study. By using this method, we wanted to verify the identity of the network of coupling residues.

2. Methods

2.1 Amino acid composition frequency

To calculate the measure of correlated mutation, we first have to extract useful evolutionary information from MSA. This is obtained by using the observed amino acid composition frequencies, f_{x_i} , which stands for the composition frequency of amino acid *x* at position *i*, and x_i is the amino acid types on position *i*. *x* denotes 20 kinds of amino acids.

The frequency can simply be obtained by

$$f_{x_i} = \frac{n_{x_i}}{N} \tag{1}$$

Where n_{x_i} means the number of sequences with amino acid *x* at position *i; N* means the total number of aligned sequences.

2.2 Conservation entropy

With the amino acid composition frequency, the conservation entropy *H(i)* can be obtained by

$$H(i) = -\sum_{x} f_{x_i} \ln f_{x_i} \tag{2}$$

Conservation entropy is a well known measure of uncertainty for a random variable. It follows that when f_{x_i} can only take on one value, i.e., with $f_{x_i} = 1$, then *H(i)* = 0, i.e. there is

no uncertainty. When the frequency is equally distributed over all members of the amino acids, i.e., f_{x_i} = 1/20 for all kinds of *x*, then *H(i)* is maximized.

2.3 Coupling entropy

As for dealing with two positions in MSA, coupling entropy can be applied:

$$H(i,j) = \sum_{x,y} f_{x_i y_j} \ln \frac{f_{x_i y_j}}{f_{x_i} f_{y_j}} + \sum_{x,y} f_{x_i} f_{y_j} \ln \frac{f_{x_i} f_{y_j}}{f_{x_i y_j}} \quad (3)$$

Here, f_{x_i, y_j} denotes the frequency of amino acids *x* and *y* observed simultaneously in position *i* and *j*, respectively. The magnitude of *H(i, j)* can be accounted by the consistency of two functions, that is, f_{x_i} and f_{y_j} in our case. To justify the ability of *H(i, j)*, we calculate a small set of artificial testing alignment in Fig. 1.

We can see that with *H(i, j)* itself can only report little about the correlation between two columns of the alignment. By the conclusion made by Wahl and Dunn *et al.*[3,4], the performance of entropy calculation would become better after normalization. Thus, we tried to normalize *H(i, j)* as well. We normalized *H(i, j)* by divide each term with its own numerator:

$$H'(i,j) = 2\left(\frac{\sum_{x,y} f_{x_i y_j} \ln \frac{f_{x_i y_j}}{f_{x_i} f_{y_j}}}{\sum_{x,y} f_{x_i y_j} \ln f_{x_i y_j}} + \frac{\sum_{x,y} f_{x_i} f_{y_j} \ln \frac{f_{x_i} f_{y_j}}{f_{x_i y_j}}}{\sum_{x,y} f_{x_i} f_{y_j} \ln f_{x_i} f_{y_j}}\right) \quad (4)$$

The constant “2” is simply put for the purpose of making the completely correlated positions to be normalized to 1(See Fig.1. (A) and (B)). The comparison between before and after normalization is shown in Fig.1. We use *H’(i, j)* instead of *H(i, j)* throughout this work unless otherwise noted.

The higher the *H(i, j)*, the more the degree of correlated mutation between position *i* and *j*.

2.4 Network searching strategy

The searching strategy is started from an initial residue, with an input of PDB file is necessary. The overview of the searching strategy is shown in Fig.2. After the target residue

been selected, a cutoff sphere is defined around the target residue by any atom of the residues with distance with any atom of the target residue below 4Å. The 4Å criterion is chosen by empirically trials. The next step is to calculate the *H(i, j)* of all residues inside the cutoff sphere. With a given threshold, we then choose the residue with largest *H(i, j)* exceed threshold as the next target residue. The termination of this network searching is executed when the next target residue is already in the network.

2.5 Datasets

We began our calculation with two examples: (1) the third PDZ domain from the mammalian synaptic protein PSD-95 (PDB_ID: 1BE9); (2) the Staphylococcal Nuclease (PDB_ID: 1EY0). The reason why pick these proteins as our examples is that they all have been studied by previous work and thus suitable to be compared[5,18]. Also, both the PDZ domain and the Staphylococcal Nuclease have published double mutant cycle data[5,19]. The MSA file of 1BE9 and 1EY0 are downloaded from the HSSP database[20]. The number of aligned sequence is 432 in 1BE9, while 67 in 1EY0.

PDZ domains are found in many cell junction-associated proteins, mediating the clustering of membrane ion channels by binding to their C-terminus. An important function of PDZ domains was implicated when the two N-terminal PDZ domains (PDZ-1 and PDZ-2) of PSD-95 and its close relatives were shown to act as specific binding modules for the peptide motif (Thr/Ser-X-Val) found at the very C-terminus of Shaker-type K+ channels and of NR2 subunits of N-methyl-D-aspartic acid (NMDA) receptor ion channels. The interaction between the PDZ domains of PSD-95 and the C-terminus of K+ channel subunits results in the co-clustering of both proteins when they are co-expressed in heterologous cells. Furthermore, Shaker-type K+ channels and NMDA receptor ion channels co-localize with PSD-95 in the mammalian brain. In further support of a membrane protein localization function, mutations in Discs-large, the Drosophila homolog of PSD-95, cause abnormal morphology at synapses where the protein is known to occur[21].

PDZ domains are suitable dataset to do coupling analysis because they are small protein with only 115 residues. And they are also evolutionarily well represented protein family (many positions in alignment show distribution near to natural occurrence). In addition, PDZ

domains play key roles as organizing centers for multi-protein signaling complexes[5].

Fig.3. is a citation figure of MacKinnon et al.'s work in 1996[21] about the crystal structure of PDZ domain. The residues that directly participate in peptide-binding can be shown.

As for Staphylococcal Nuclease, we compared our results to a previous work of double mutant cycle analysis[19], which studied the stability effects of multiple packing mutations in the hydrophobic core. This study verified that there are 6 residues in the core region show interactions: 23, 25, 66, 72, 92, and, 99.

3. Results

1BE9 contains 115 residues. The numbering begins from 301 to 415. The plot of global mapping of conservation entropy is shown in Fig.4. The ranking of this global mapping of top conserved residues is listed in Table.1, which we can see that there is not much residues located on binding site in those highly conserved region. We then generate global mapping plot of coupling entropy *H(i, j)* with certain positions in binding site. Fig.5~8 shows some results with *j* position in 318, 372, 322, and 339. Table 2~5 shows the ranking of these *H(i, j)*. The results of network searching are listed in Table 6~8, with different thresholds. We selected initial targets as those residues in binding-sites. The comparison between different thresholds is shown in Table 9. Fig. 11~15 demonstrate the mapping with those high *H(i, j)* to three-dimensional structure.

1EY0 contains 136 residues. The numbering begins from 6 to 141. Fig. 19~21 demonstrate the mapping with those high *H(i, j)* to three-dimensional structure, with *j* positions on hydrophobic core.

4. Discussion

4.1 PDZ domain

From Fig.4 and Table.1, we can find that many positions in 1BE9 show low degree of conservation. This is consistent with the previous statement that PDZ domains are evolutionarily well represented protein family[5]. Although global mapping of some *H(i)* can verify many binding site residues, it's still difficult to verify the coupling degree between

them.

As for the global mapping of *H(i, j)* of 1BE9, we pick 4 binding-site residues, 318, 372, 322, 339, to be our examples. From the results we could make a roughly suggestion that *H(i, j)* may be good in three-dimension prediction. From Table 3, we can find that the highest residue except the 372 itself is residue 329. We thus compare this result to MacKinnon's schematic plot of binding site in Fig.12. It is interesting to find that position 372 and 329 are in direct contact in three-dimension structure. Furthermore, they even bind to the same residue of the substrate with hydrogen bind. This result of coupling can be accounted by either direct contact or functional interaction between residues. In other example, such as residue 339, also shows a highest *H(i, j)* partner residue 326, which is also directly contact in space. For even more, residue 339 and 326 also binds to the same residue of peptide substrate (Fig. 13). Fig.15 shows the results of residue 318. In this result, the top *H(i, j)* value residues are few in binding site, but show spatial arrangement, instead. This result may suggest that in some positions *H(i, j)* may only identify those residues in proximal but without functional correlation.

Network searching methods can verify "strong coupling pairs", which responsible for terminating the network. With our searching strategy, the network would always been terminated in the coupling pairs with in average high *H(i, j)*. Fig.16 shows these "Strong coupling pairs" mapping on three-dimensional structure. The meaning of these coupling pairs needs more study to realize.

Threshold in searching network may serve as a "filter" that screens out "strong coupling pairs." From Table 9, we can easily discover that when threshold set to as much high as 0.5 in 1BE9, the searched network would only those linked with binding site residues. This suggests that coupling residues in binding site are in average with high *H(i, j)* relative to other positions. The strong coupling of these binding site residues may give a hint that these residues are not only with three-dimensional contact, but also with "functionality attractions".

4.2 Staphylococcal Nuclease

Fig. 19~21 show the *H(i, j)* mapping results with *j* position at 92,23, and 72, respectively. From the results of Stites and Chen in 2001[19], six residues in the hydrophobic core are highly coupled, which are 23, 25, 66, 72, 92, and, 99. Our result can only identify very few of them (Fig. 19, 92 to 23). This suggests that analysis in correlated mutation may have little power to identify energetic connections, at least in the Staphylococcal Nuclease dataset. This result agrees with the conclusion made by Aldrich and Fodor in 2004 that analysis in correlated mutations can hardly find energetically connect residues[18].

5. Conclusions

Globally mapping of some coupling entropy can verify some residues that contact in 3D structure. Some can even identify functional correlated residues. Network searching methods can verify “Strong coupling pairs”, which responsible for terminating the networks. Threshold in searching network may serve as an useful “filter” that screens out “Strong coupling pairs.” The coupling residues verified by correlated mutation can be accounted by directly contact. And some coupling residues in functional sites can be explained by functional correlation. Our results do not support the hypothesis that analysis in correlated mutations can identify those with energetic connections.

6. References

1. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. BIOINFORMATICS 2003;19(1):163–164.
2. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Engineering 2002;15(9):721-725.
3. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. BIOINFORMATICS 2005;21(22):4116–4124.
4. Martin LC, Gloor GB, Dunn SD, Wahl LM. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. Biochemistry 2005;44:7156-7165.
5. Lockless SW, Ranganathan R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science 1999;286:295-299.
6. Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. nature structural biology 2003;10(1):59-69.
7. Gether U. Uncovering molecular mechanisms involved in activation of G proteincoupled receptors. Endocr Rev 2000;21:90-113.
8. Menon ST, Han M, Sakmar TP. Rhodopsin: structural basis of molecular physiology. Physiol Rev 2001;81:1659-1688.
9. Carter PJ, Winter G, Wilkinson AJ, Fersht AR. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). Cell 1984;38:835-840.
10. Hidalgo P, MacKinnon R. Revealing the architecture of a K+ channel pore through mutant cycles with a peptide inhibitor Science 1995;268:307-310.
11. Horovitz A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. Fold Des 1996;1:R121-R126.
12. JANG DS, CHA HJ, CHA S-S, HONG BH, HA N-C, LEE JY, OH B-H, LEE H-S, CHOI KY. Structural double-mutant cycle analysis of a hydrogen bond network in

ketosteroid isomerase from Pseudomonas putida biotype B. Biochem J 2004;382:967-973.

13. Schreiber G, Fersht AR. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 1995;248:478-486.
14. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Genetics 1994;18(4):309-317.
15. Singer M, Oliveira L, Vriend G, Shepherd G. Potential ligand-binding residues in rat olfactory receptors identified by correlated mutation analysis. Receptors Channels 1995;3:89-95.
16. Oliveira L, ACM P, Vriend G. Correlated Mutation Analyses on Very Large Sequence Families. ChemBioChem 2002;3:1010-1017.
17. Tillier ERM, Lui TWH. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. BIOINFORMATICS 2003;19(6):750-755.
18. Fodor AA, Aldrich RW. On Evolutionary Conservation of Thermodynamic Coupling in Proteins. The Journal of Biological Chemistry 2004;279(18):19046-19050.
19. Chen J, Stites WE. Energetics of Side Chain Packing in Staphylococcal Nuclease Assessed by Systematic Double Mutant Cycles. Biochemistry 2001;40:14004-14011.
20. Dodge C, Schneider R, Sander C. The HSSP database of protein structure–sequence alignments and family profiles. Nucleic Acids Research 1998;26(1):313-315.
21. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R. Crystal Structures of a Complexed and Peptide-Free Membrane Protein–Binding Domain: Molecular Basis of Peptide Recognition by PDZ. Cell 1996;85:1067-1076.

7. Tables

Table.1. The ranking of *H(i)* on protein 1BE9 for top most conserved positions (Residues on binding site are in boldface)

Position *i*	*H(i)*
347	0.10
324	0.11
357	0.11
356	0.40
363	0.41
415	0.50
351	0.50
301	0.50
353	0.57
414	0.65
362	0.69
375	0.70
360	0.79
412	0.80
413	0.81
329	0.83
338	0.84
341	0.88
327	0.92
379	0.93
359	0.94
364	0.96
322	0.96
330	0.99

Table.2. The ranking of *H(i,318)* on protein 1BE9 for top most conserved positions (Residues on binding site are in boldface)

Position *i*	*H(i, 318)*
311	0.52
312	0.41
314	0.40
319	0.38
316	0.34
310	0.30
317	0.30
315	0.30
320	0.29
387	0.28
309	0.26
313	0.24
391	0.23
412	0.22
308	0.22
342	0.22
388	0.21
303	0.21
334	0.21
413	0.20
355	0.20
405	0.19
394	0.19

Table.3. The ranking of *H(i,372)* on protein 1BE9 for top most conserved positions (Residues on binding site are in boldface)

Position *i*	*H(i, 372)*
329	0.72
382	0.64
322	0.63
339	0.56
380	0.54
323	0.52
326	0.51
327	0.51
340	0.46
325	0.45
331	0.43
348	0.41
355	0.40
330	0.39
376	0.38
367	0.36
362	0.36
378	0.36
346	0.36
350	0.35
345	0.33
334	0.32
361	0.32
354	0.30

Table.4. The ranking of *H(i,322)* on protein 1BE9 for top most conserved positions (Residues on binding site are in boldface)

Position *i*	*H(i, 322)*
372	0.63
380	0.57
326	0.45
382	0.44
346	0.36
386	0.35
348	0.34
350	0.34
361	0.34
339	0.33
366	0.32
323	0.32
331	0.32
367	0.30
321	0.30
377	0.30
378	0.29
345	0.29
332	0.29
340	0.28
376	0.27
337	0.27
333	0.27
362	0.27

Table.5. The ranking of *H(i,339)* on protein 1BE9 for top most conserved positions (Residues on binding site are in boldface)

Position *i*	*H(i, 339)*
326	0.65
372	0.56
331	0.56
340	0.55
370	0.53
355	0.52
382	0.51
368	0.48
328	0.47
352	0.47
380	0.46
330	0.45
321	0.42
323	0.42
334	0.41
350	0.40
361	0.39
378	0.38
383	0.37
327	0.36
377	0.36
365	0.36
358	0.36
349	0.36
348	0.35

Table.6. Network initiated from binding site residues of 1BE9 with no threshold.

(Residues on binding site are in boldface)

Initial position *i*	**Network**	**Number of Network elements**
318	**318** → 319 → 317 → 319	4
321	**321** → **322** → 346 → 349 → 346	5
322	**322** → 346 → 349 → 346	4
323	**323** → 386 → **323**	3
324	**324** → **325** → 340 → 342 → 340	5
325	**325** → 340 → 342 → 340	4
326	**326** → **327** → **326**	3
327	**327** → **326** → **327**	3
329	**329** → **372** → **329**	3
339	**339** → **326** → **327** → **326**	4
372	**372** → **329** → **372**	3
379	**379** → **323** → 386 → **323**	4
383	**383** → 384 → **383**	3

Table.7. Network initiated from binding site residues with threshold = 0.3.

(Residues on binding site are in boldface)

Initial position *i*	**Network**	**Number of Network elements**
318	**318** → 319 → 317 → 319	4
321	**321** → **322** → 346 → 349 → 346	5
322	**322** → 346 → 349 → 346	4
323	**323** → 386 → **323**	3
324	**324**	1
325	**325** → 340 → 342 → 340	4
326	**326** → **327** → **326**	3
327	**327** → **326** → **327**	3
329	**329** → **372** → **329**	3
339	**339** → **326** → **327** → **326**	4
372	**372** → **329** → **372**	3
379	**379** → **323** → 386 → **323**	4
383	**383** → 384 → **383**	3

Table.8. Network initiated from binding site residues with threshold = 0.5.

(Residues on binding site are in boldface)

Initial position *i*	Network	Number of Network elements
318	**318**	1
321	**321**	1
322	**322**	1
323	**323**	1
324	**324**	1
325	**325**	1
326	**326 → 327 → 326**	3
327	**327 → 326 → 327**	3
329	**329 → 372 → 329**	3
339	**339 → 326 → 327 → 326**	4
372	**372 → 329 → 372**	3
379	**379**	1
383	**383**	1

Table.9. Comparison between networks with different threshold

Initial position *i*	**No threshold**	**Threshold = 0.3**	**Threshold = 0.5**
318	**318** → 319 → 317 → 319	**318** → 319 → 317 → 319	**318**
321	**321** → 322 → 346 → 349 → 346	**321** → **322** → 346 → 349 → 346	**321**
322	**322** → 346 → 349 → 346	**322** → 346 → 349 → 346	**322**
323	**323** → 386 → **323**	**323** → 386 → **323**	**323**
324	**324** → 325 → 340 → 342 → 340	**324**	**324**
325	**325** → 340 → 342 → 340	**325** →340 → 342 → 340	**325**
326	**326** → **327** → **326**	**326** → **327** → **326**	**326** → **327** → **326**
327	**327** → **326** → **327**	**327** → **326** → **327**	**327** → **326** → **327**
329	**329** → **372** → **329**	**329** → **372** → **329**	**329** → **372** → **329**
339	**339** → **326** → **327** → **326**	339 → 326 → 327 → **326**	**339** → **326** → **327** → **326**
372	**372** → **329** → **372**	372 → **329** → **372**	**372** → **329** → **372**
379	**379** → **323** → 386 → **323**	**379** → **323** → 386 → **323**	**379**
383	**383** → 384 → **383**	**383** → 384 → **383**	**383**

8. Figures

	(A)	(B)	(C)	(D)	(E)
	G. . M G. . M G. . M C. . T C. . T C. . T	G. . M G. . M C. . T C. . T D. . F D. . F	G. . M G. . M C. . M C. . M D. . F D. . F	G. . M G. . M C. . M D. . T D. . F D. . F	G. . M G. . T C. . M C. . T C. . M C. . T
$H(i, j)$	0.058	0.122	0.056	0.058	0
$H'(i, j)$	1	1	0.496	0.357	0

Figure.1

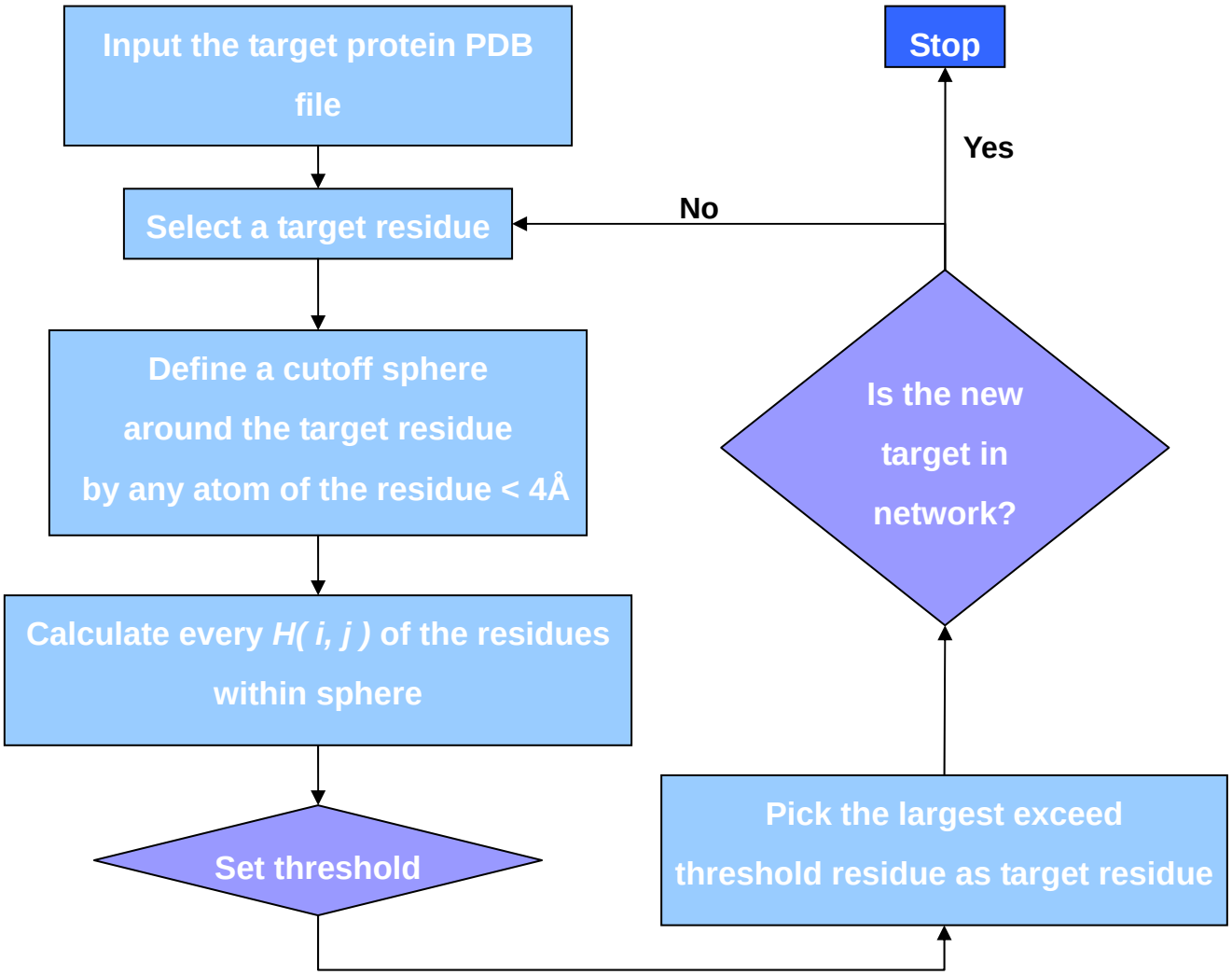
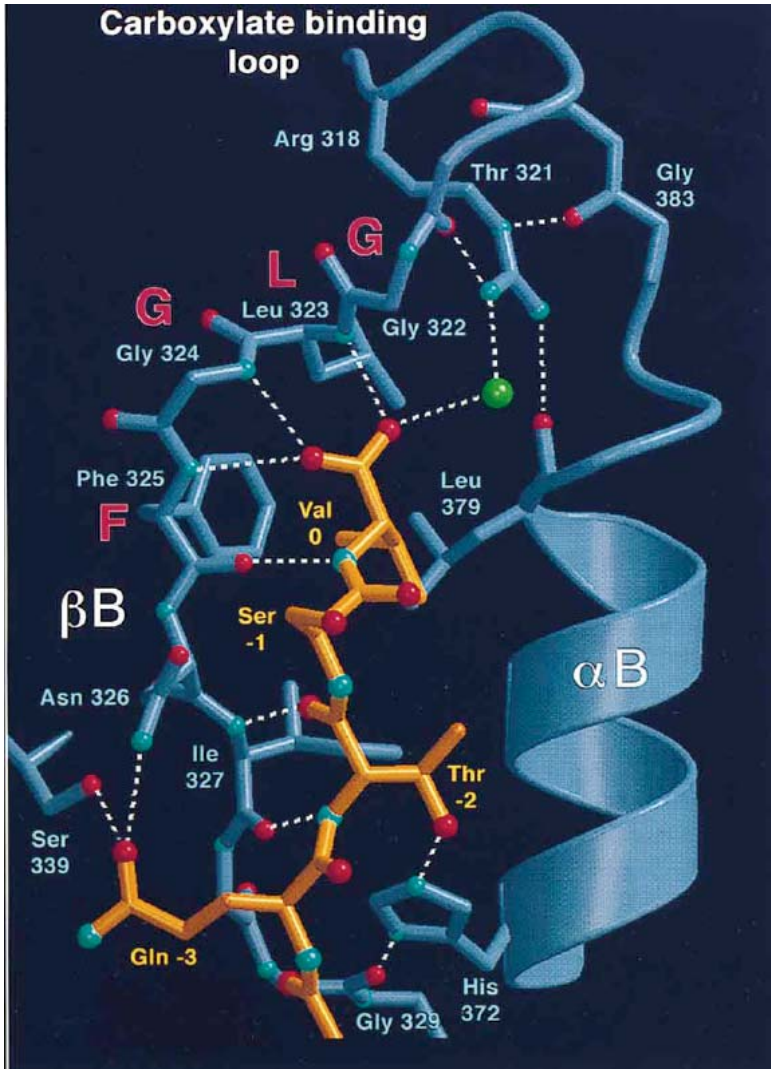

Input the target protein PDB file
Select a target residue
Define a cutoff sphere around the target residue by any atom of the residue < 4Å
Calculate every H(i, j) of the residues within sphere
Set threshold
Pick the largest exceed threshold residue as target residue
Is the new target in network?
Yes
No
Stop

Figure.2



αB region
372H
379L
383G
βB region
322G
323L
324G
325F
326N
327I
329G
339S
Loop region
318K
321T

Figure.3

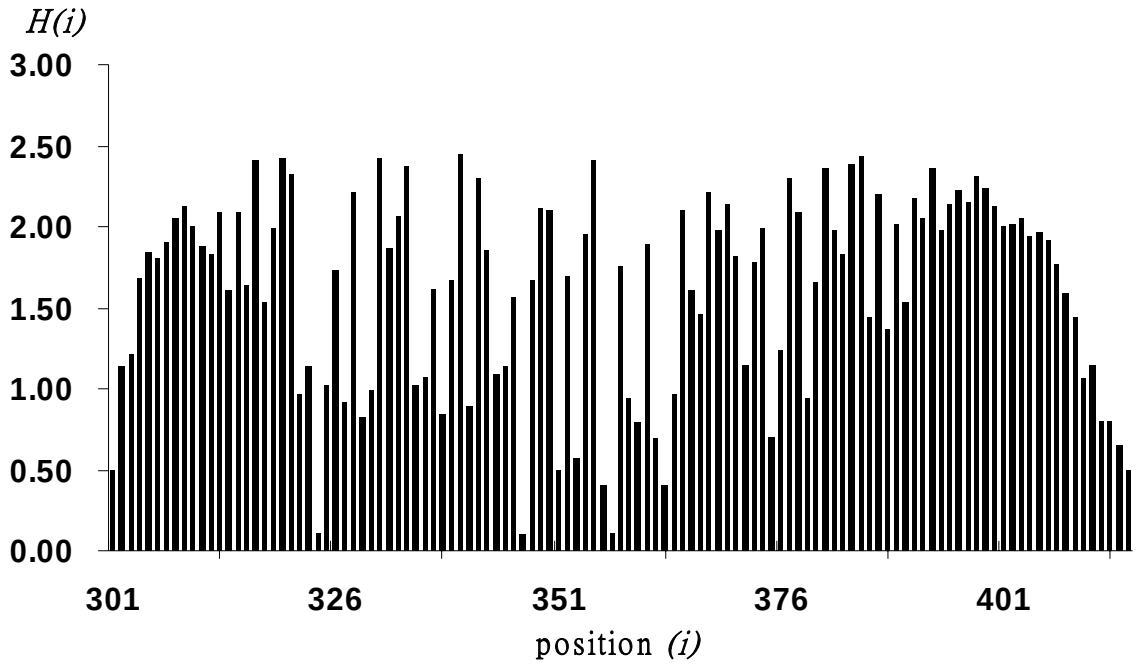


Figure.4

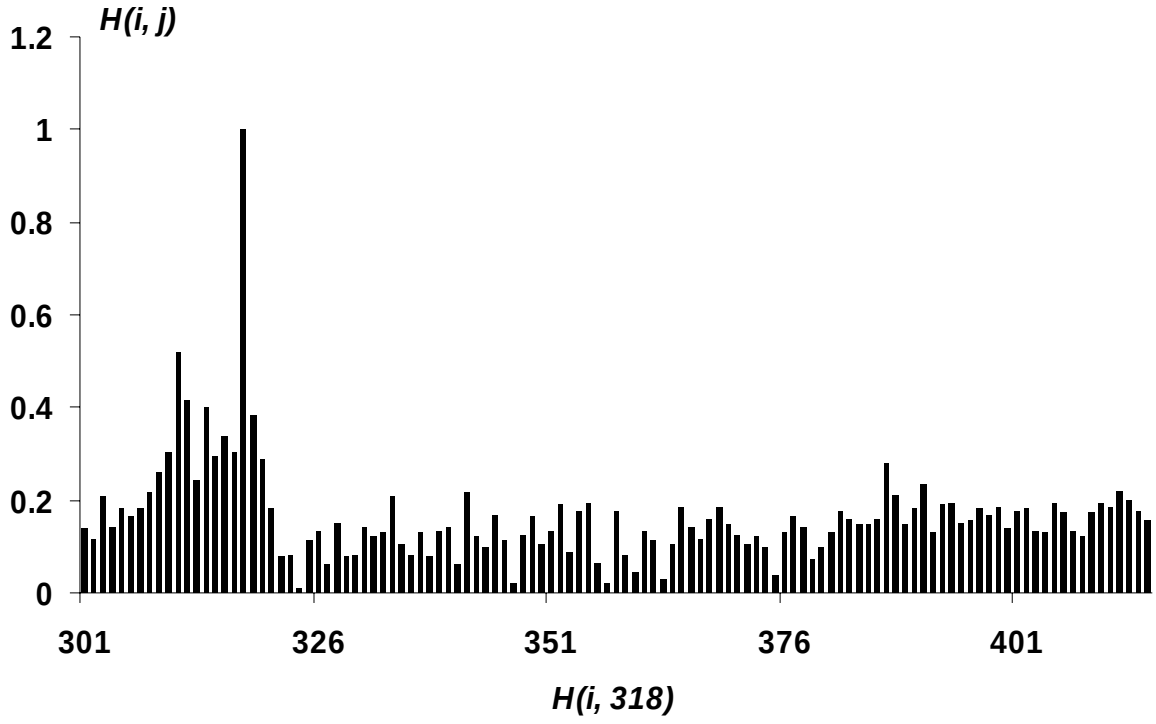


Figure.5

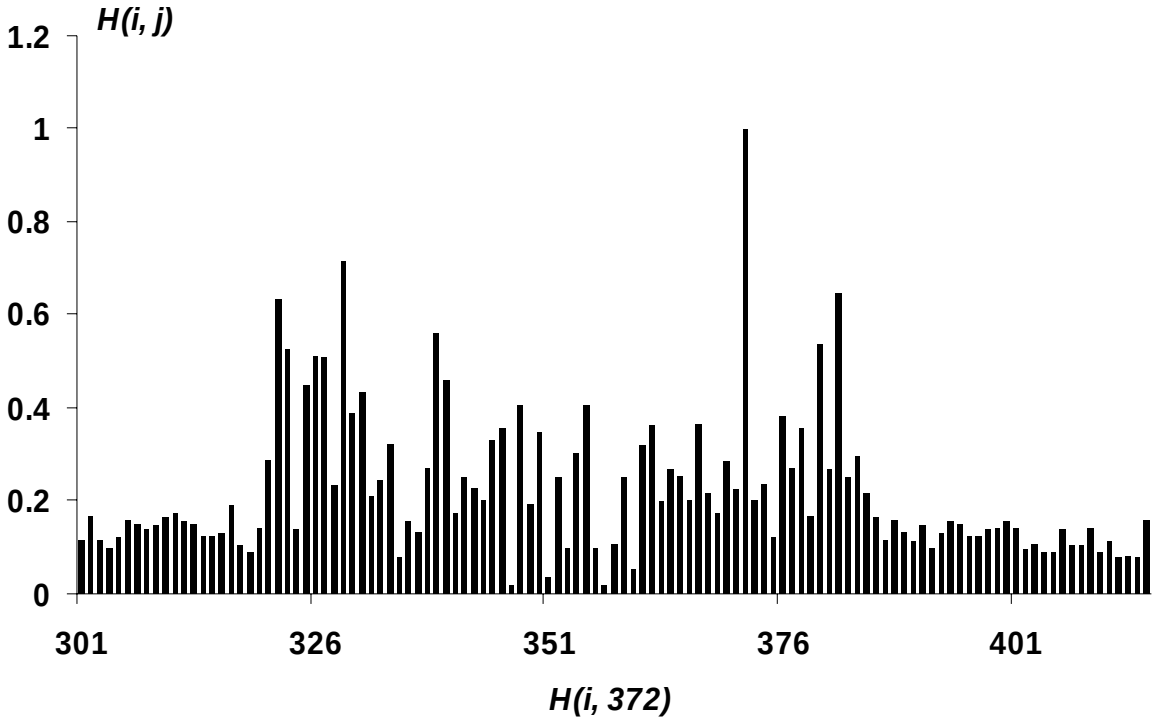


Figure.6

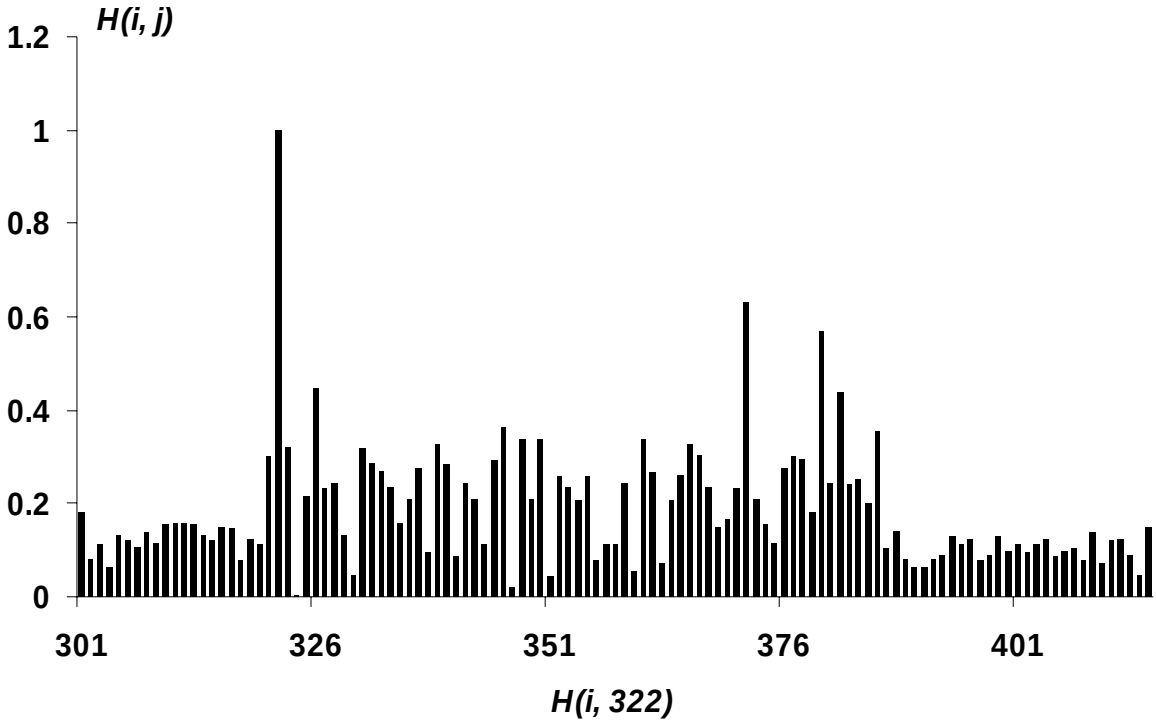


Figure.7

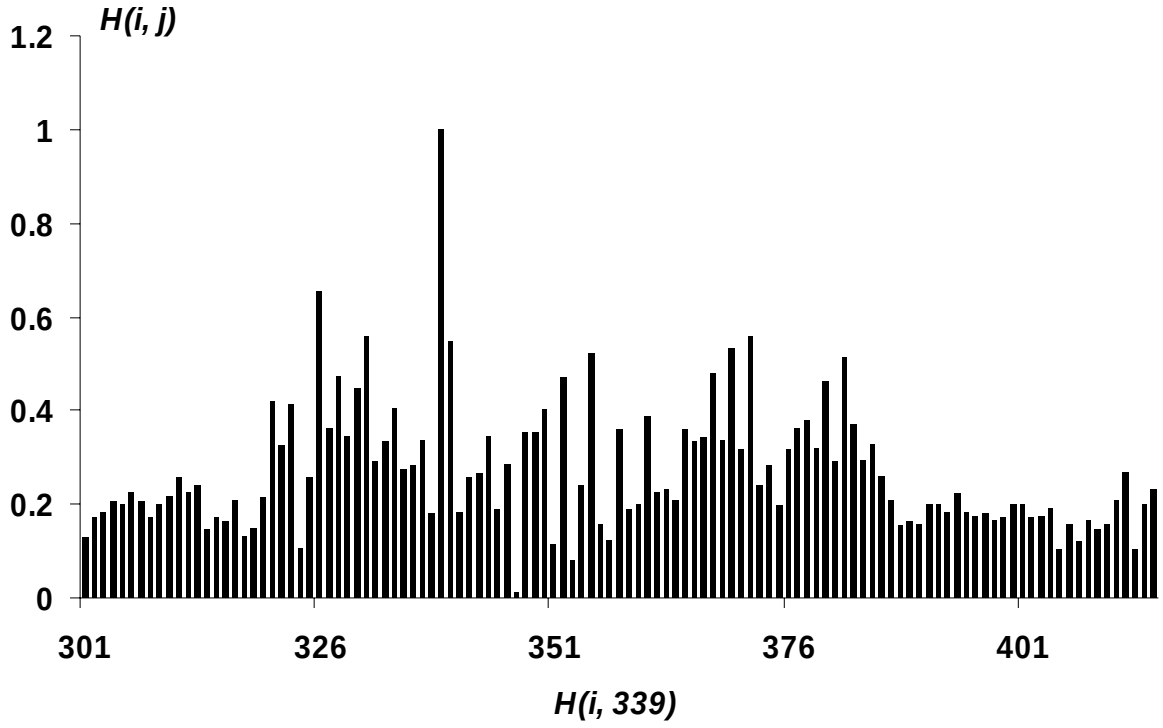


Figure.8

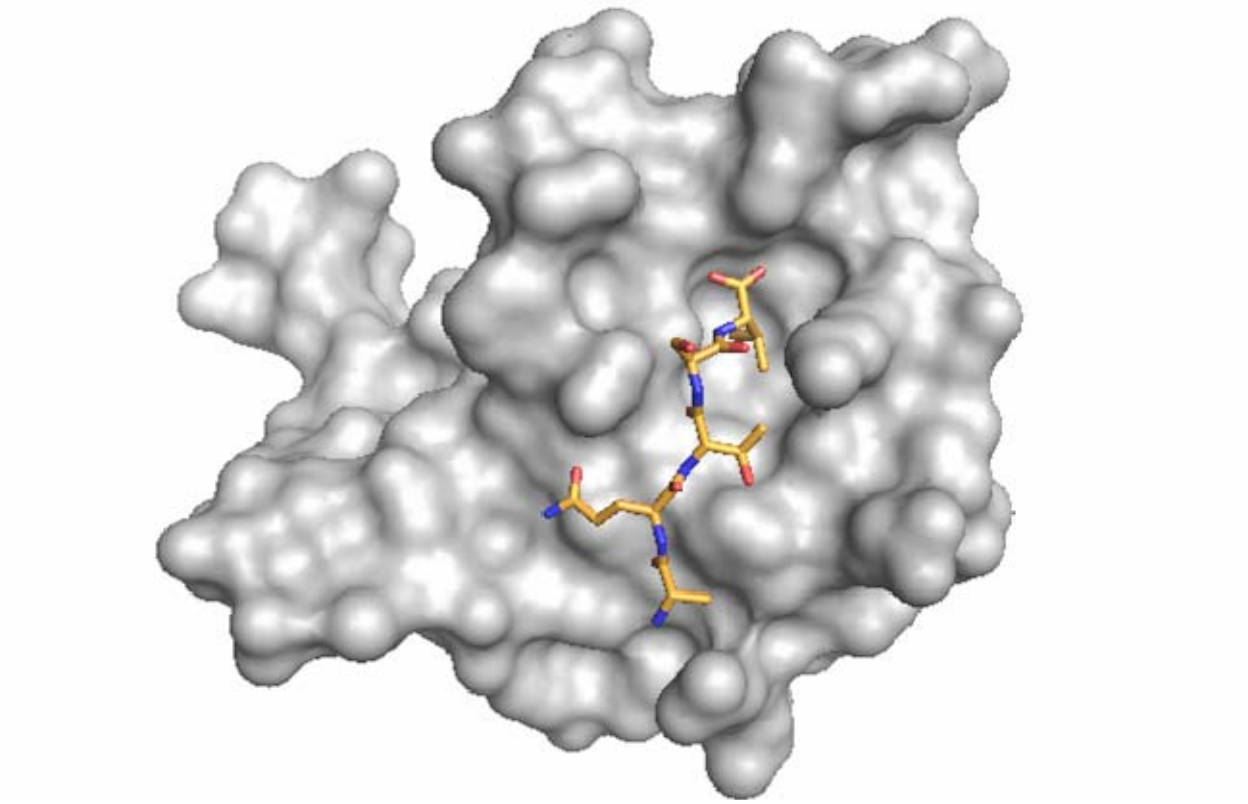


Figure.9

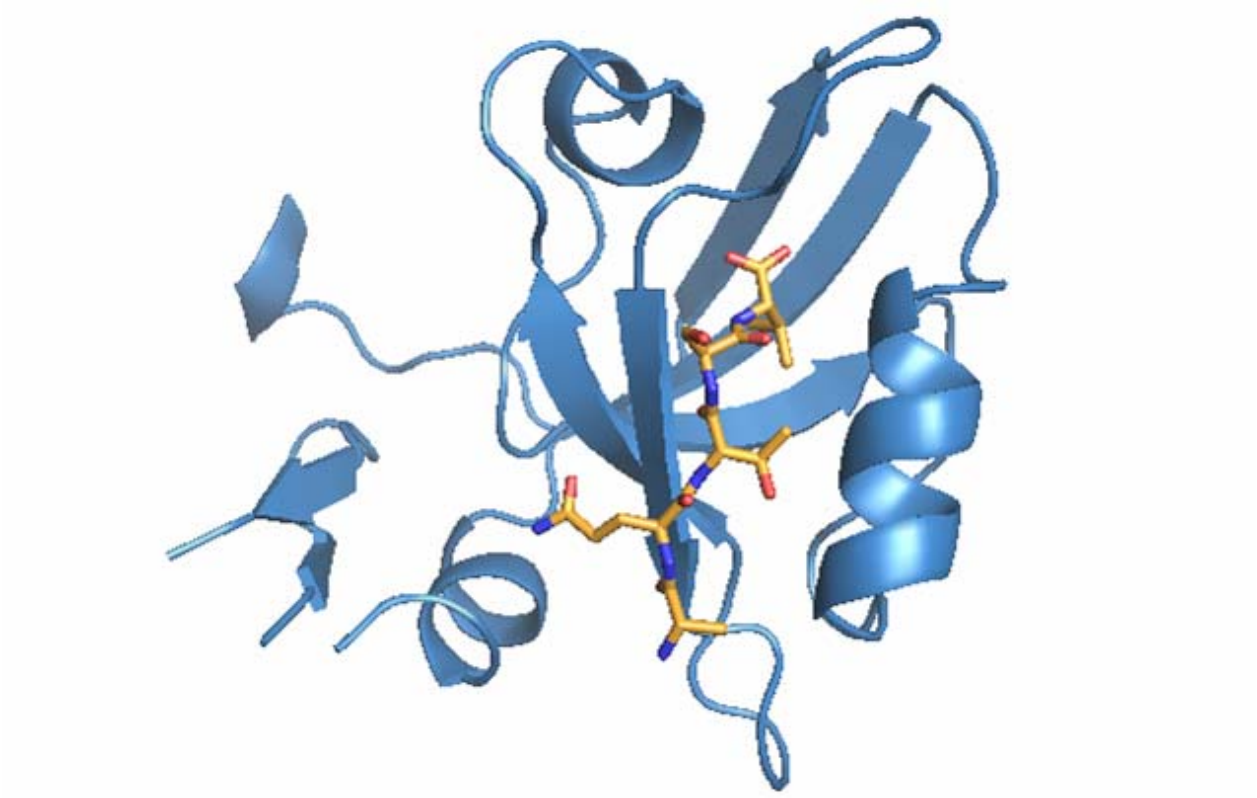


Figure.10

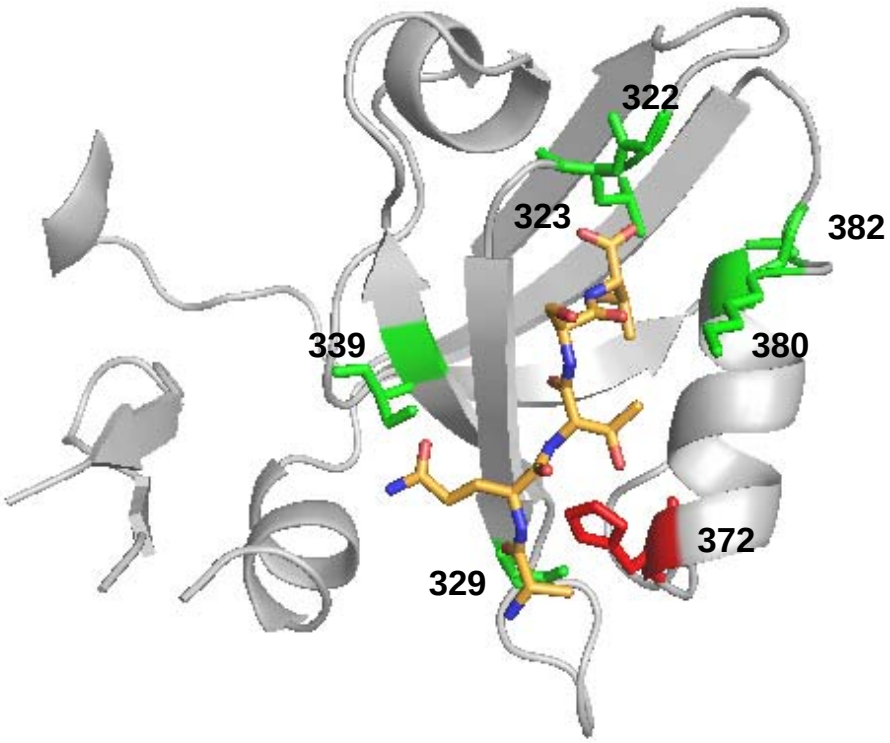
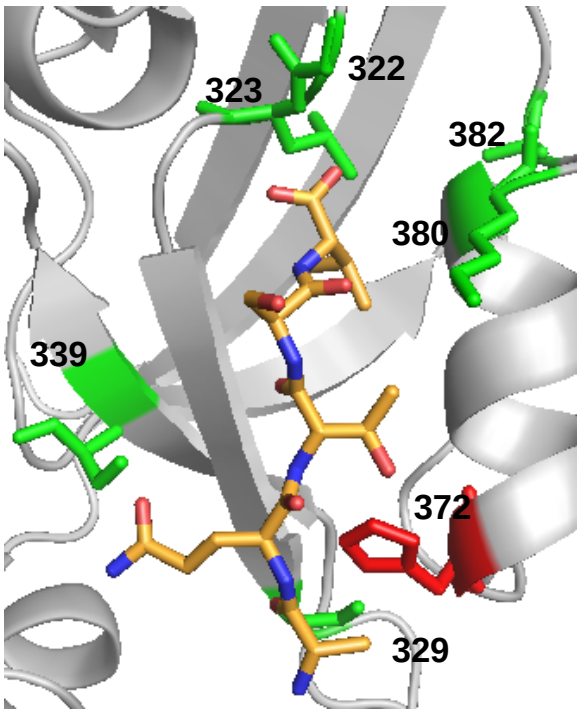


Figure.11


323
322
382
380
339
372
329

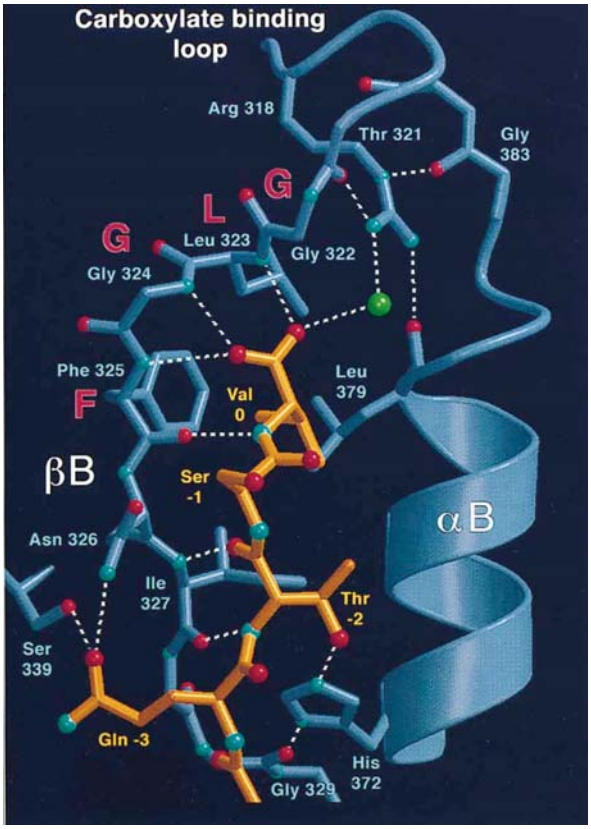

Carboxylate binding loop
Arg 318
Thr 321
Gly 383
G
L
G
G
Leu 323
Gly 322
Gly 324
Phe 325
F
Val 0
Leu 379
βB
Ser -1
αB
Asn 326
Ile 327
Thr -2
Ser 339
Gln -3
His 372
Gly 329

Figure.12

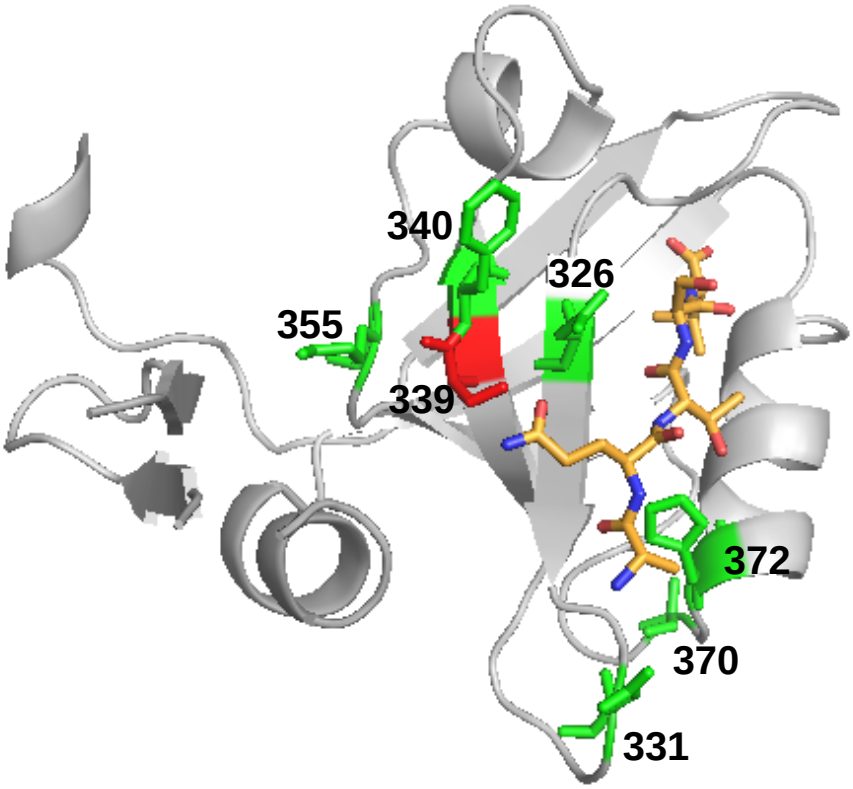
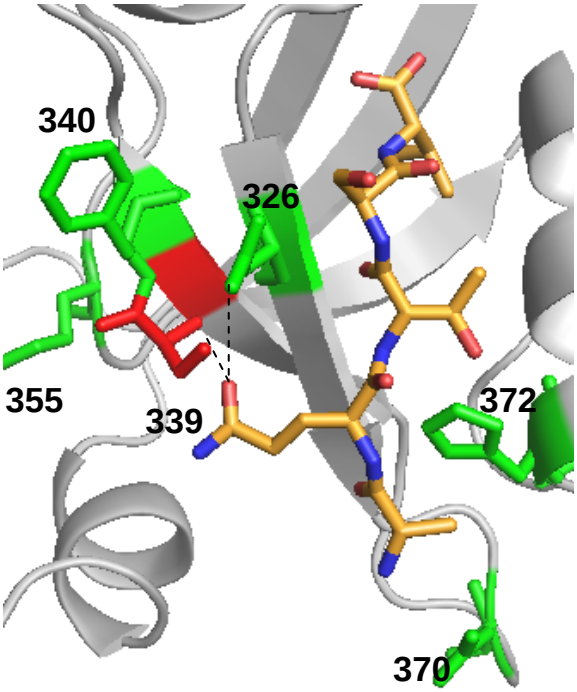


Figure.13


340
326
355
339
372
370

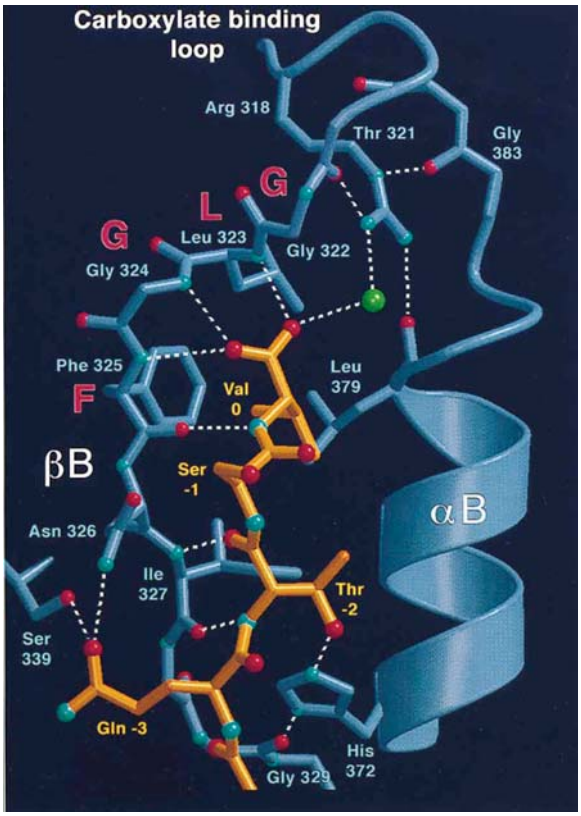

Carboxylate binding loop
Arg 318
Thr 321
Gly 383
G
L
G
G
Leu 323
Gly 322
Gly 324
Phe 325
Leu 379
F
Val 0
βB
Ser -1
αB
Asn 326
Ile 327
Thr -2
Ser 339
Gln -3
His 372
Gly 329

Figure.14

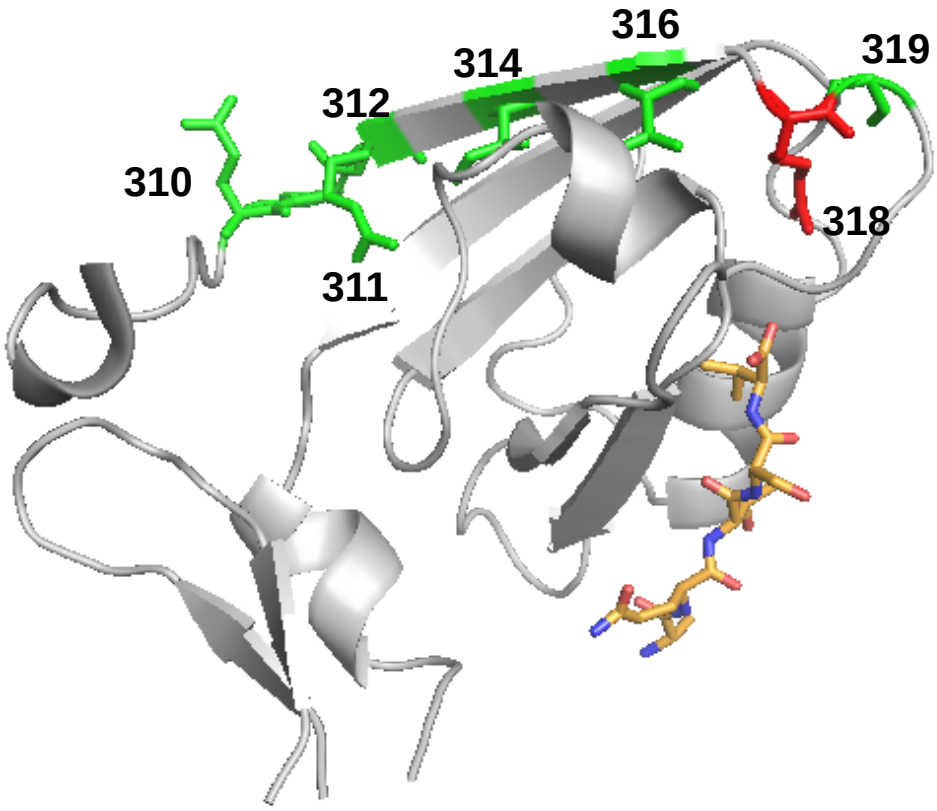


Figure.15

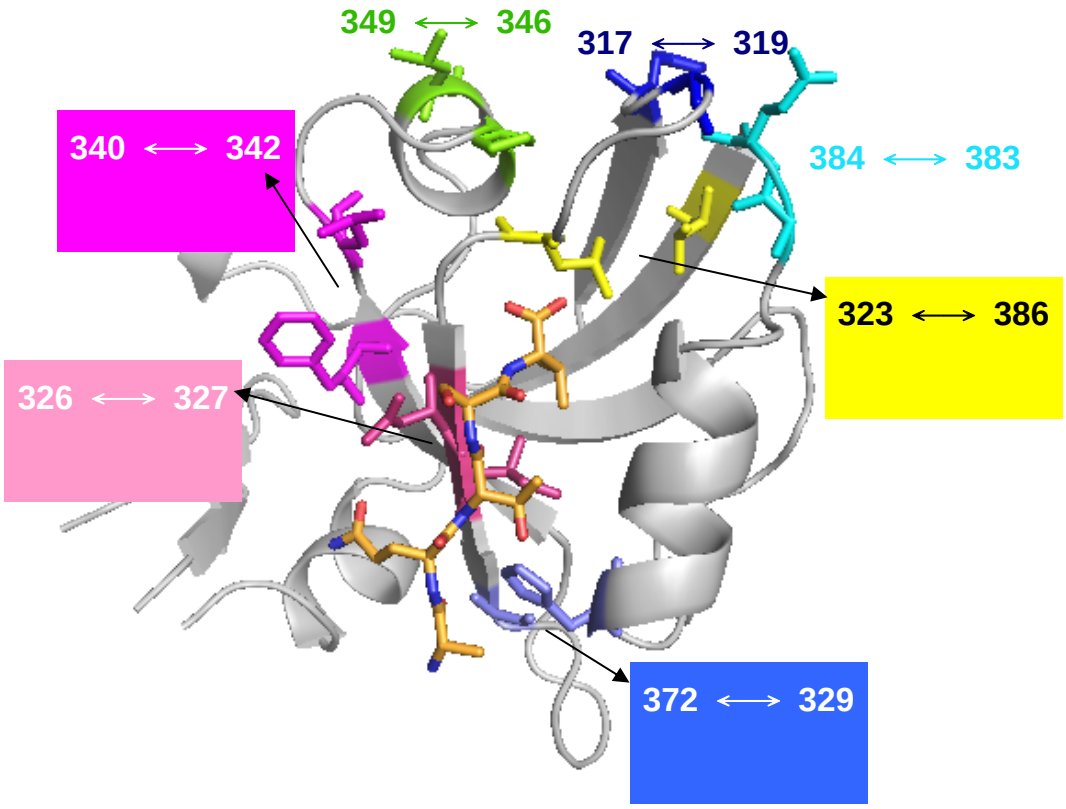


Figure.16

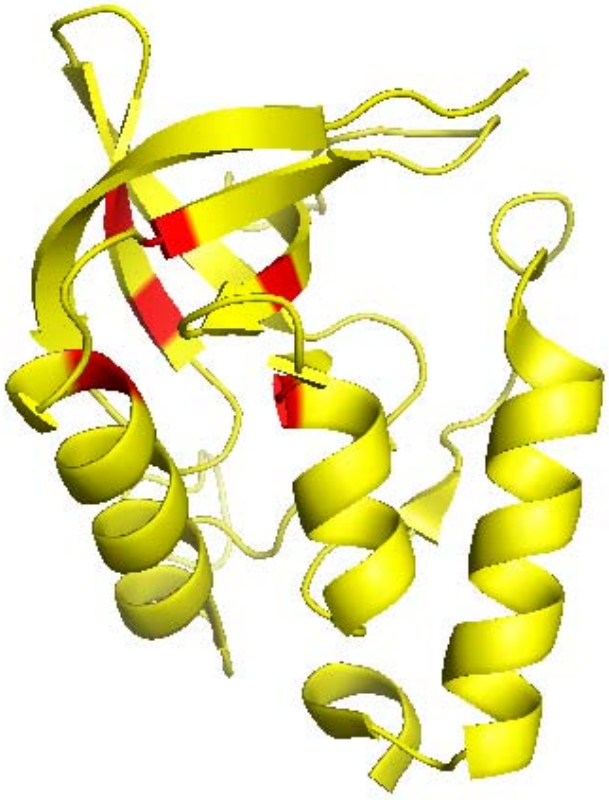


Figure.17

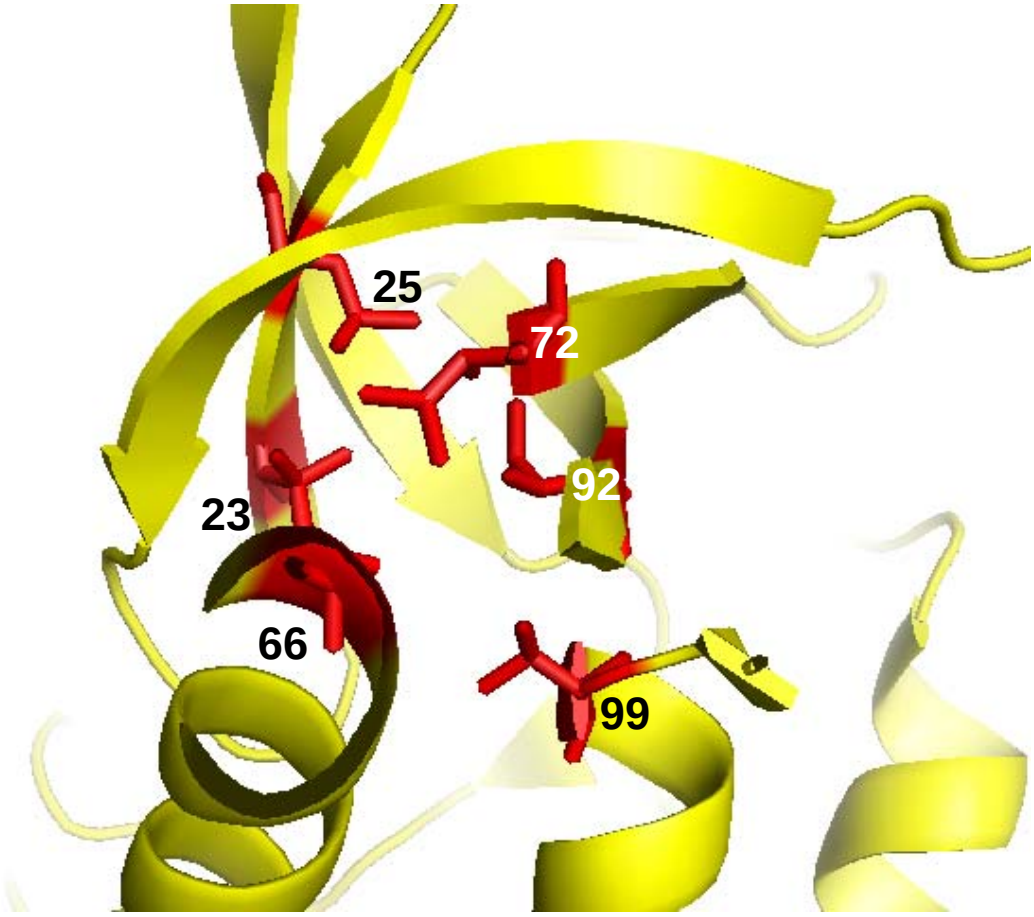


Figure.18

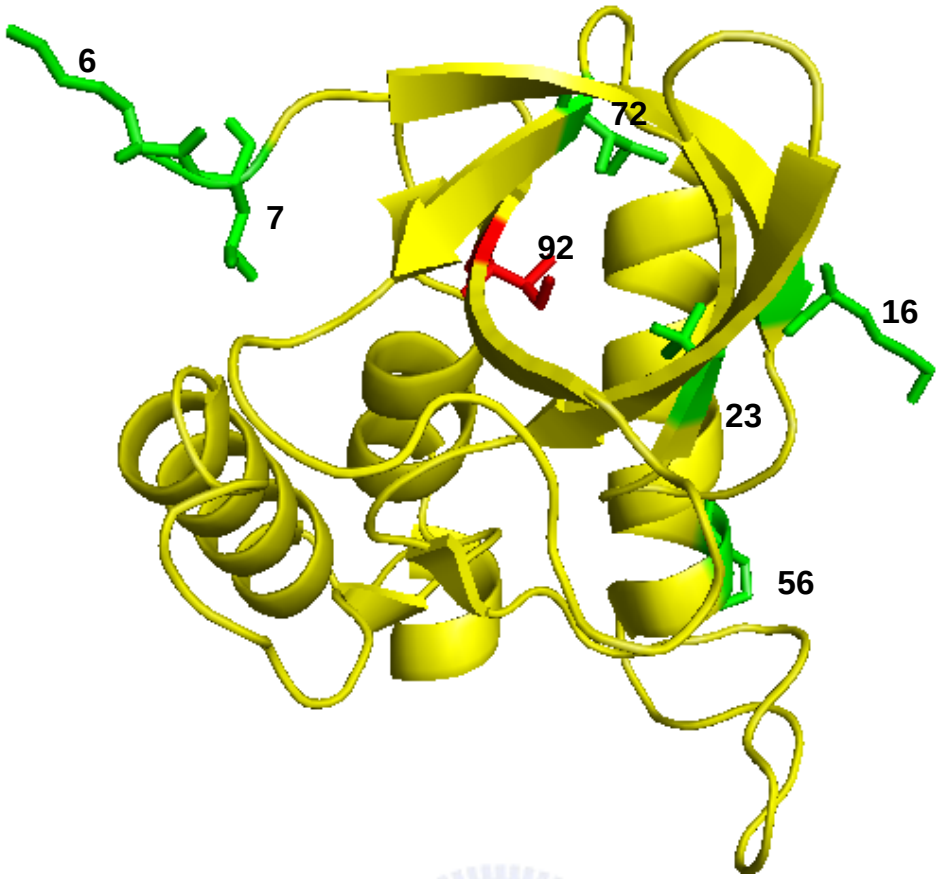


Figure.19

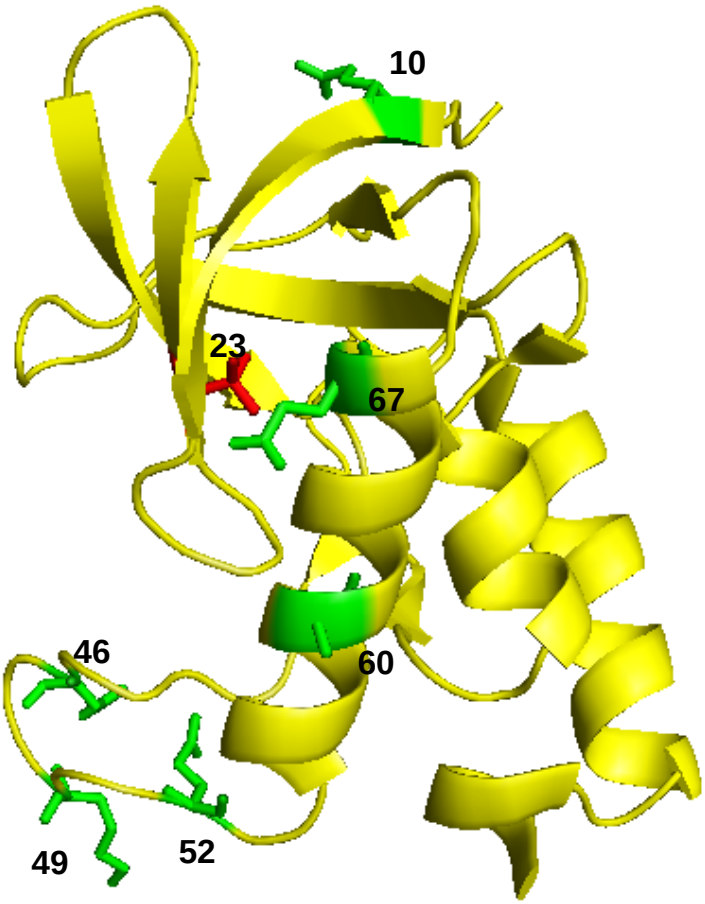


Figure.20

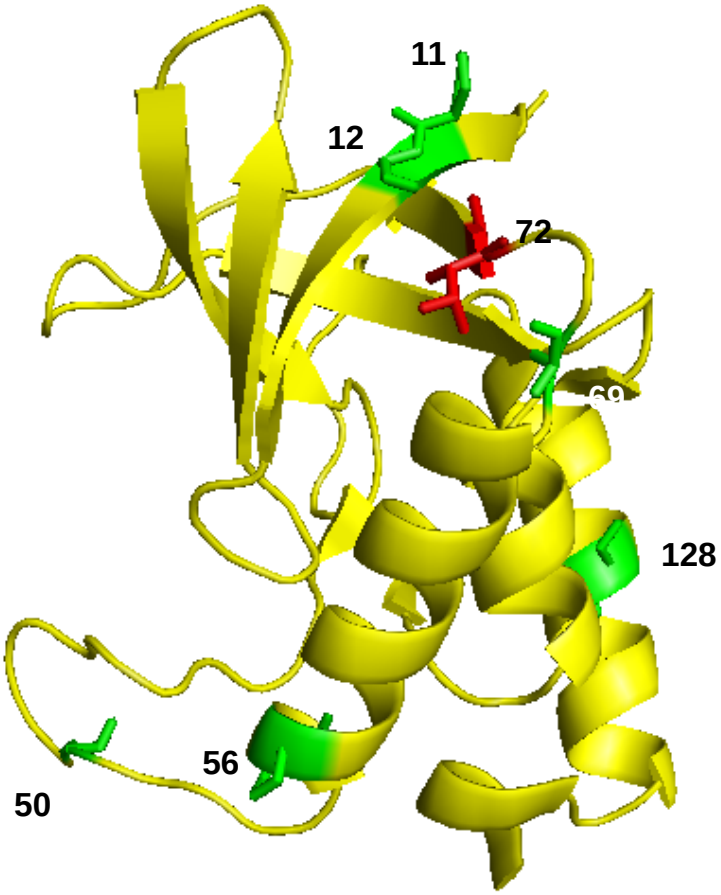


Figure.21