Introduction

In recent years, the molecular simulations has increased to display the conformational change and function in proteins. The molecular simulations are all-atom simulations, so it needs much more computer computing and time. The coarse-grained models such as elastic network model(ENM), which only uses α carbon atoms to analize proteins motion . Therefore, the ENM model is more effectively to process the large proteins information than all-atom molecular simulation.

Here we develop the methods, for example, CM、COS、WCN、PCN, whose autocorrelations approximately correspond with the crystallographic B factor(temperature factors or Debye-Waller factors). We use these methods to discuss the dynamical correlation and functions in proteins and compare the normalized and non-normalized data. Finally, we find the correlation between binding sites and hot spot residues.

Method

Gaussian network model (GNM)

GNM(Gaussian network model)[5,6] is derived from the elastic network theory which is assumed that every residue in a protein can be represented by a node in elastic

network model. Each node contacts as springs and fluctuates with Gaussian distribution in the uniform harmonic potential. In this model, protein is coarse-grained by adopting C_α atoms. The overall conformational potential of a structure is given by

$$V = (\gamma / 2)\Delta R^T \Gamma \Delta R \qquad (1)$$

where ΔR is the N-dimentional vector of fluctuations of individual residues and Γ is a symmetric matrix known as the Kirchhoff connectivity matrix defined below :

$$\Gamma_{ij} = \begin{cases} -1 & if \quad i \neq j \quad , \quad r_{ij} \leq R_c \\ 0 & if \quad i \neq j \quad , \quad r_{ij} > R_c \\ -\sum_{k,k\neq 1}^{N} \Gamma_{ik} & if \quad i = j \end{cases} \qquad (2)$$

The cross-correlation between residue fluctuation are found from

$$\langle \Delta R_i \cdot \Delta R_j \rangle = \frac{3k_B T}{\gamma} \left[\Gamma^{-1} \right]_{ij} \qquad (3)$$

and mean-square fluctuation of individual residues can be readily found from analogous expression :

$$\langle \Delta R_i \cdot \Delta R_i \rangle = \left\langle (\Delta R_i^{\,2}) \right\rangle = \frac{3k_B T}{\gamma} \left[\Gamma^{-1} \right]_{ii} \qquad (4)$$

The inverse of the Kirchhoff matrix can be expressed as an expansion of eigenvalues λ and eigenvectors u of Γ

$$\langle \Delta R_i \cdot \Delta R_j \rangle = \frac{3k_B T}{\gamma} \sum_k \left[\lambda_k^{-1} \quad u_k \quad u_k^T \right]_{ij} \qquad (5)$$

Centroid model (CM)

The distance between atom i and the center of mass of a protein is expressed as :

$$r_i^2 = \langle X_i - X_0 \rangle \langle X_i - X_0 \rangle \qquad (6)$$

where X_0 is equal to the center of mass of a protein. If a protein with Cα atoms of size N, and the square of distance of each Cα atom distribution is given by

$$(r_1^2, r_2^2, ..., r_n^2) \qquad (7)$$

The r^2 profile closely accord with the crystallographic temperature factors (also named thermal B factors or Debey-Waller factors). The thermal B factor is given as

$$B_i = 8\pi^2 / 3 \langle \delta X_i \delta X_i \rangle \qquad (8)$$

and the results of CM[1] suggest the following relation :

$$\langle \delta X_i \delta X_i \rangle \sim \langle X_i - X_0 \rangle \langle X_i - X_0 \rangle \qquad (9)$$

This equation(9) can offer us the information about the fluctuation of a residue usually is proportional to the distance between its position and center of mass.

The correlation between atoms in proteins can be written as :

$$c_{ij} = (X_i - X_0)(X_i - X_0) \qquad (10)$$

when i = j , $c_{ij} = r_i^2$

From eq.(10), we can also get the following equation according to the normal mode of

analysis(NMA)[2,3] :

$$\langle \delta X_i . \delta X_j \rangle \sim \sum_k \frac{V_{ik} V_{kj}}{L_k} \qquad (11)$$

In eq.(11), i, j : α carbon atom number, k : mode number, V : eigen vector, L : eigen value. It can represent the situation of fluctuation of each atom in every mode.

The normalization of centroid model(COS)

The correlation between two atoms is relative to the position of two atoms, which can be expressed as :

$$C_{ij}^{\cos} = \frac{(\vec{x}_i - \vec{x}_0) \cdot (\vec{x}_j - \vec{x}_0)}{|\vec{x}_i - \vec{x}_0||\vec{x}_j - \vec{x}_0|} \qquad (12)$$

In eq.(12), i, j : α carbon atom number, $\vec{x}_0$ is the vector to the center of mass of a protein. The normalization of CM is equal to COS.

Weighted contact-number model(WCN)

It has much relationship between the thermal B factor and neighboring contact of atoms. When the neighboring contact number of a atom is more large, the lesser fluctuation it will has.Thus we can define WCN model[4] as :

$$W_{ii} = \left(\sum_{k}^{N} \frac{1}{r_{ik} r_{ik}} \right)^{-1} \hat{x}_i \cdot \hat{x}_i$$
$$= \left(\sum_{k}^{N} \frac{1}{r_{ik}^{\,2}} \right)^{-1} \qquad (13)$$

where i, k = atom number of protein, $\hat{x}_i$ is the unit vector in the direction of $\sum_{k}^{N} r_i - r_k$.

We can find this relationship :

$$\langle \delta X_i \cdot \delta X_i \rangle \sim \left(\sum_{k}^{N} \frac{1}{r_{ik}^{\,2}} \right)^{-1} \qquad (14)$$

We can define the correlation term W_{ij} between residue i and residue j as :

$$W_{ij} = \left(\sum_{k \neq i,j}^{N} \frac{1}{r_{ik} r_{jk}} \right)^{-1} \hat{x}_i \cdot \hat{x}_j \qquad (15)$$

$\hat{x}_i$ and $\hat{x}_j$ are the unit vectors in the direction of $\sum_{k}^{N} r_i - r_k$ and $\sum_{k}^{N} r_j - r_k$

The correlation between fluctuation can also be computed with the method of normal mode analysis(NMA) :

$$\langle \delta r_i \cdot \delta r_j \rangle \sim \sum_{k \neq i,j}^{N} \frac{U_{ik} U_{jk}}{\lambda_k} \qquad (16)$$

λ_k : the eigenvalues of the kth mode, U_{ik} : the eigenvector of the kth mode.

The improvement of weighted contact-number model (iWCN)

iWCN is the improvement of WCN which formula can be expressed as :

$$iW_{ij} = \frac{\min\{r_i^2, r_j^2\} \cdot \cos\theta}{(r_i^2 \cdot r_j^2)^{\frac{1}{2}}} \qquad (17)$$

In eq. 17, $r_i = \left(\sum_k^N \frac{1}{r_{ik}^2} \right)^{-1}$, $r_j = \left(\sum_k^N \frac{1}{r_{jk}^2} \right)^{-1}$, i, j, k = atom number of protein, r is equal to the distance between two atoms.

Normalization

We normalize the correlation matrix by this equation :

$$C_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}}\sqrt{C_{jj}}} \qquad \textbf{(18)}$$

$C_{ij} = C_{ij}^{GNM}, C_{ij}^{CM}, C_{ij}^{COS}, C_{ij}^{WCN}, C_{ij}^{iWCN}$

i, j : atom number of protein.

Prediction Accuracy

The overall prediction accuracy(accu) is given by :

$$accu = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP : true positive , FP : false positive, TN : true negative, FN : false negative.

Data sets

The DNA binding sites of PDB code 1A1V protein and high sequence conserved residues are obtained from PDBsum[7], the ATP binding sites is from the literature

studied by Zheng et al.[8]. The experimental hot spot residues refer to Zheng et al.[8] and Darnell et al.[9] We sort the correlation data in order of correlation value and take the top percentage correlation value as high correlation with binding sites.

Result

Correlation Map

NS3 helicase is an enzyme in hepatitis C virus (HCV) that can unwind double-stranded DNA or RNA in an ATP-dependent reaction. Thus research on this enzyme can develop an effective therapeutic drugs. We take the NS3 helicase for an example (PDB code : 1A1V), and compare the dynamical correlation with these methods CM, COS, GNM, WCN and iWCN (Figure 1) which are non-normalized, and find their eigenvalues (Figure 2) and eigenvectors (Figure 3). In figure 1 and in figure 5, each pattern of every method is very similar, but the GNM is somewhat different. We can also observe that there are three main parts with high correlation, which implies three domains exist independently in space. The same methods with normalized data is in the figure 4、figure 5 and figure 6.

Eigenvalue

The eigenvalue of protein dynamical correlation is assumed to be relative to the frequency of protein fluctuation. In figure 2 , the eigenvalue is very large in CM, COS

and WCN. However, the eigenvalue is rather small in GNM and iWCN. Thus, GNM and iWCN model are more accurate to describe proteins fluctuation in low-frequency motions. The lowest-frequency motions also called global motions (as opposed to local motions subject to high-requency modes) which are usually of functional importance[10].

Eigenvector

The eigenvector can represent the amplitude of fluctuation and relative direction of vibration for each atom in the elastic network models of proteins. Relative to GNM, the correlation coefficient in CM, COS, WCN, iWCN increase from mode 1 to mode 3 in figure 3. We can see that the fluctuation types are similar in these models in higher frequency mode 1 and mode 2 . CM, COS, WCN and iWCN have much more correlation for each other than GNM. In figure 7, relative to GNM, the correlation coefficient in COS and iWCN increae from mode 1 to mode 2, but decrease from mode 2 to mode 3 . Relative to GNM, the correlation coefficient in WCN increase from mode 1 to mode 3. The correlation coefficient is more large in normalized data than in non-normalized data.

In figure 4, the main difference of fluctuation type is in GNM mode 1, which fluctuation type is more accurate than others. Because according to Bahar et al. [6], hinge motions may contribute ligand binding, the mode 1 in GNM can show that ATP binding sites is at the global hinge which is identified from the crossover between the positive and negative of the eigenvectors and ATP binding in the crossover is of functional

importance. In figure 14-B, ATP binding in the interface between domain 1 and domain 2 plays a critical role, so it is the reason that high conservation residues is convergent in this interface.

The improvement of eigenvector

Because the direction of eigenvector and eigenvalue fluctuation is more accurate in GNM which view domain 1 and domain 3 as a whole to fluctuate, we calculate the center of mass with domain 1 plus domain 3, and calculate the center of mass with domain 2, then average the 2 centers as the new center. In figure 9, we can see the new center is shown in yellow sphere, and the center of mass in CM is shown in magenta sphere. By using new center to calculate COS, CM and iWCN and compare their eigenvectors correlation coefficient with GNM, we can see that the correlation coefficient is increase in figure 11 campared to figure 10. In figure 12, the COS and iWCN accord with GNM, but CM does not.

Hot spot residues

A residue is defined as a hot spot residue when it is mutated to alanine and gives rise to a distinct drop in the binding constant (typically tenfold or higher). [11] The alanine-scanning mutagenesis can study the contributions of individual amino acid side chains of protein-protein binding interface. Bogan and Thorn[12] in their hot spot anatomy work indicate that hot spot residues tend to cluster in the center of interfaces.

In figure 3 and in figure 7, the red squares represeant hot spot residues, residue number 293 is ATP binding site, and residue number 324 is beside ATP binding site 323. Residue number 432 is DNA binding site. According to the literature studied by Yang et al.[10], the flexibility of binding sites are more lower than others, but higher than catalytic sites. We can see the absolute value of eigenvectors of residue number 324 is more small in GNM than other methods in mode 1. The absolute value of eigenvectors of residue number 432 in WCN is more smaller than other methods in mode 1. So is the same at residue number 460 in mode 2. The fluctuation of hot spots in mode 3 are almost more large than mode 1 and mode 2.

It is well known that residues have tendency to be conserved, they possibly play some important role in proteins. They may contribute to structural stability, catalysis or recognition. Hot spot residues have protein-protein or protein-ligand binding affinity function and are usually high conservation. Binding sites are also with high conservation and play an critical role in modulate the function of proteins. Thus we discuss the correlation between conserved residues and binding sites in proteins. In figure 13-A, as the conservation increase, the residues conservation larger than 9 is convergent in the interface between domain 1 and domain 2. This distribution is more consistent in iWCN top 2% than in COS top 2%, and it is more obvious in ATP binding site than in DNA binding site in table I and table II. In table III, we can observe that iWCN has much more correlation with high

conservation residues than COS. However, CM and GNM can not identify the hot spot residues, so we do not find their correlation with high conserved residues .

Hot spot residues can help us understand the affinity and specificity in protein interfaces. Only a small fraction of interface residues can dramatically affect the binding affinity by the alanine scanning. In order to make experimental design more efficient, we develop a model, like iWCN and COS, to identify the hot spot residues. In table IV and table V, iWCN is more accurate than COS in top 2%, maybe it is the reason that iWCN can recognize large percentage of high conservation residues.

Discussion

The function of proteins has much relation with geometic structures of proteins. Although GNM is more accurate in presenting the fluctuation in mode 1, it can not find the high relation between binding site and hot spot residues. The iWCN is the better method to find the correlation between binding sites and high conserved residues which include hot spot residues. This may be the reason that there are more atoms around binding sites and high conserved residues, so that proteins can effectively execute their function from the binding sites. Corresponsively, we may identify binding sites by high conserved residues in the future work.

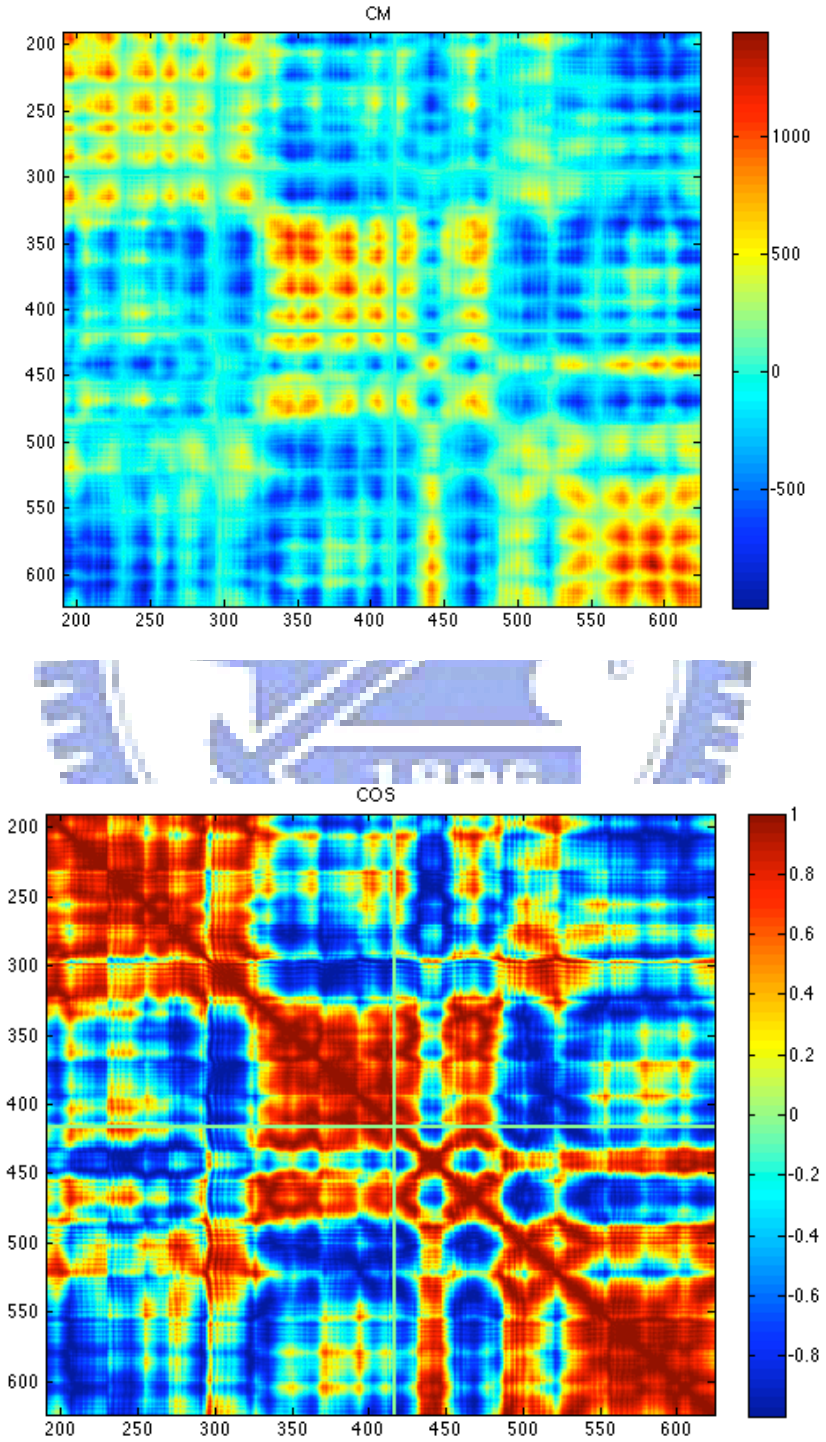
Reference

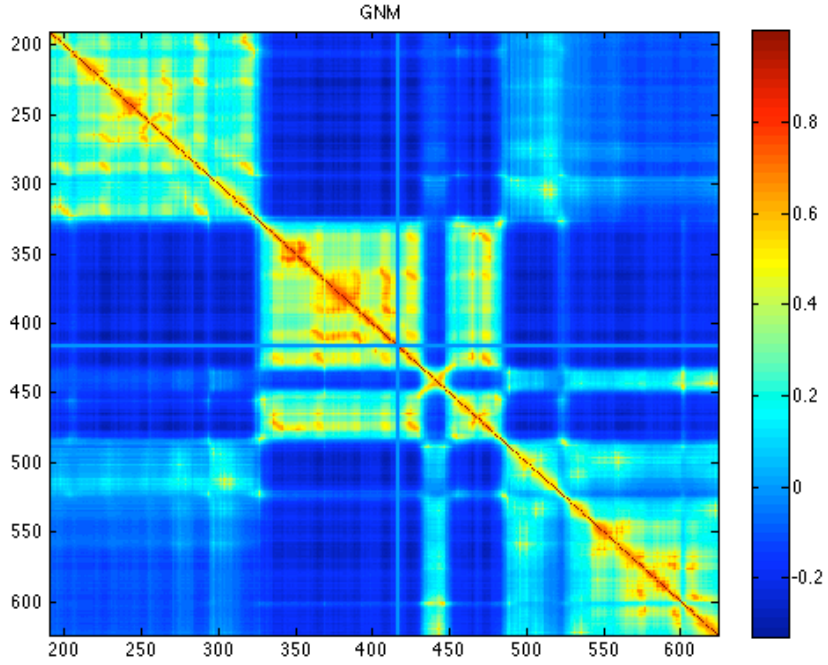
1. Shih CH, Huang SW, Yen SC, Lai YL, Yu SH, Hwang JK. A simple way to compute protein dynamics without a mechanical model. Proteins 2007;68(1):34-38.
2. Brooks B, Karplus M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci U S A 1983;80(21):6571-6575.
3. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. Journal of molecular biology 1985;181(3):423-447.
4. Lin CP, Huang SW, Lai YL, Yen SC, Shih CH, Lu CH, Huang CC, Hwang JK. Deriving protein dynamical properties from weighted protein contact number. Proteins 2008.
5. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2(3):173-181.
6. Bahar I, Jernigan RL. Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms. Journal of molecular biology 1998;281(5):871-884.
7. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic acids research 2005;33(Database issue):D266-268.
8. Zheng W, Liao JC, Brooks BR, Doniach S. Toward the mechanism of dynamical couplings and translocation in hepatitis C virus NS3 helicase using elastic network model. Proteins 2007;67(4):886-896.
9. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. Proteins 2007;68(4):813-823.
10. Yang LW, Bahar I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 2005;13(6):893-904.
11. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges.

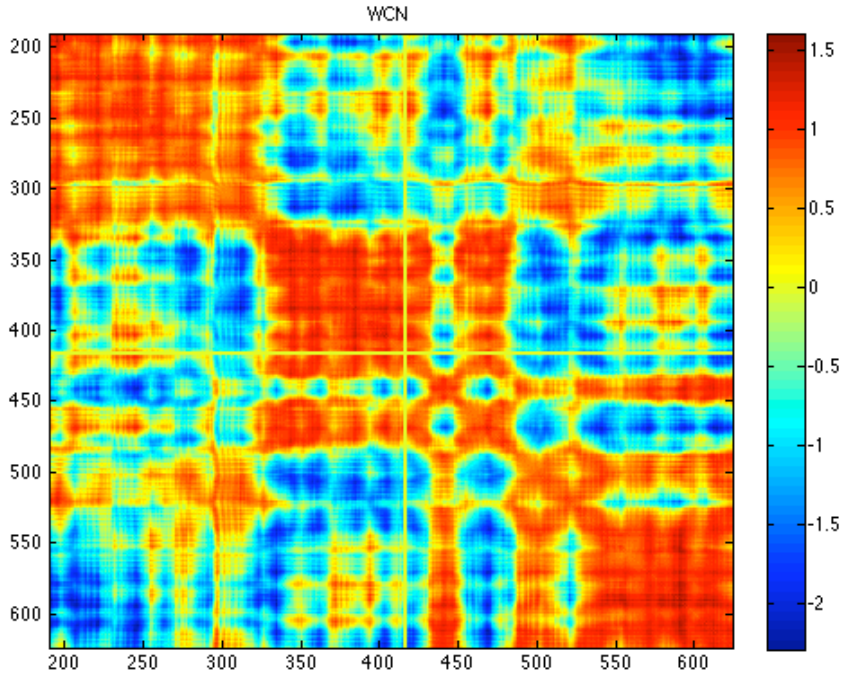
Current opinion in structural biology 2002;12(1):14-20.

12. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science (New York, NY 1995;267(5196):383-386.

Figure 1 : The non-normalize correlation map of CM, COS, GNM, WCN, iWCN.




GNM
200
250
300
350
400
450
500
550
600
0.8
0.6
0.4
0.2
0
-0.2


WCN
200
250
300
350
400
450
500
550
600
1.5
1
0.5
0
-0.5
-1
-1.5
-2

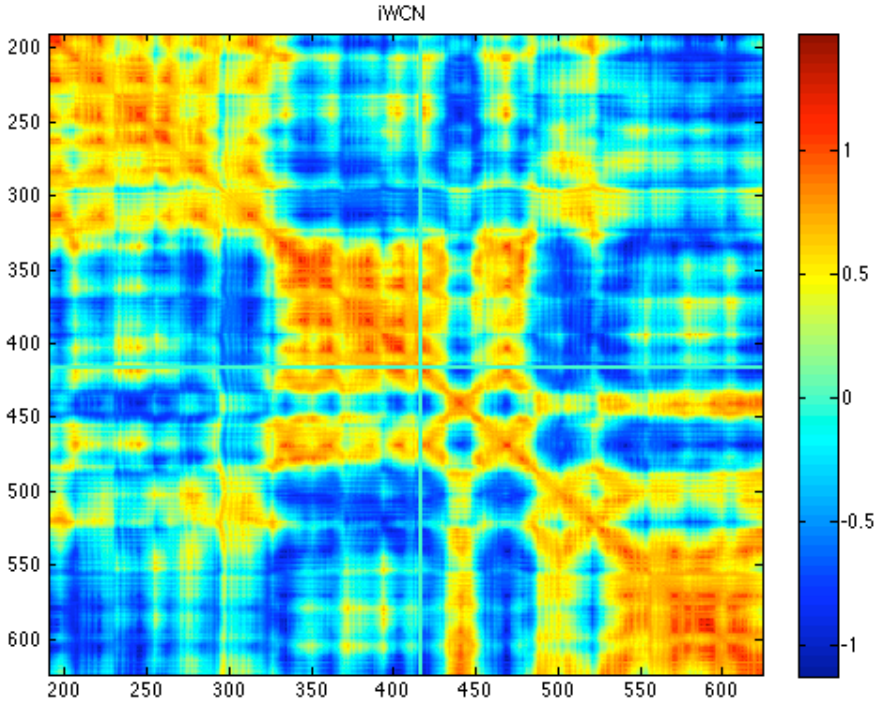
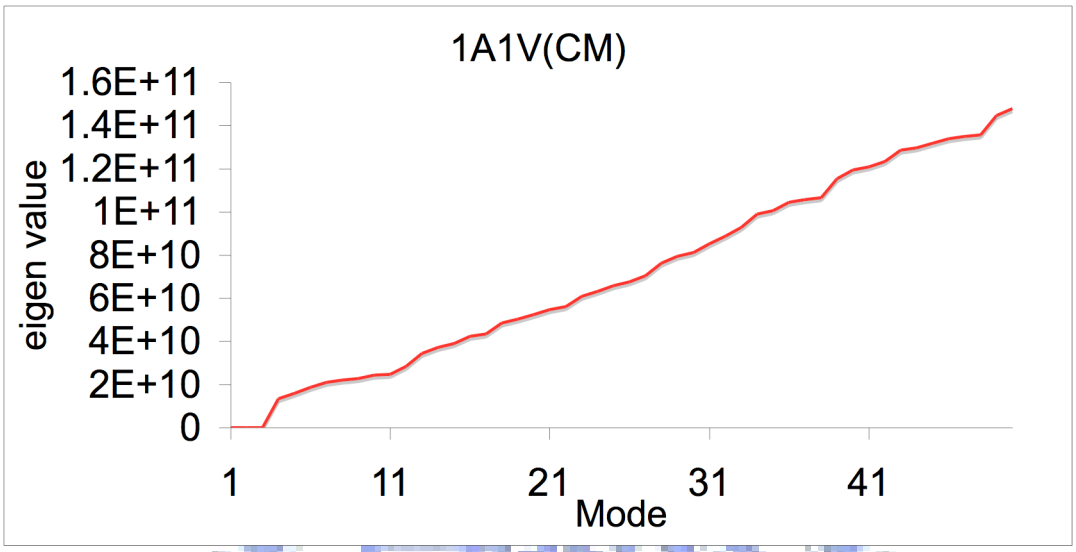
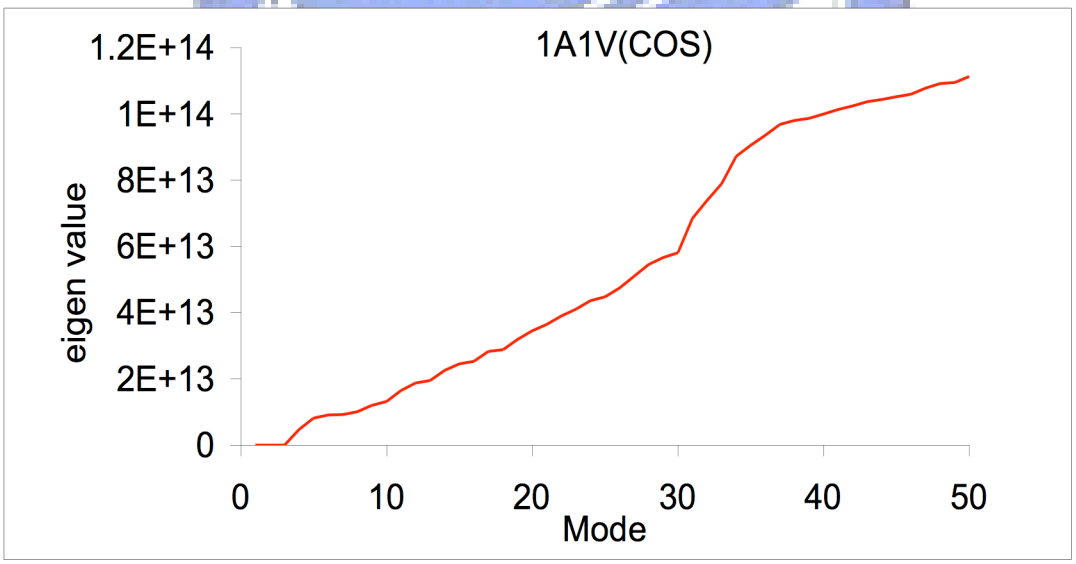
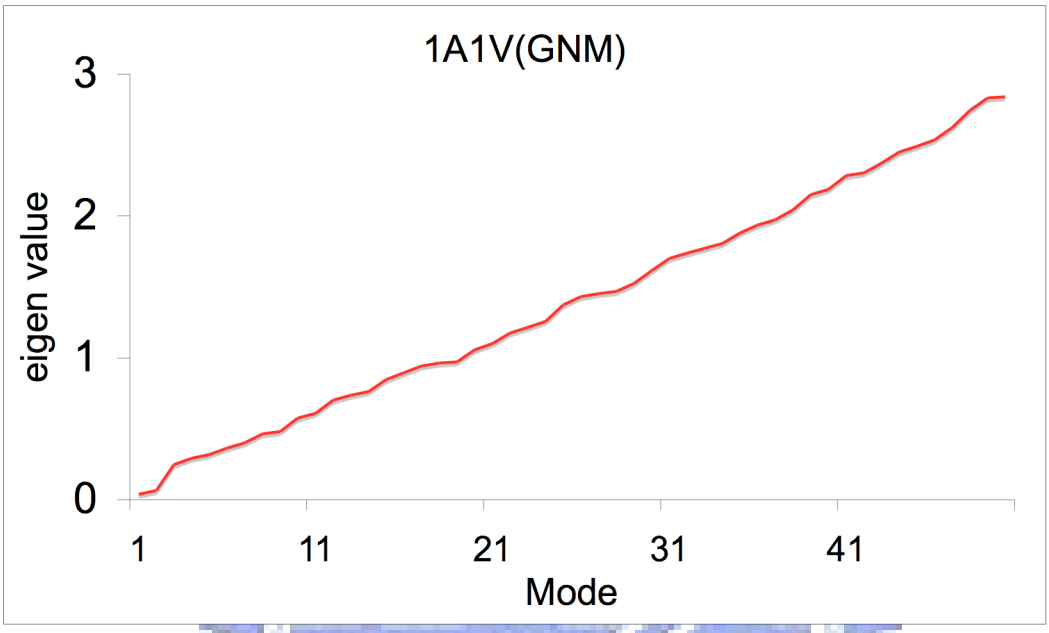
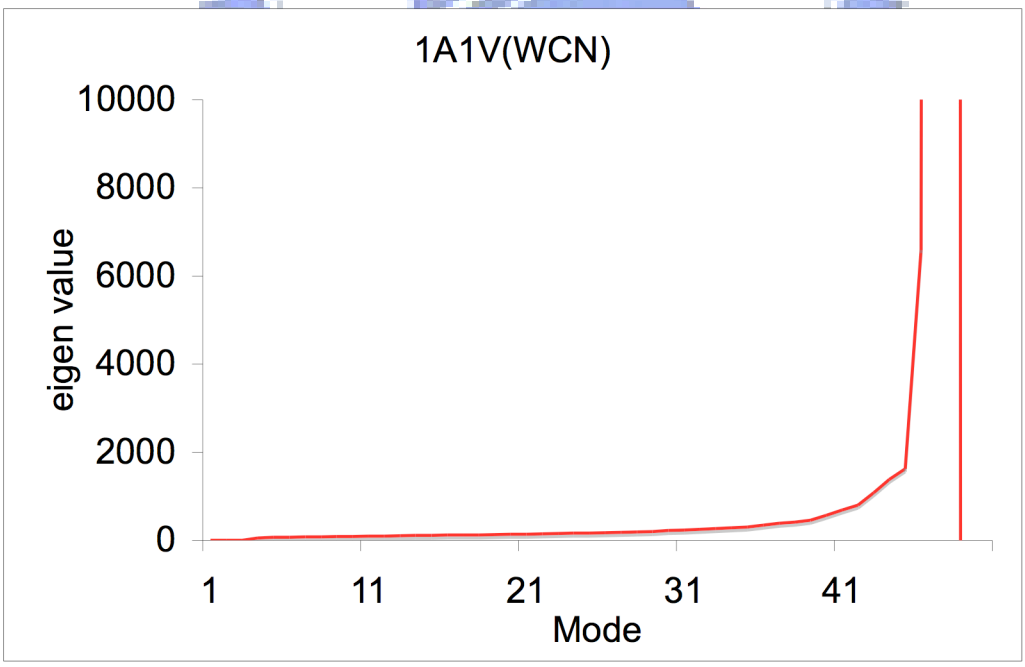

iWCN
200
250
300
350
400
450
500
550
600
1
0.5
0
-0.5
-1

Figure 2 : The eigen value of CM, COS, GNM, WCN, iWCN, which are non-normalized.






1A1V(GNM)
3
2
1
0
eigen value
1
11
21
31
41
Mode


1A1V(WCN)
10000
8000
6000
4000
2000
0
eigen value
1
11
21
31
41
Mode

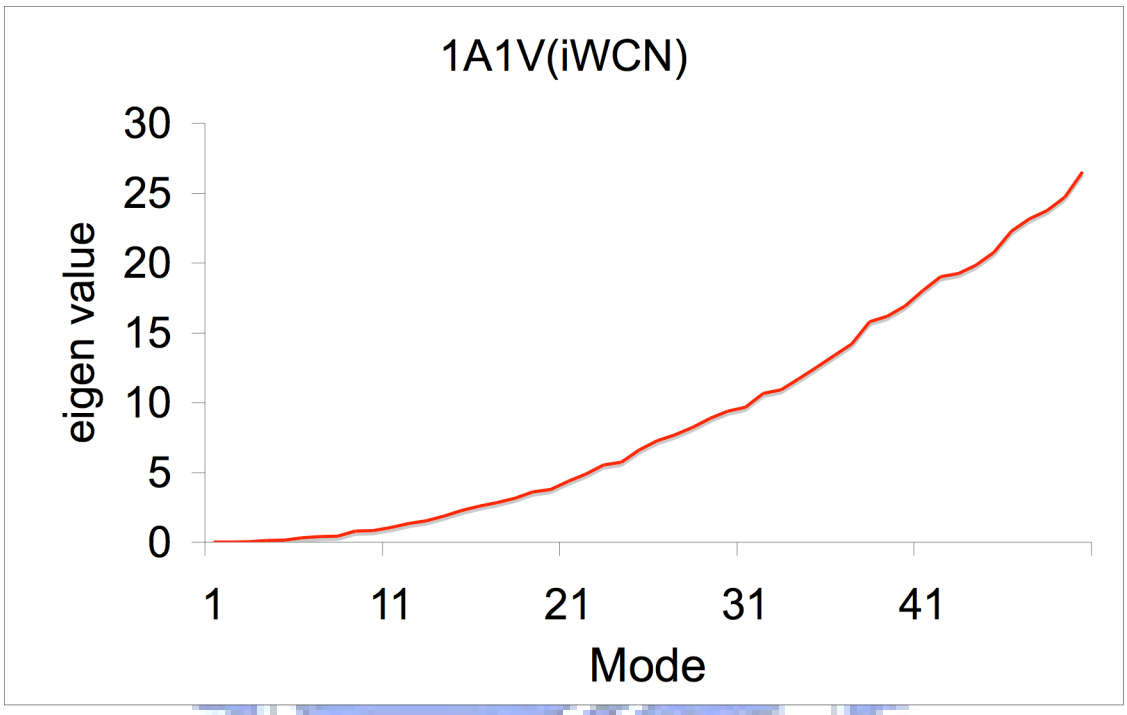
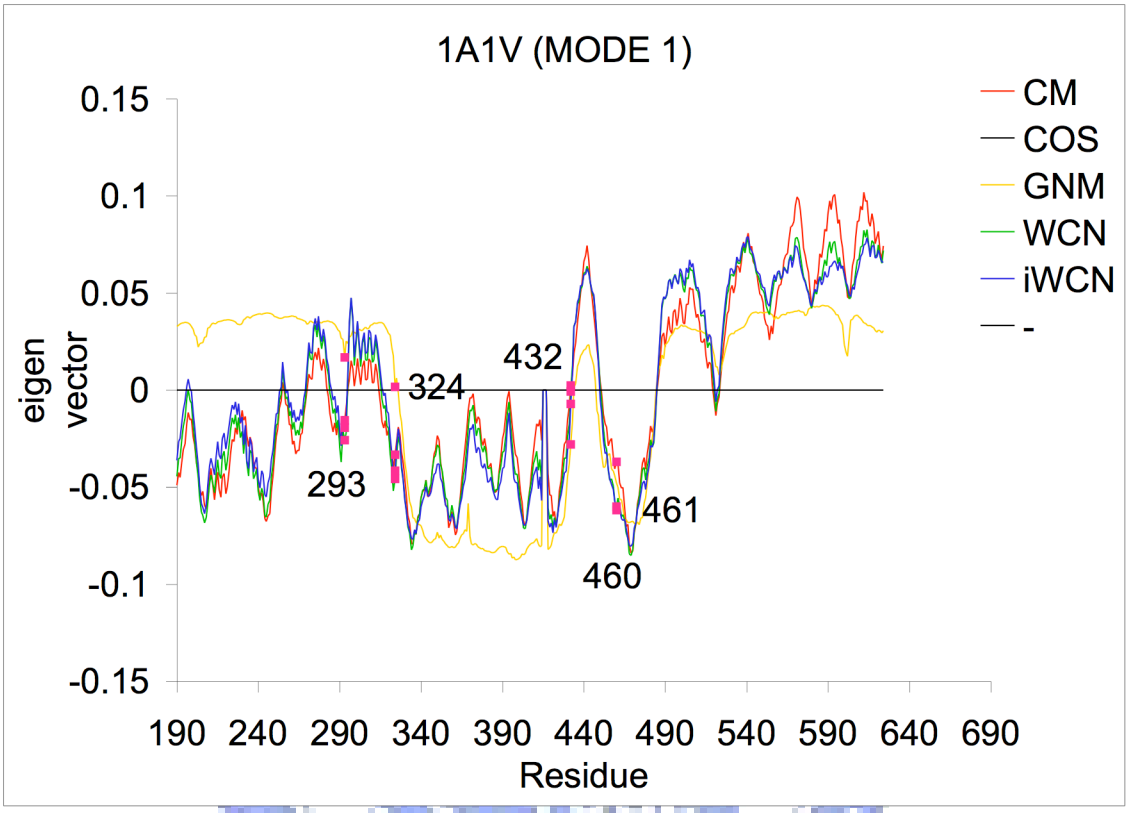

1A1V(iWCN)
eigen value
30
25
20
15
10
5
0
1
11
21
31
41
Mode

Figure 3 : The eigen vector of CM, COS, GNM, WCN, iWCN, which are non-normalized. Red squares are hot spot residues.

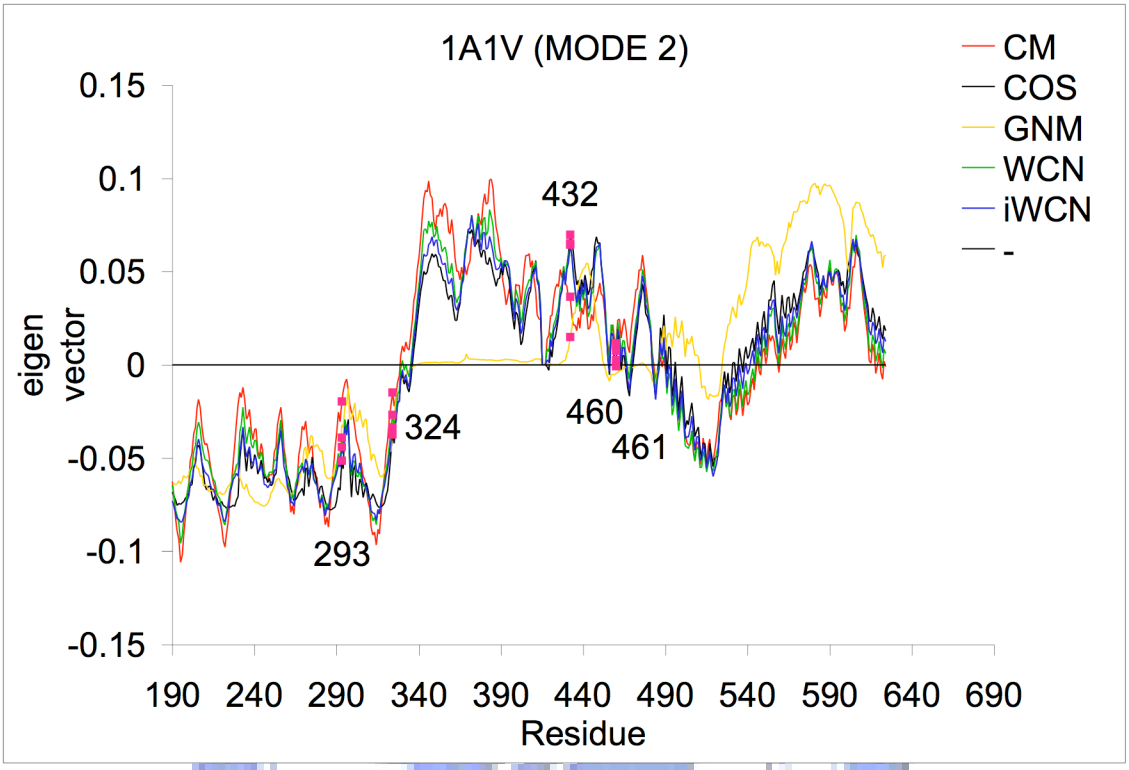


Correl	CM	COS	GNM	WCN	iWCN
CM		0.922	0.589	0.976	0.959
COS	0.922		0.747	0.978	0.992
GNM	0.589	0.747		0.650	0.717
WCN	0.976	0.978	0.650		0.994
iWCN	0.959	0.992	0.717	0.994	

Correl : correlation coefficient

residue No.	CM	COS	GNM	WCN	iWCN
293	-0.016	-0.022	0.017	-0.026	-0.019
324	-0.033	-0.044	0.002	-0.046	-0.042
432	0.002	-0.014	-0.028	-0.001	-0.007
460	-0.037	-0.070	-0.048	-0.060	-0.062
461	-0.038	-0.070	-0.050	-0.056	-0.059

residue No.: hot spot residue number.

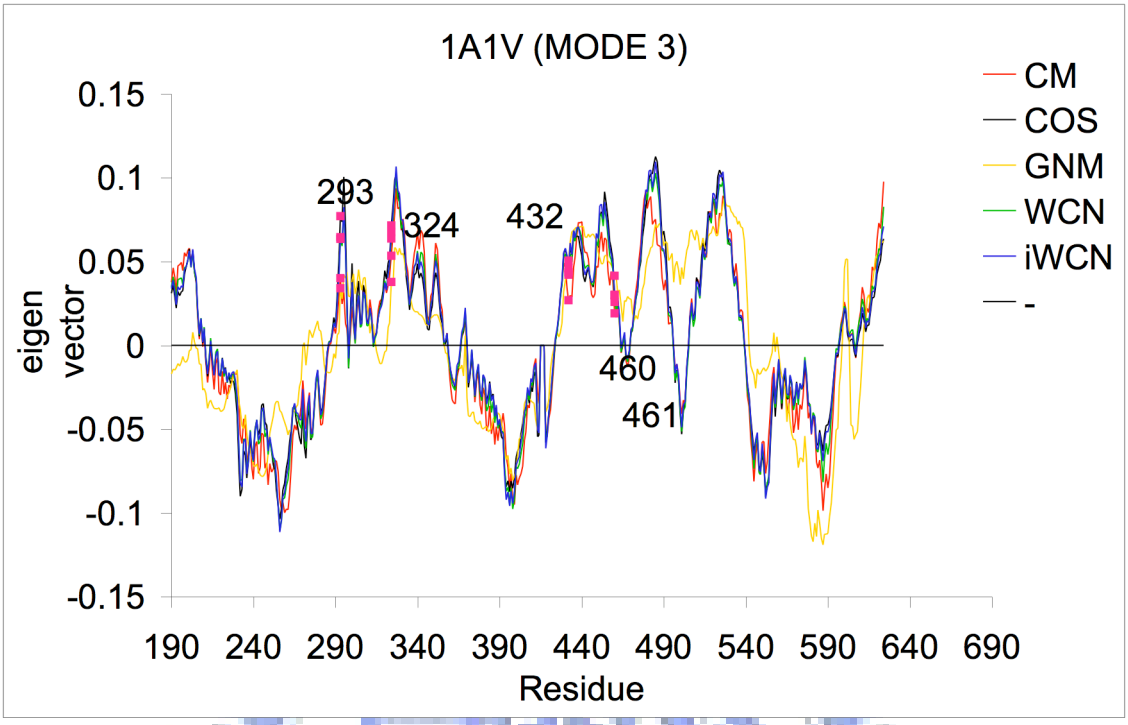


Correl	CM	COS	GNM	WCN	iWCN
CM		0.918	0.600	0.974	0.955
COS	0.918		0.785	0.976	0.991
GNM	0.600	0.785		0.679	0.740
WCN	0.974	0.976	0.679		0.994
iWCN	0.955	0.991	0.740	0.994	

Correl : correlation coefficient

residue No.	CM	COS	GNM	WCN	iWCN
293	-0.020	-0.051	-0.032	-0.039	-0.044
324	-0.015	-0.037	-0.028	-0.027	-0.034
432	0.037	0.070	0.015	0.065	0.065
460	0.012	-0.001	-0.005	0.008	0.003
461	0.024	0.018	-0.004	0.023	0.019

residue No.: hot spot residue number.



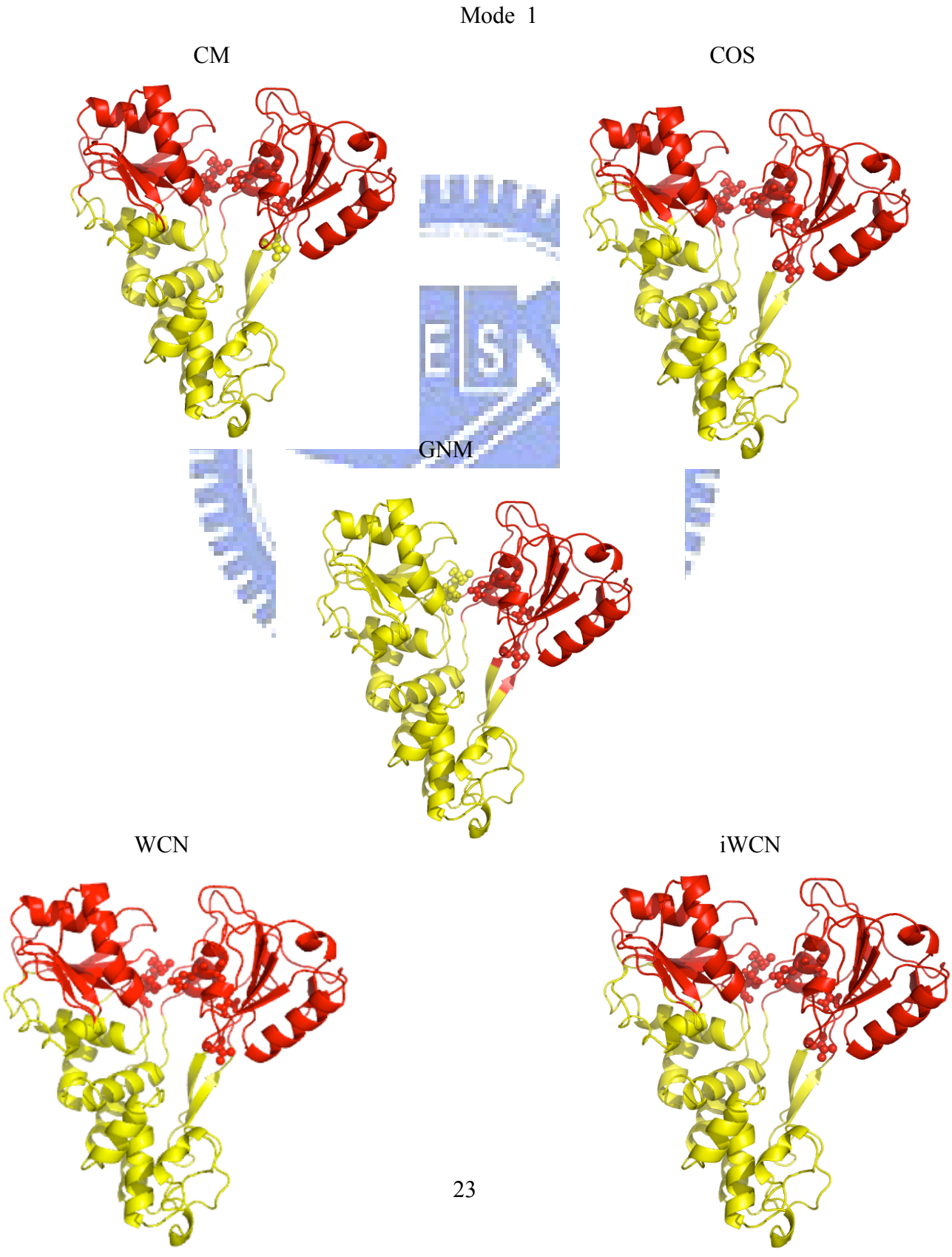
Correl	CM	COS	GNM	WCN	iWCN
CM		0.955	0.797	0.979	0.971
COS	0.955		0.771	0.993	0.996
GNM	0.797	0.771		0.774	0.779
WCN	0.979	0.993	0.774		0.998
iWCN	0.971	0.996	0.779	0.998	

Correl : correlation coefficient

residue No.	CM	COS	GNM	WCN	iWCN
293	0.040	0.077	0.034	0.064	0.065
324	0.054	0.072	0.038	0.064	0.068
432	0.027	0.044	0.051	0.042	0.046
460	0.019	0.030	0.042	0.026	0.030
461	0.017	0.023	0.041	0.020	0.024

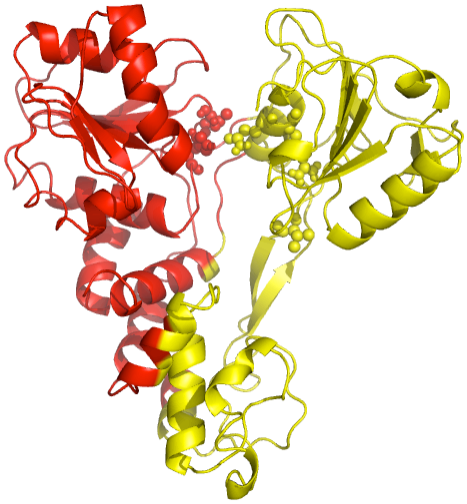
residue No.: hot spot residue number.

Figure 4 : The direction of eigenvalue of CM, COS, GNM, WCN, iWCN , which are non-normalize . Red and yellow color represent contrarious direction of fluctuation .

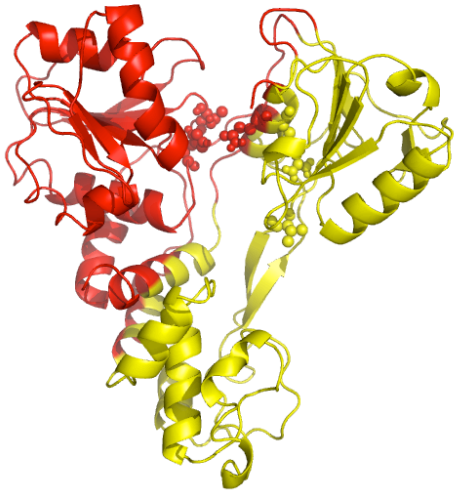


Mode 2

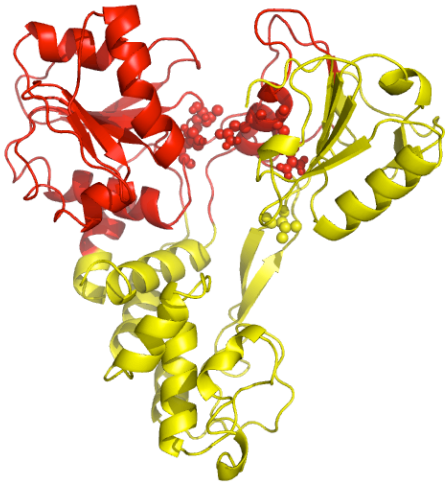
CM



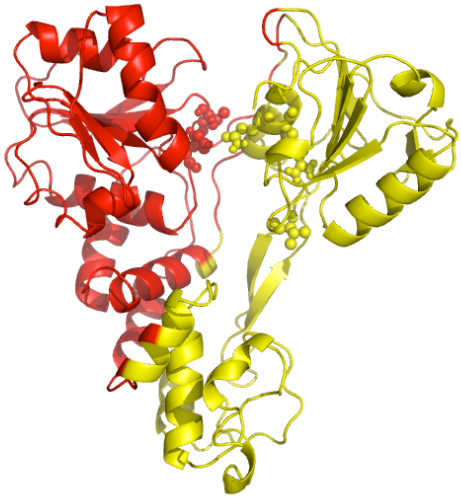
COS



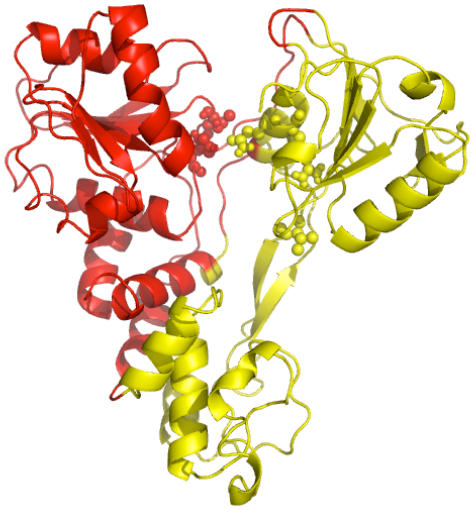
GNM



WCN

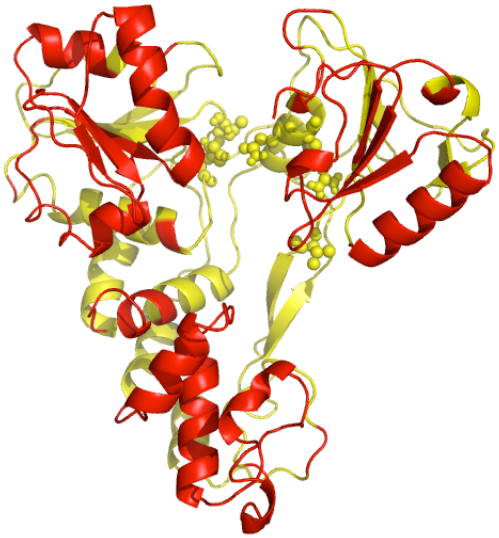


iWCN

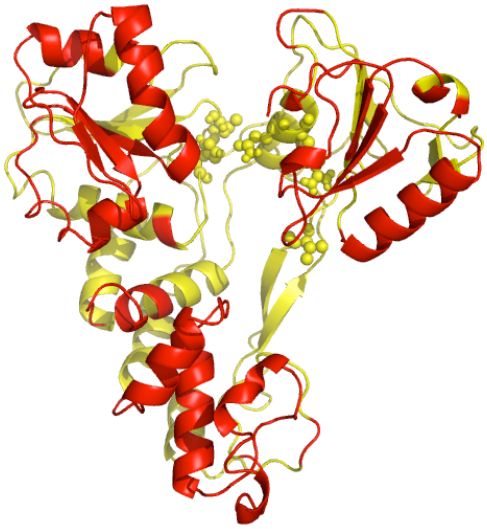


Mode 3

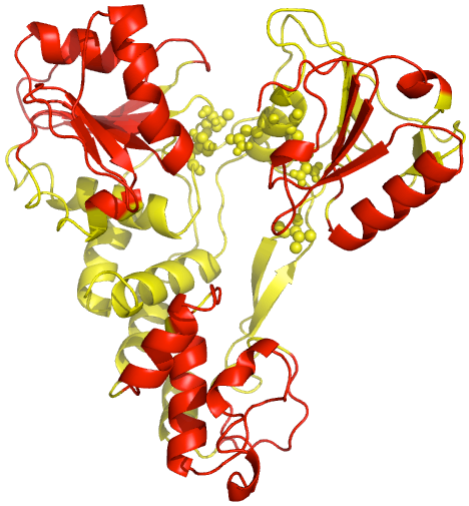
CM



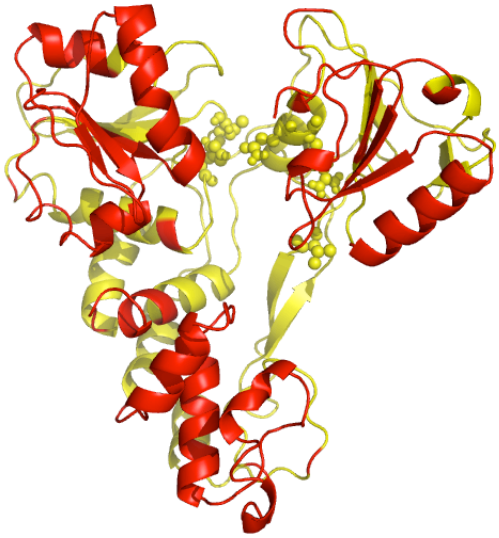
COS



GNM



WCN



iWCN

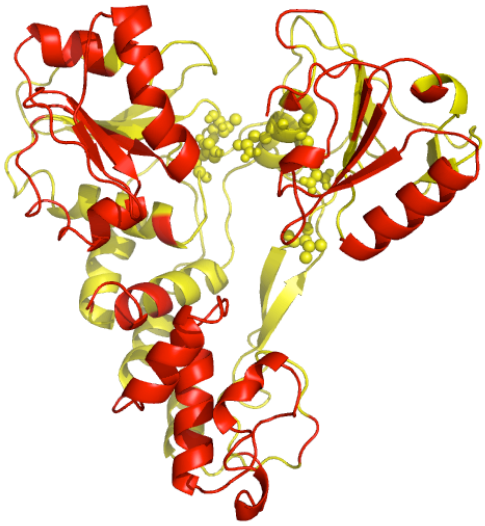
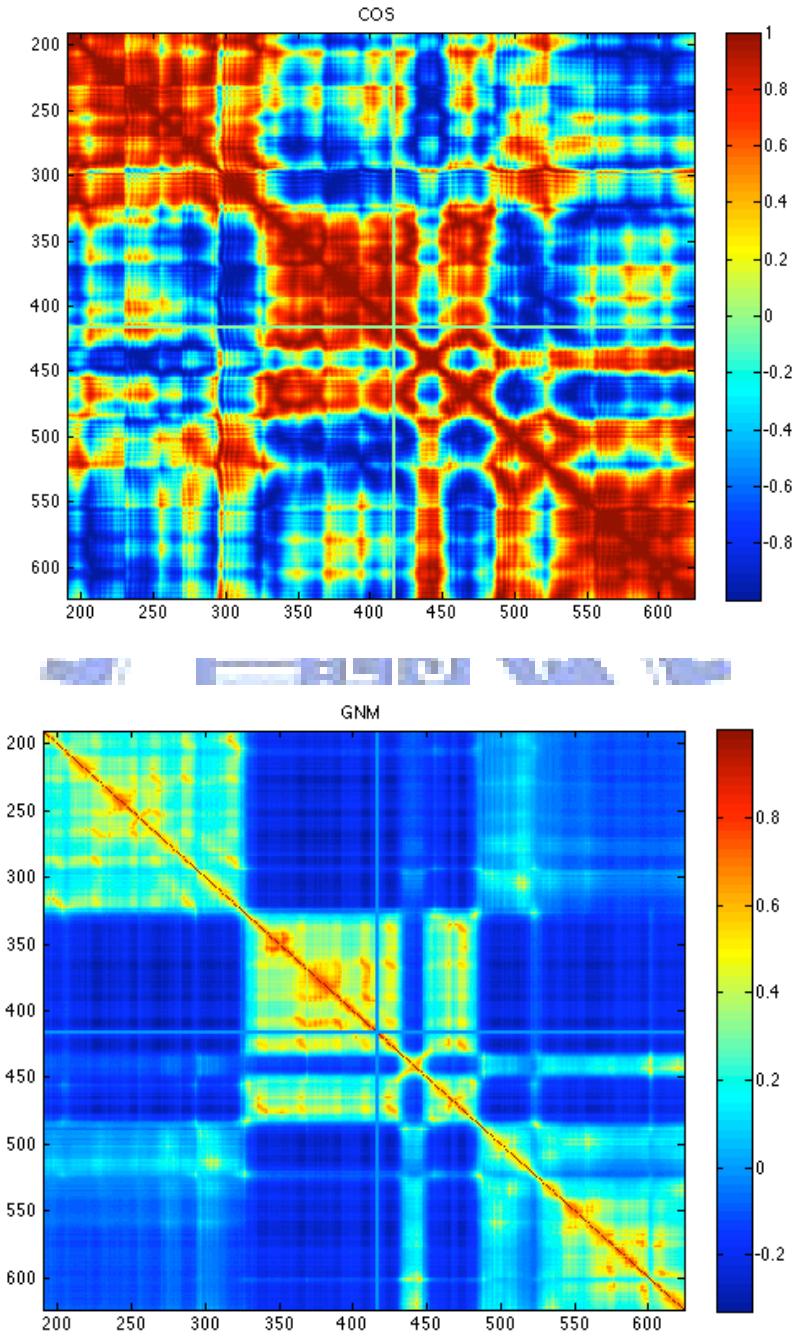
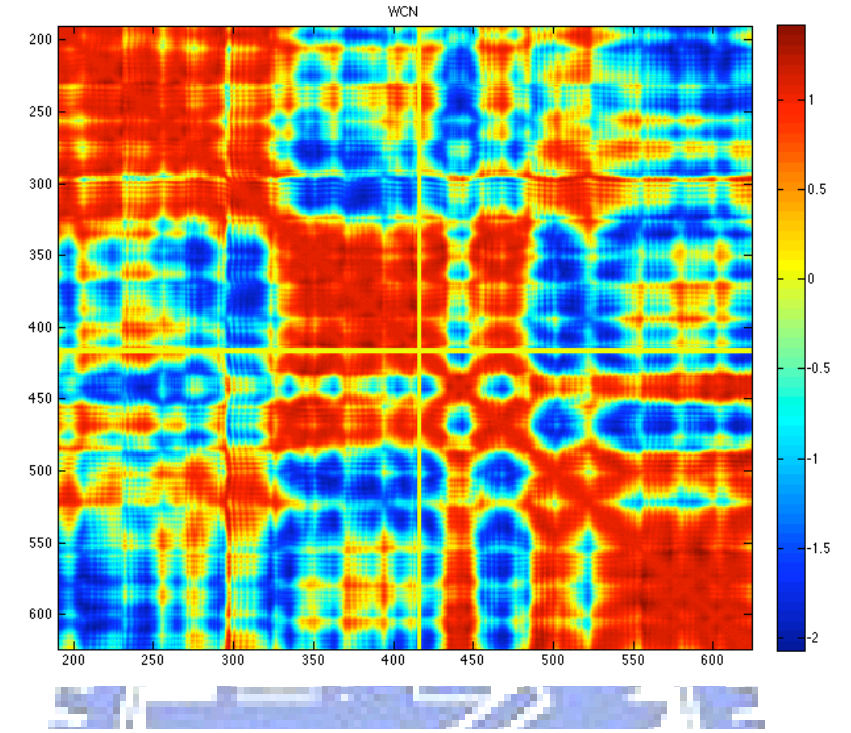


Figure 5 : The correlation map of COS, GNM, WCN, iWCN, which are normalized.




WCN
200
250
300
350
400
450
500
550
600
1
0.5
0
-0.5
-1
-1.5
-2

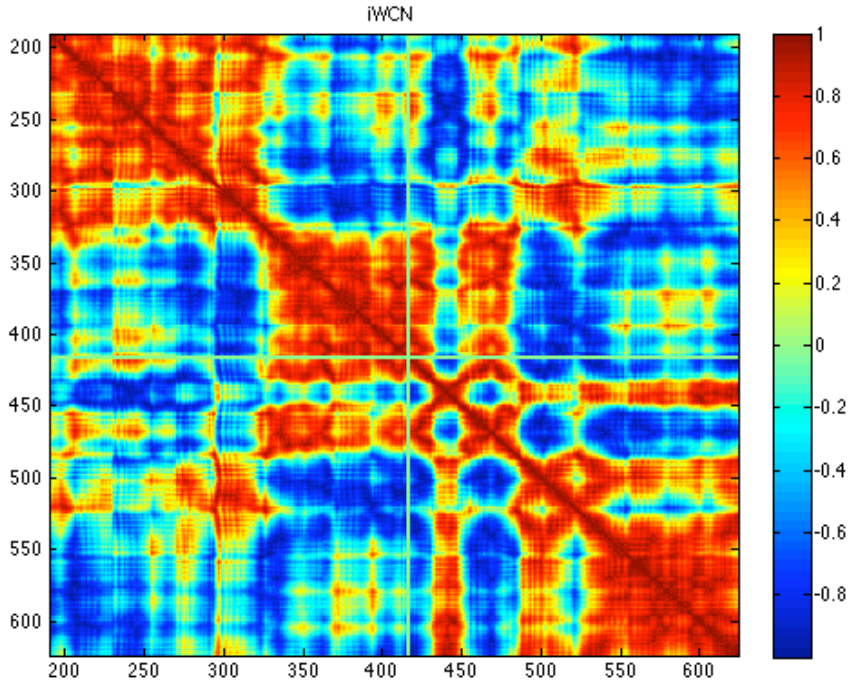
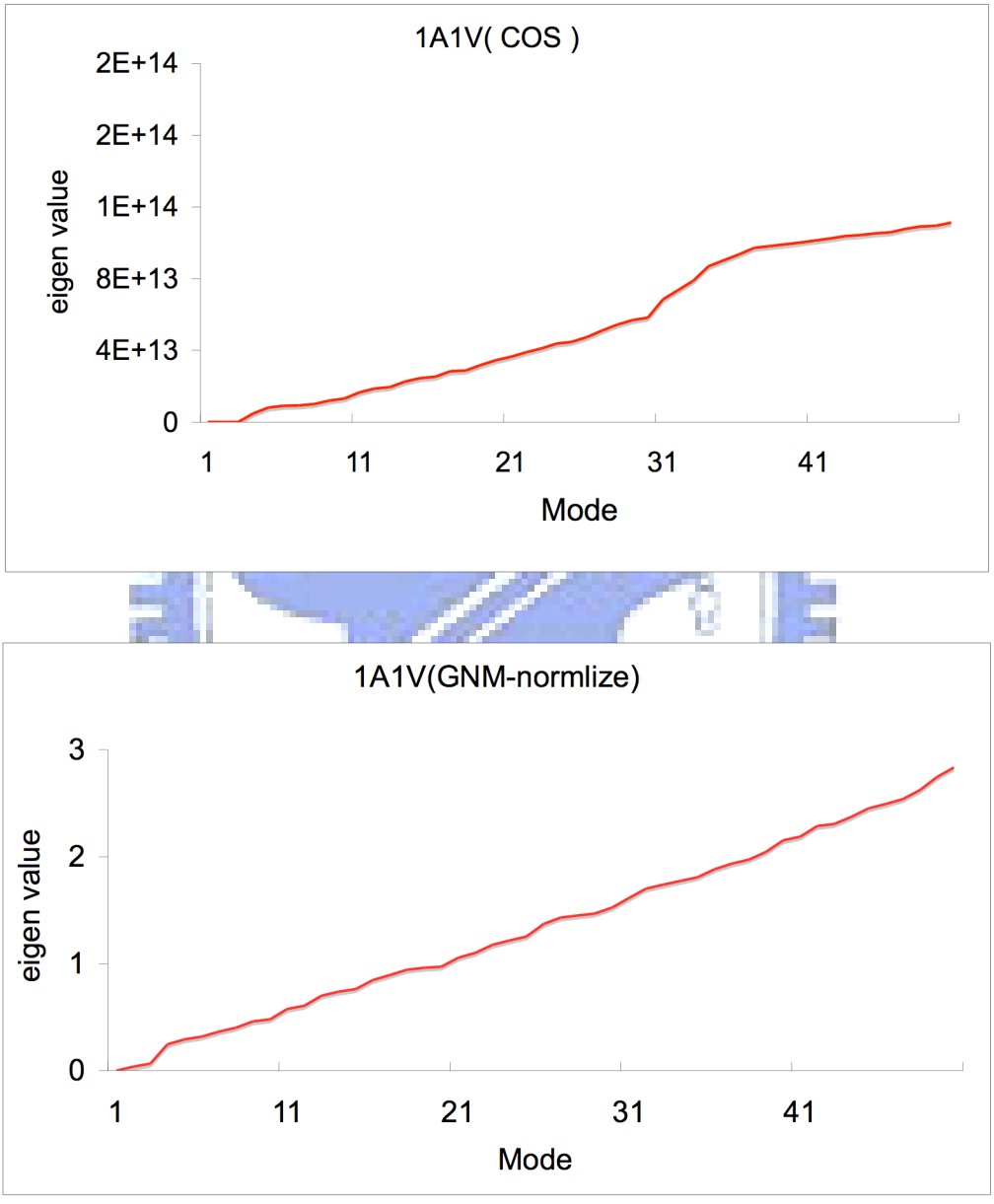
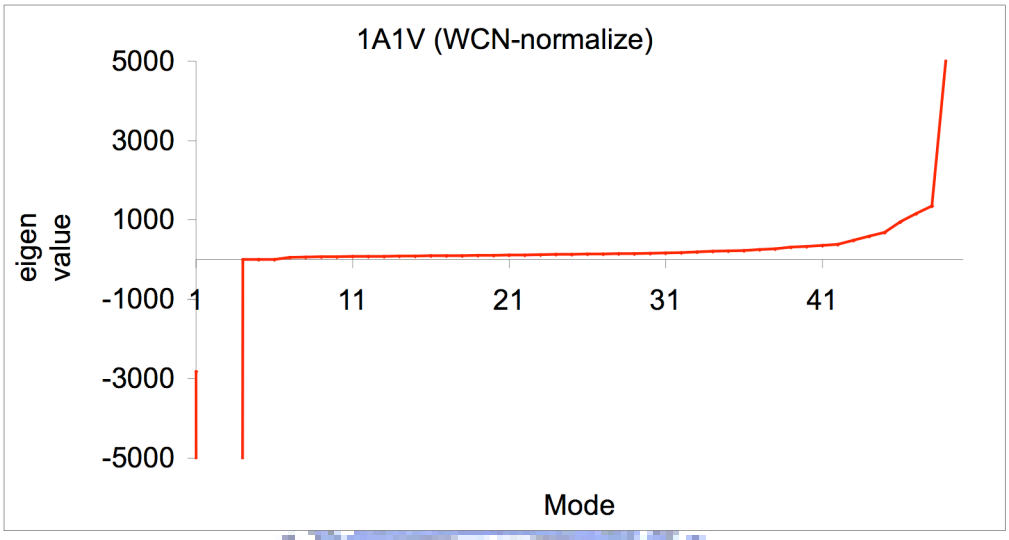

iWCN
200
250
300
350
400
450
500
550
600
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8

Figure 6 : The eigenvalue of COS, GNM, WCN, iWCN, which are normalized.




1A1V (WCN-normalize)
5000
3000
1000
-1000
-3000
-5000
eigen value
1
11
21
31
41
Mode

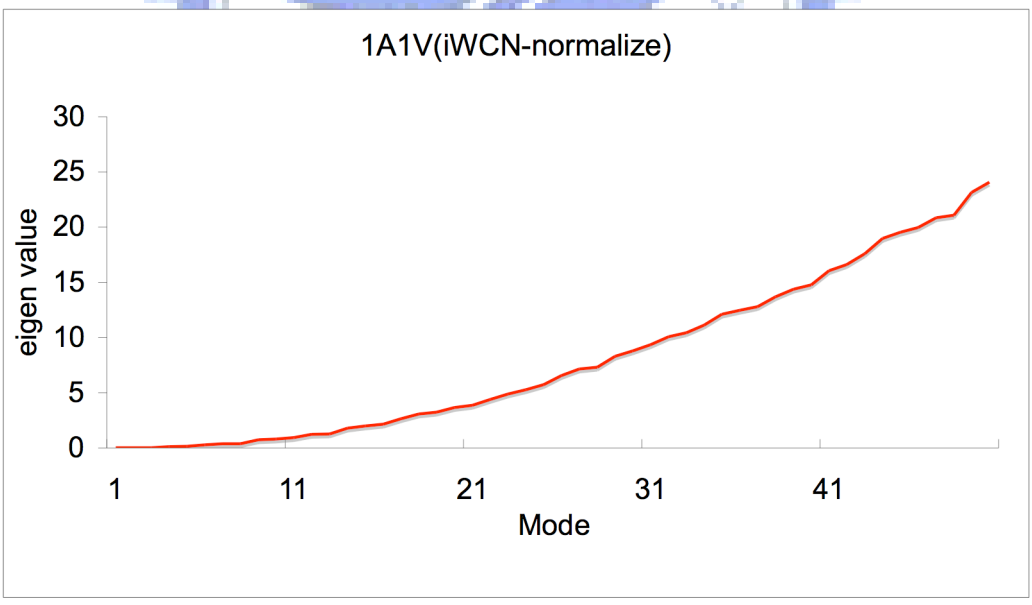
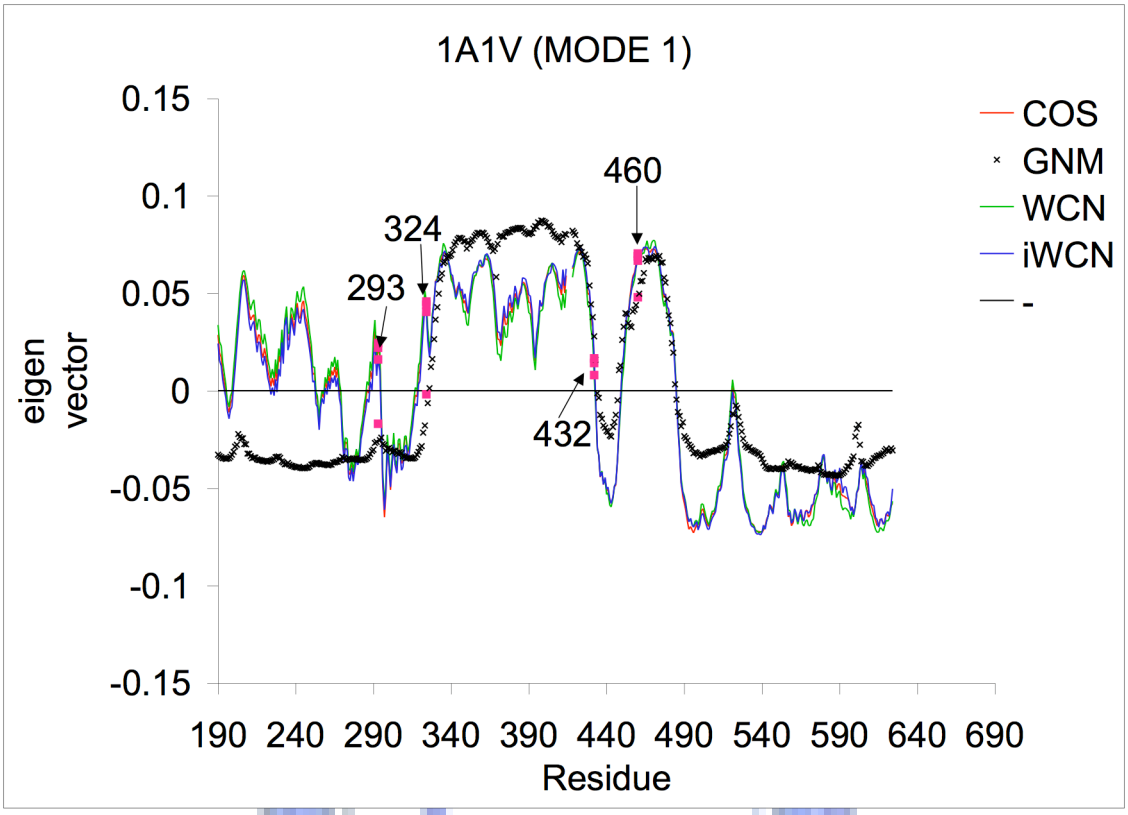

1A1V(iWCN-normalize)
30
25
20
15
10
5
0
eigen value
1
11
21
31
41
Mode

Figure 7 : The eigenvector of COS, GNM, WCN, iWCN, which are normalized. Red squares are hot spot residues.

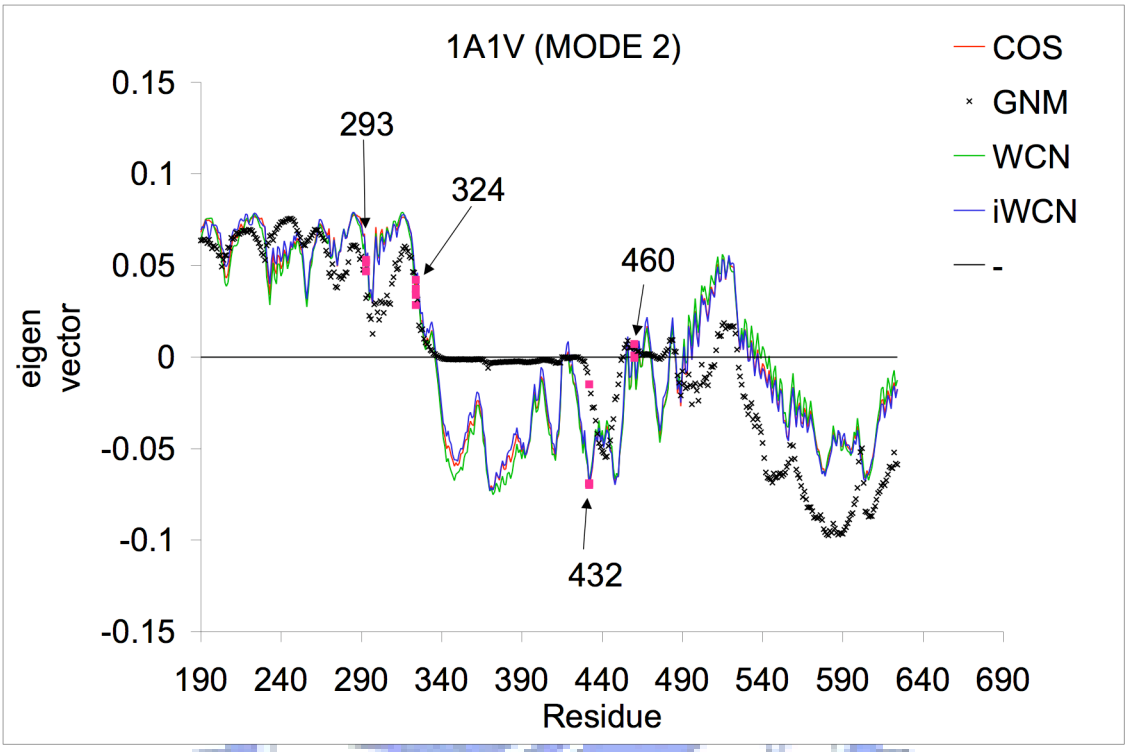


Correl	COS	GNM	WCN	iWCN
COS		0.747	0.997	0.998
GNM	0.747		0.709	0.774
WCN	0.997	0.709		0.992
iWCN	0.998	0.774	0.992	

Correl : correlation coefficient

residue No.	CM	GNM	WCN	iWCN
293	0.022	-0.017	0.025	0.016
324	0.044	-0.002	0.046	0.041
432	0.014	0.028	0.008	0.017
460	0.070	0.048	0.067	0.069
461	0.070	0.050	0.065	0.069

residue No.: hot spot residue number.

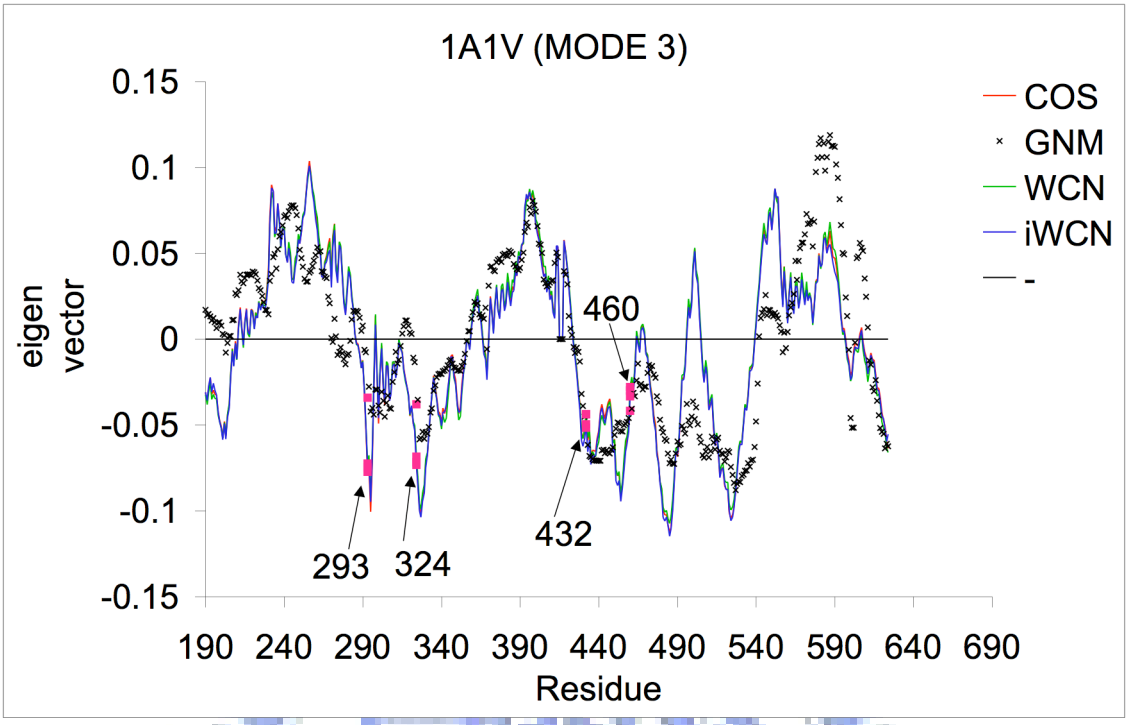


Correl	COS	GNM	WCN	iWCN
COS		0.785	0.996	0.998
GNM	0.785		0.748	0.807
WCN	0.996	0.748		0.992
iWCN	0.998	0.807	0.992	

Correl : correlation coefficient

residue No.	CM	GNM	WCN	iWCN
293	0.051	0.032	0.047	0.053
324	0.037	0.028	0.034	0.042
432	-0.070	-0.015	-0.069	-0.069
460	0.001	0.005	0.000	0.007
461	-0.018	0.004	-0.018	-0.012

residue No.: hot spot residue number.



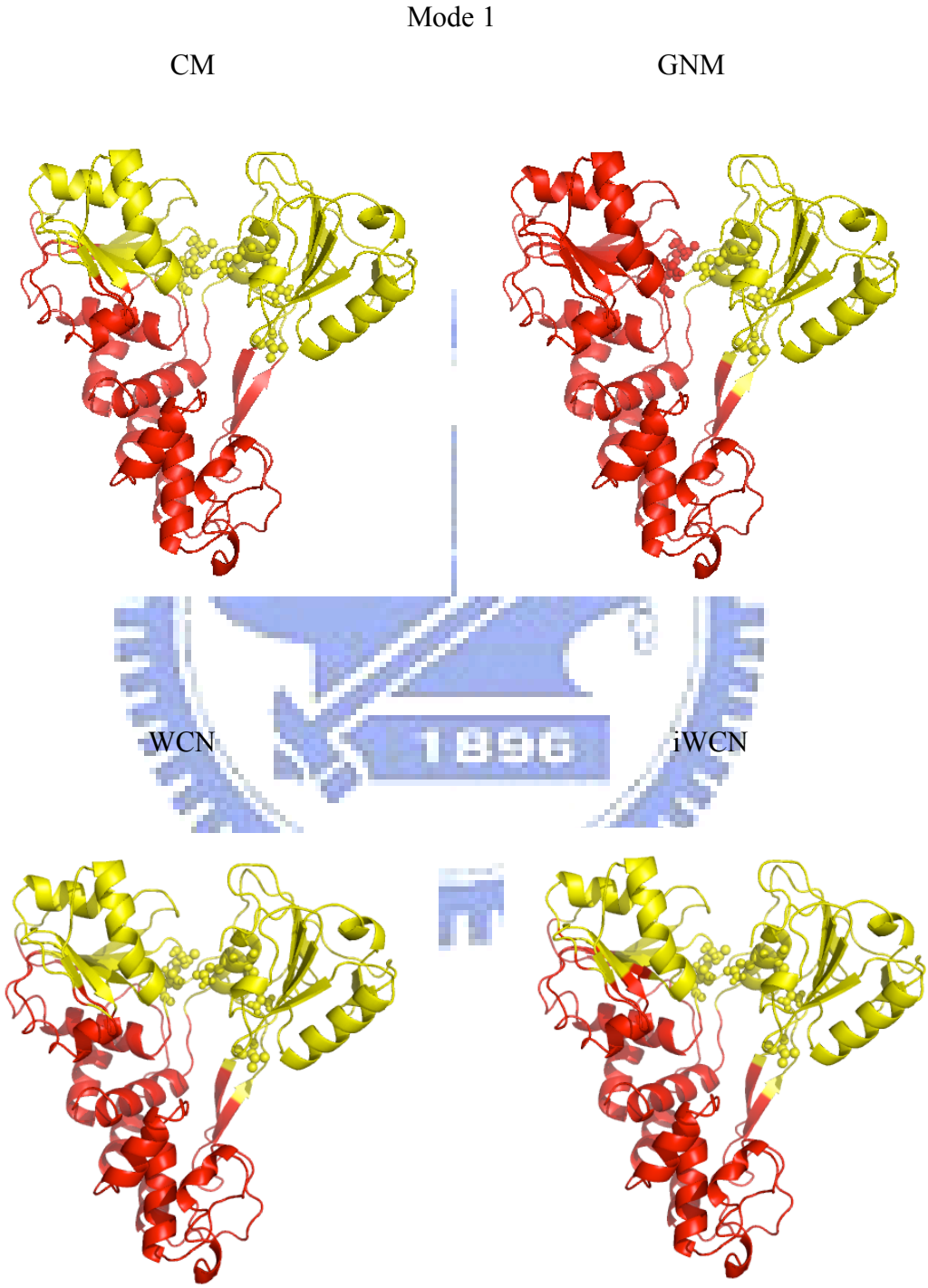
Correl	COS	GNM	WCN	iWCN
COS		0.771	0.999	0.999
GNM	0.771		0.771	0.772
WCN	0.999	0.771		0.998
iWCN	0.999	0.772	0.998	

Correl : correlation coefficient

residue No.	CM	GNM	WCN	iWCN
293	-0.077	-0.034	-0.072	-0.075
324	-0.072	-0.038	-0.068	-0.073
432	-0.044	-0.052	-0.044	-0.050
460	-0.030	-0.042	-0.028	-0.033
461	-0.023	-0.041	-0.022	-0.026

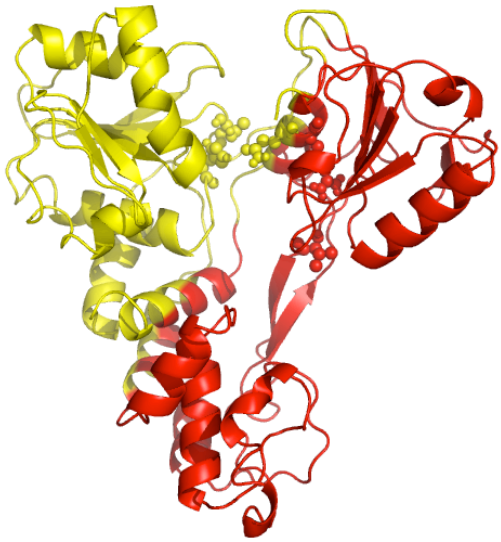
residue No.: hot spot residue number.

Figure 8 : The mode of eigen vector of COS, GNM, WCN, iWCN, which are normalized.

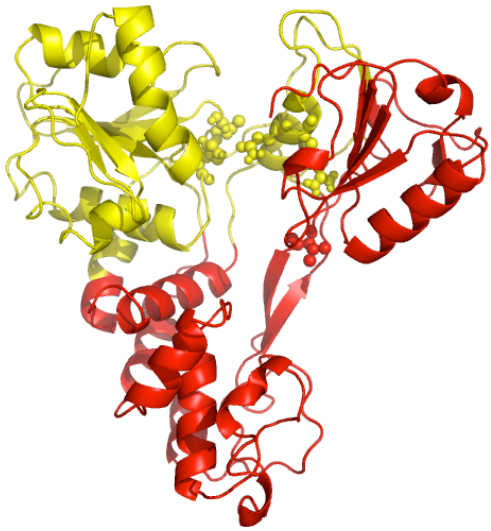


Mode 2

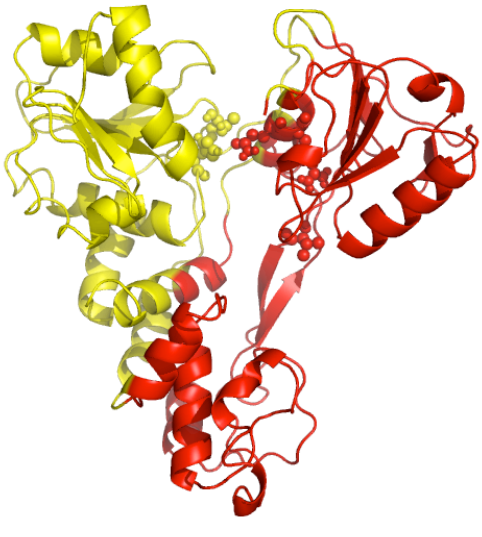
CM



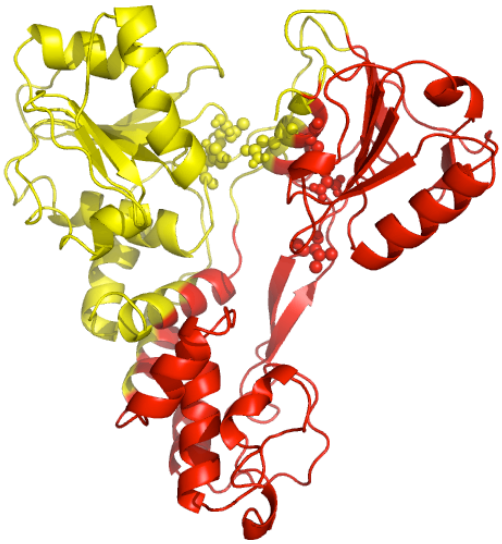
GNM



WCN

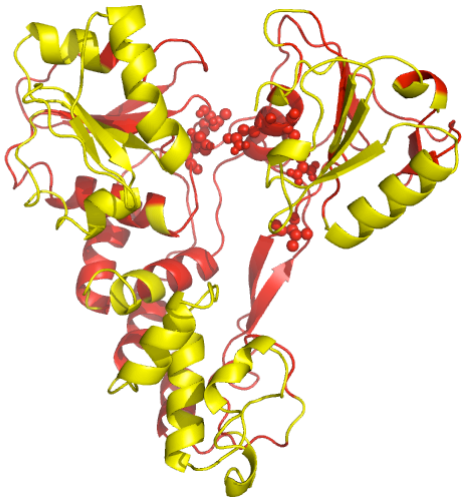


iWCN

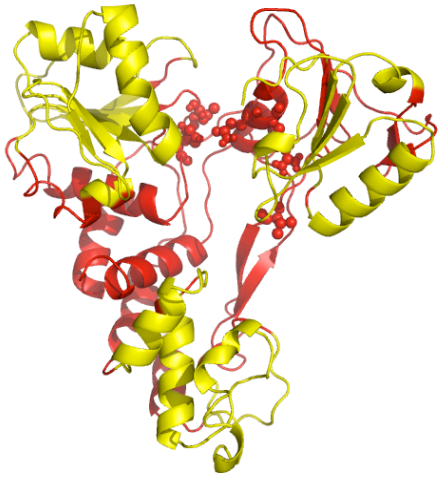


Mode 3

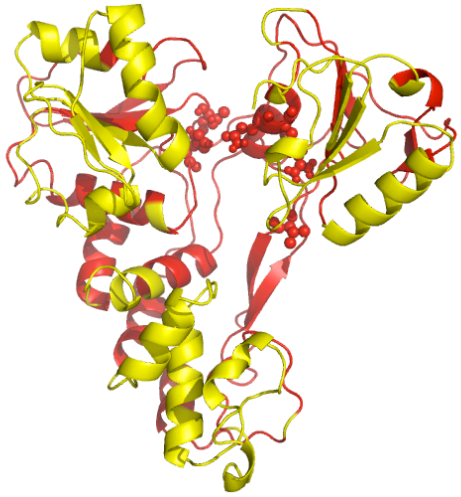
CM



GNM



WCN



iWCN

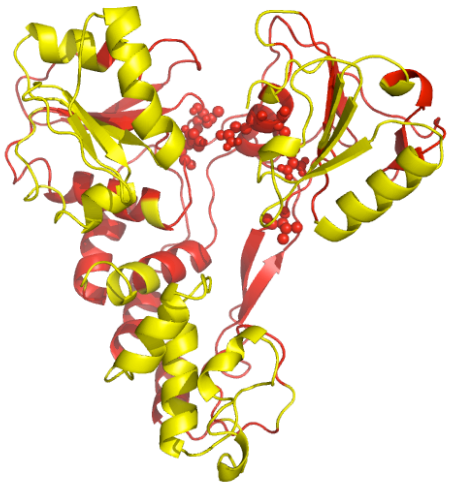


Figure 9 : Blue color denotes domain1 , green color denotes domain 2, red color denotes domain 3, magenta sphere denotes the center of CM, yellow sphere denotes the new center.

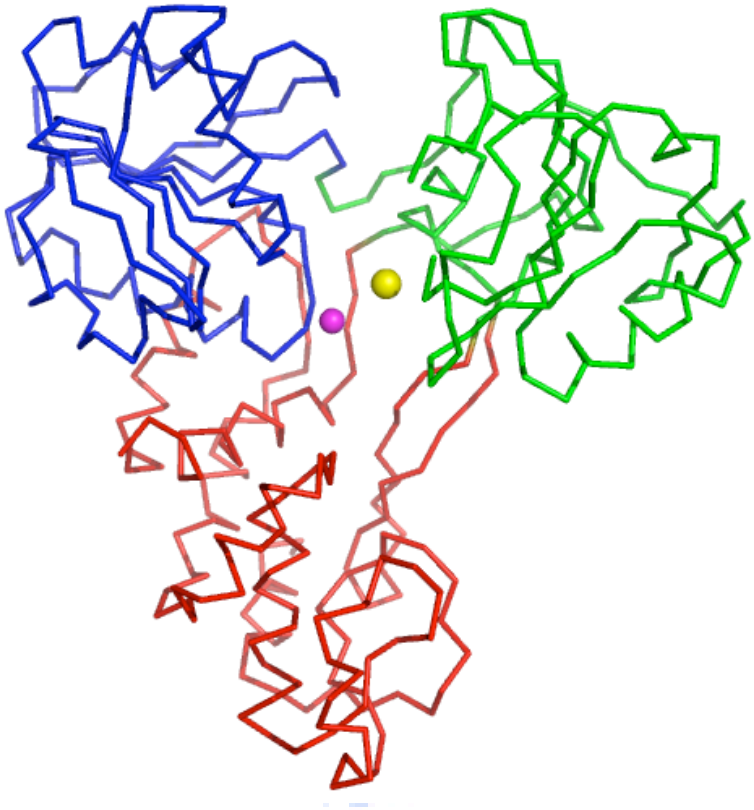
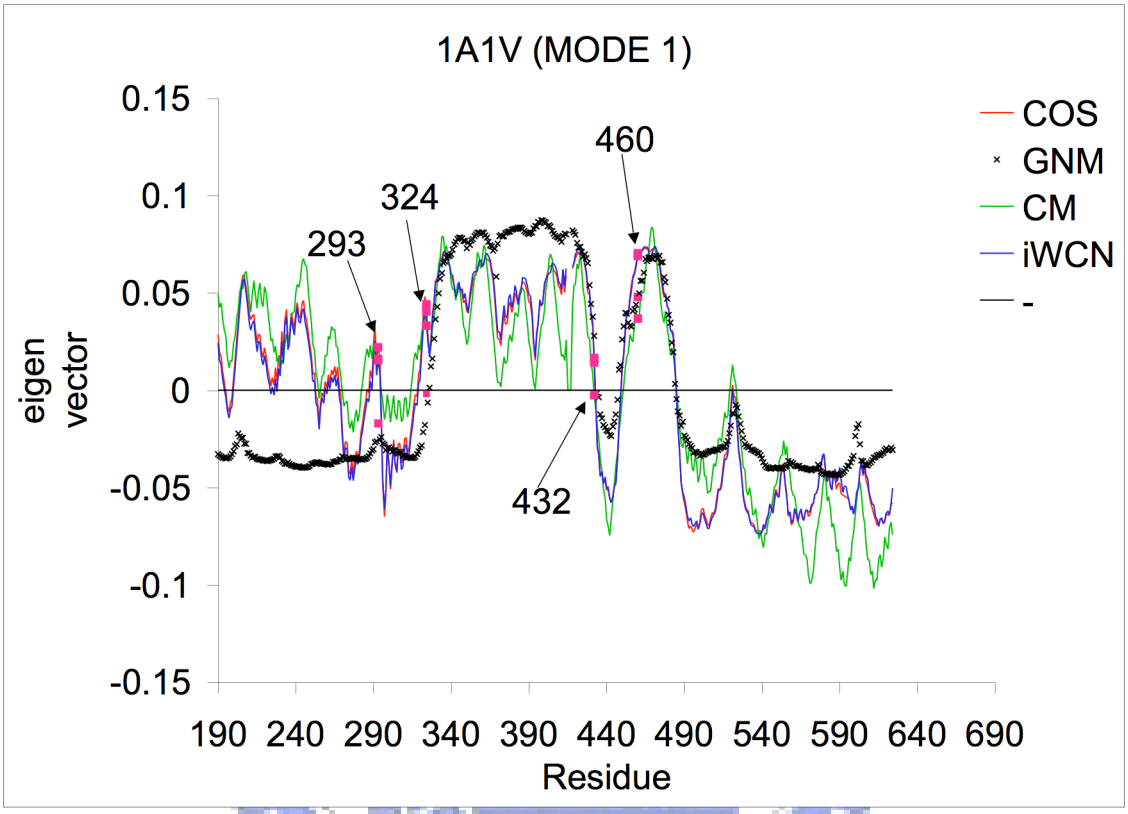


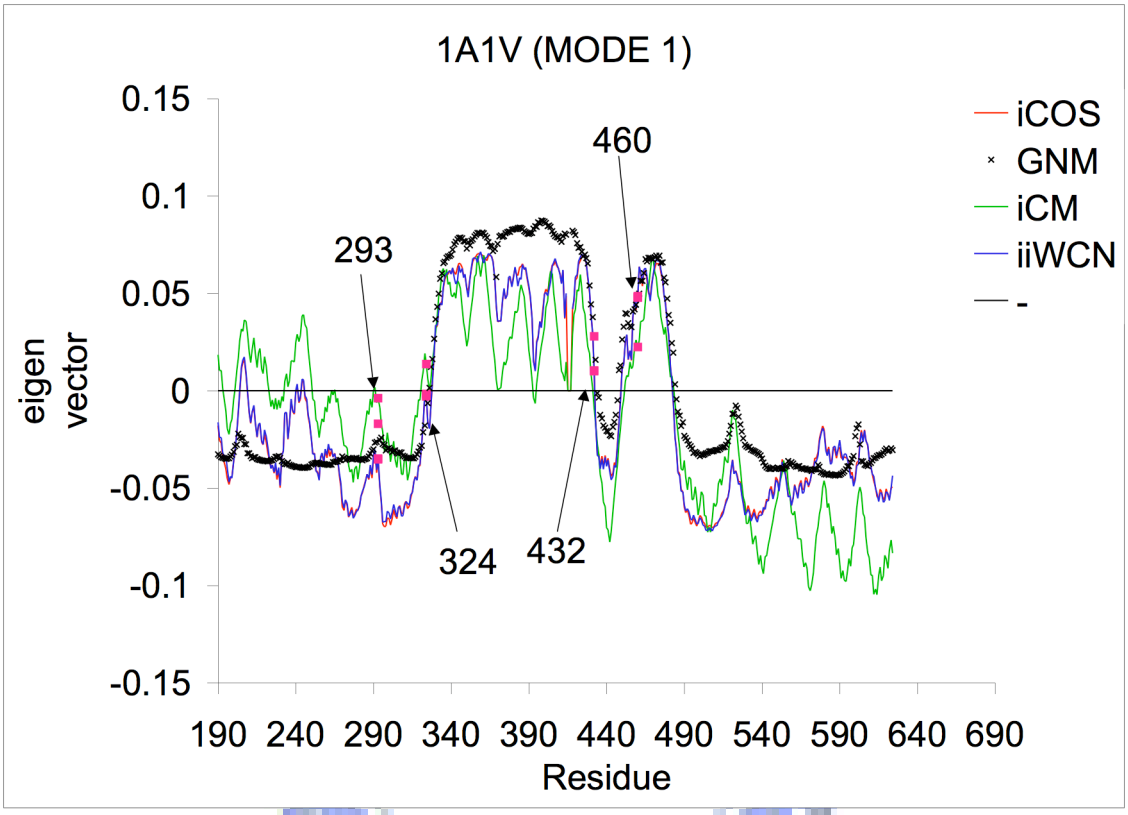
Figure 10 : The comparison of each method before improvement.



Correl	COS	GNM	CM	iWCN
COS		0.747	0.922	0.998
GNM	0.747		0.589	0.774
CM	0.922	0.589		0.908
iWCN	0.998	0.774	0.908	

Correl : correlation coefficient

Figure 11 : The comparison of each method after improvement.



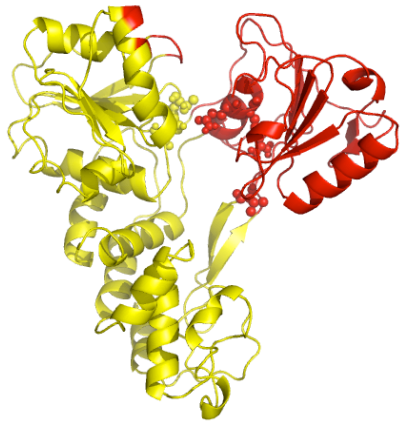
iCOS, iCM, iiWCN : improvement of COS, CM and iWCN.

Correl	iCOS	GNM	iCM	iiWCN
iCOS		0.919	0.821	0.999
GNM	0.919		0.725	0.916
iCM	0.821	0.725		0.825
iiWCN	0.999	0.916	0.825	

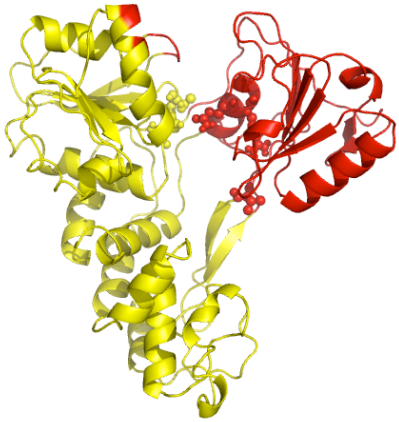
Correl : correlation coefficient

Figure 12 : The mode 1 comparison of each method after improvement.

COS



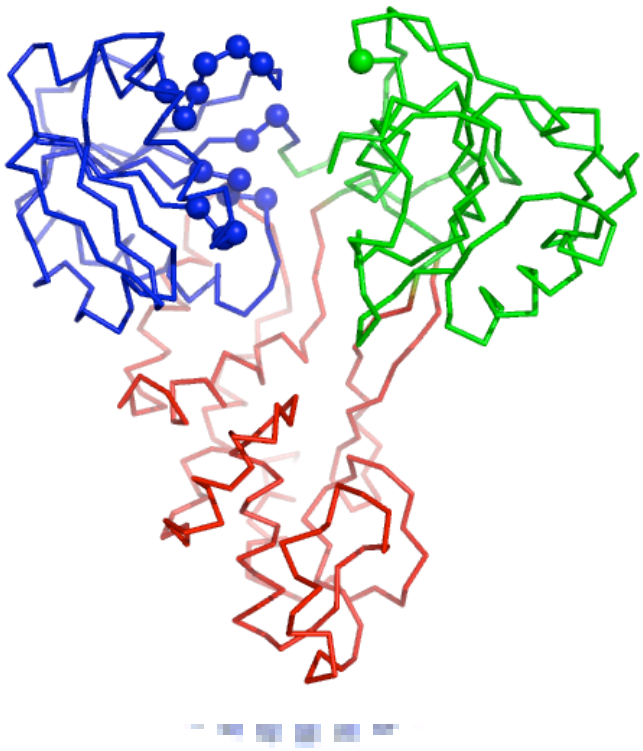
iWCN



CM

Figure 13 : A. ATP binding site in protein of PDB code 1A1V . B. High conservation residues in 1A1V . C. The high correlation with ATP binding sites in COS . D. The high correlation with ATP binding sites in PCN. Here we only show α carbon spheres. Domain 1 : blue color. Domain 2 : green color. Domain 3 : red color.

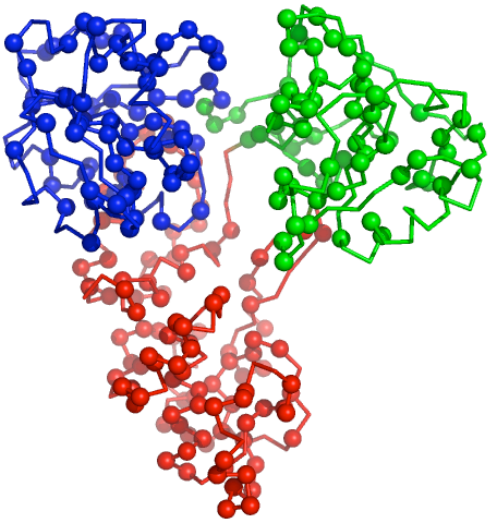
A.



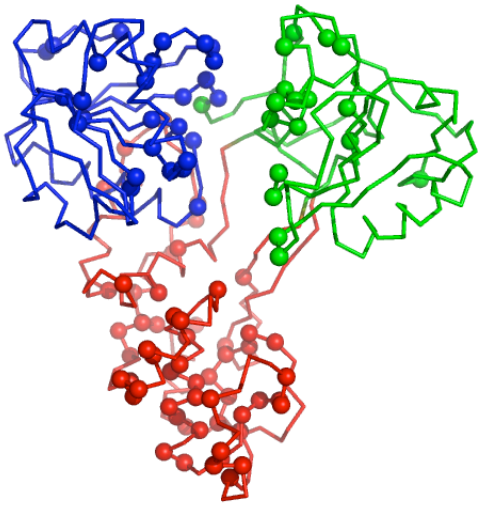
The residues number of ATP binding site : 207-212, 229-231, 290-293, 322, 323, 467

B.

conservation > 7

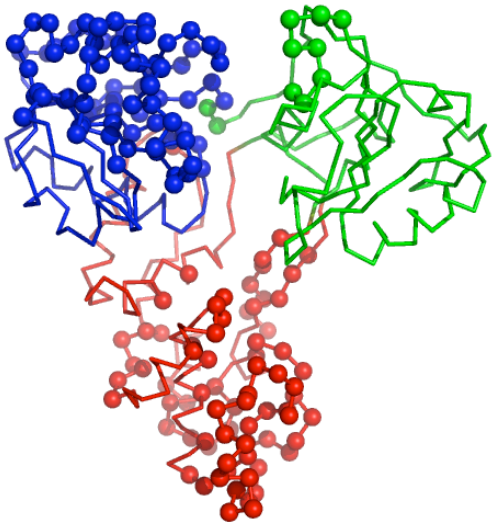


conservation > 9

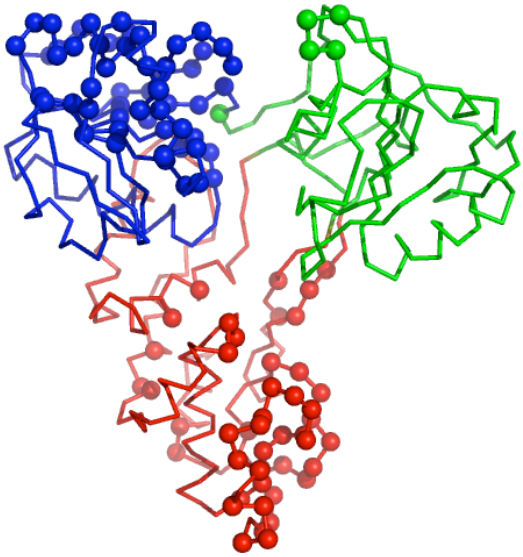


C.

COS top 5%

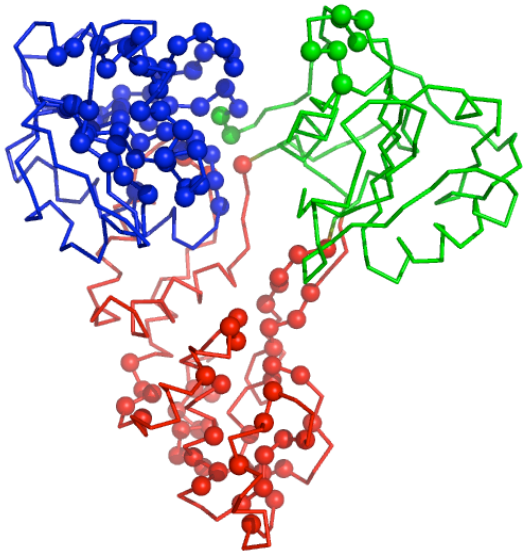


COS top 2%



D.

iWCN top 5%



iWCN top 2%

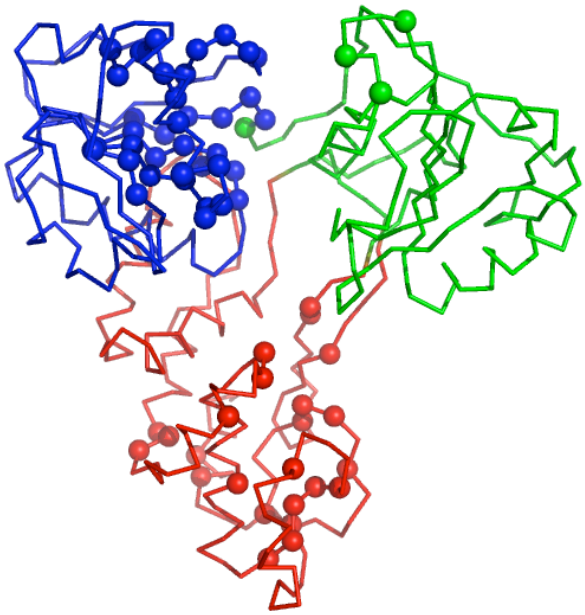
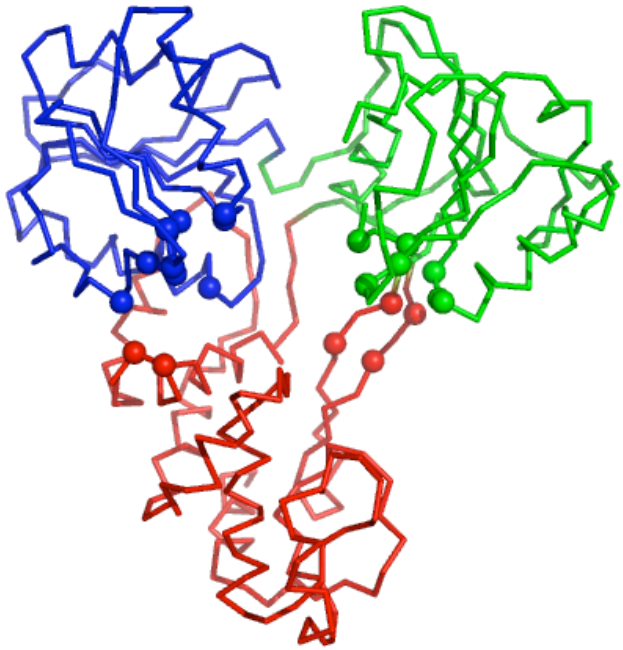


Figure 14 : A. DNA binding site in protein of PDB code 1A1V . B. The high correlation with DNA binding sites in COS . C. The high correlation with DNA binding sites in iWCN. Here we only show α carbon spheres. Domain 1 : blue color. Domain 2 : green color. Domain 3 : red color.

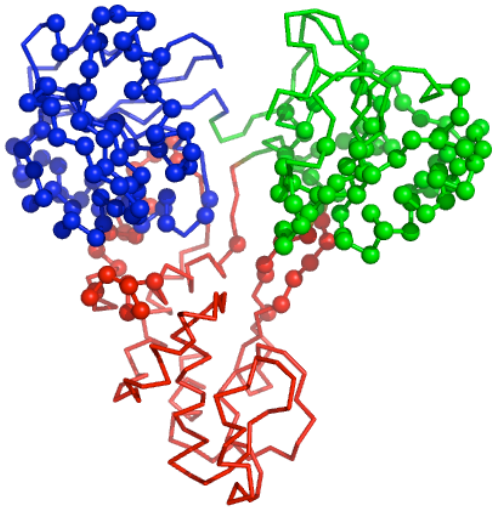
A.



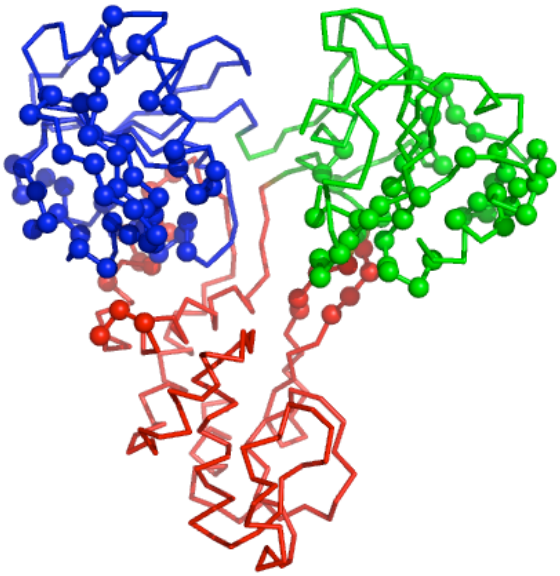
The residues number of DNA binding site : 230, 232, 254, 255, 269, 271, 272, 275, 298, 369-371, 392, 393, 411, 413, 432, 434, 448, 450, 501, 502.

B.

COS top 5%

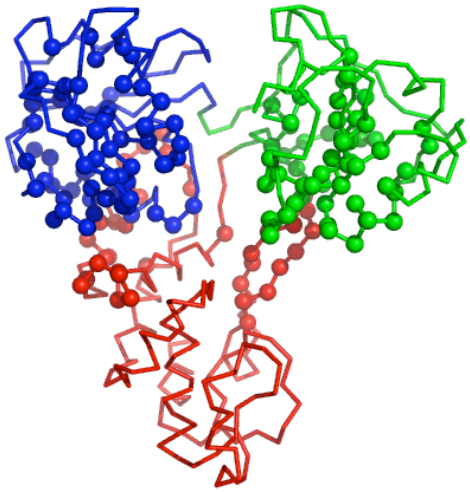


COS top 2%



C.

iWCN top 5%



iWCN top 2%

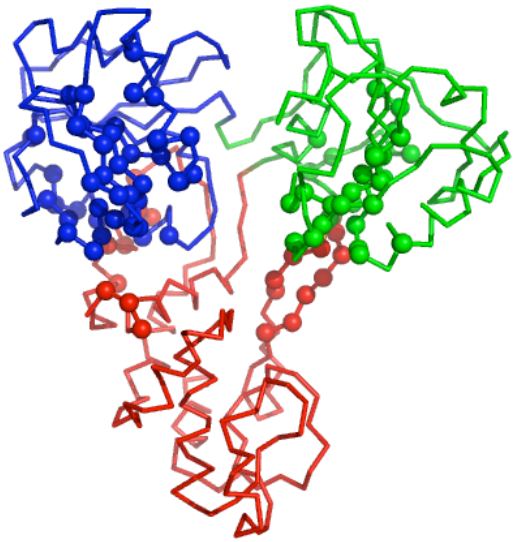
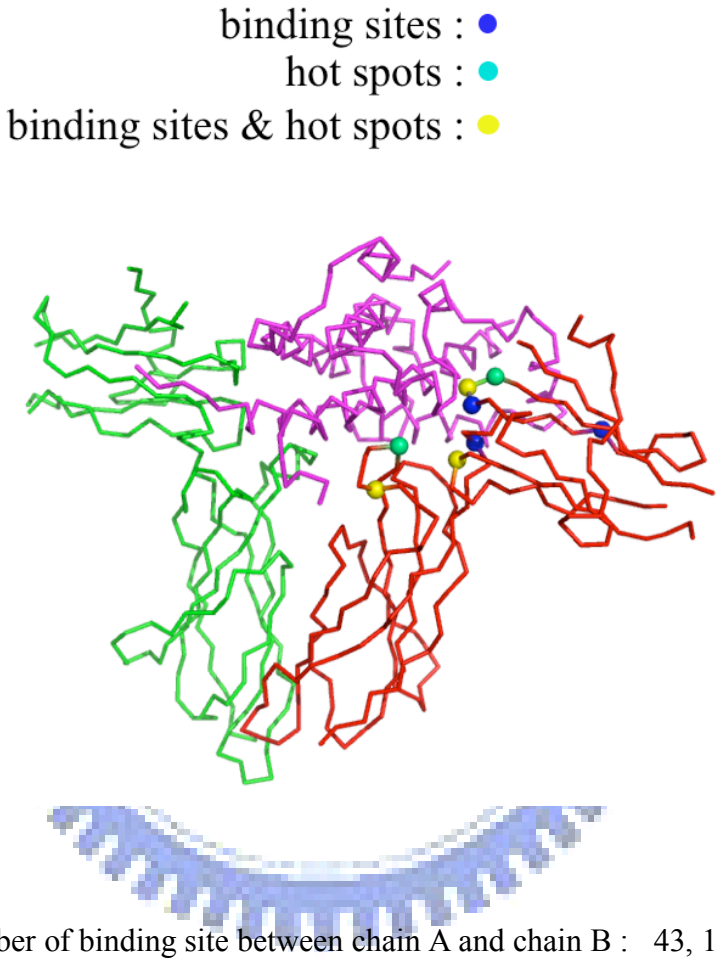


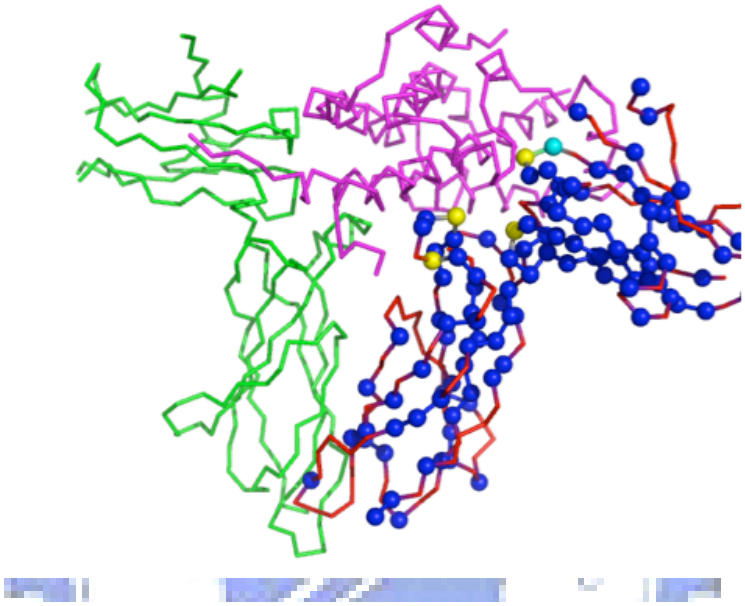
Figure 15 : A. Human hormone receptor (PDB code 3HHR) binding site in chain b and its highly conserved residues . B. iWCN with high correlation . C. COS with high correlation .

A.

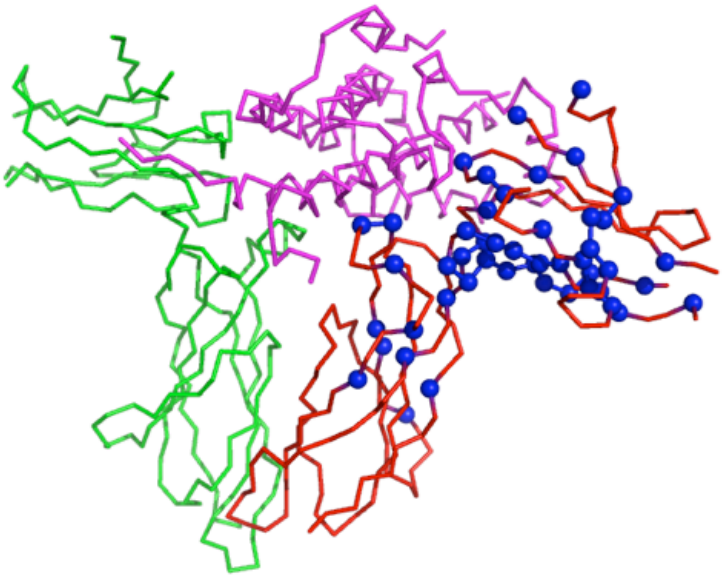


The residues number of binding site between chain A and chain B : 43, 103, 104, 120, 127, 165.

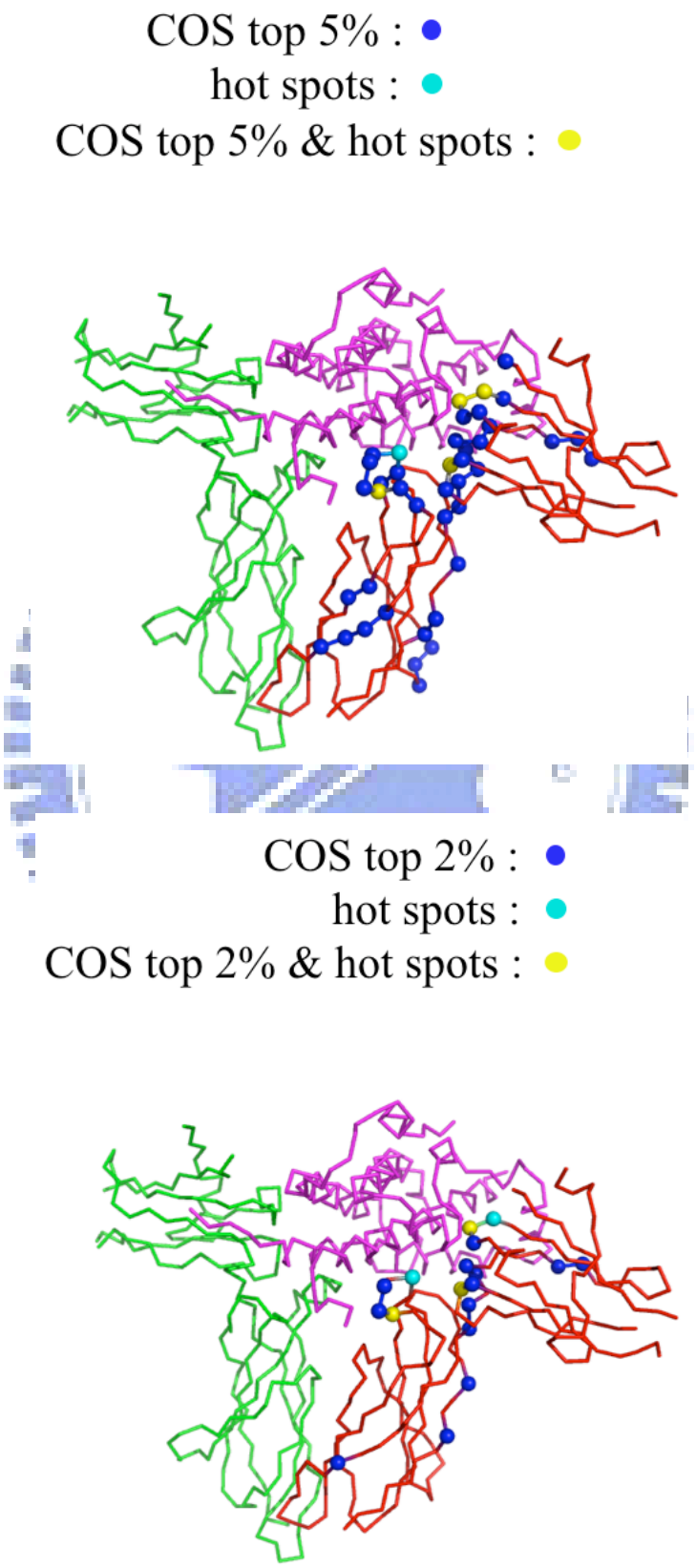
conservation > 7 : ●
hot spots : ●
conservation > 7 & hot spots : ●



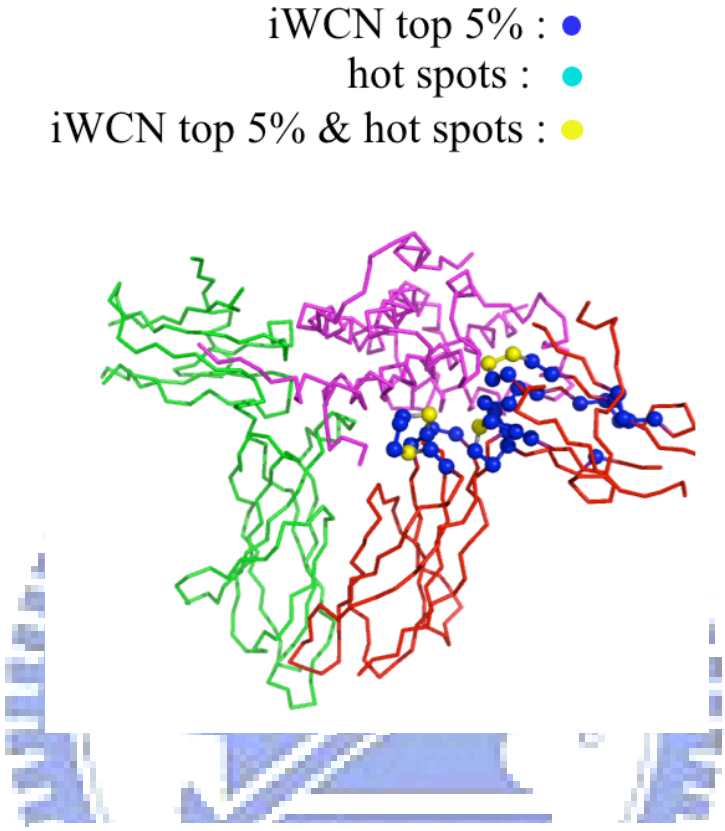
conservation > 9 : ●



B.


COS top 5% :
hot spots :
COS top 5% & hot spots :
COS top 2% :
hot spots :
COS top 2% & hot spots :

C.


iWCN top 5% :
hot spots :
iWCN top 5% & hot spots :

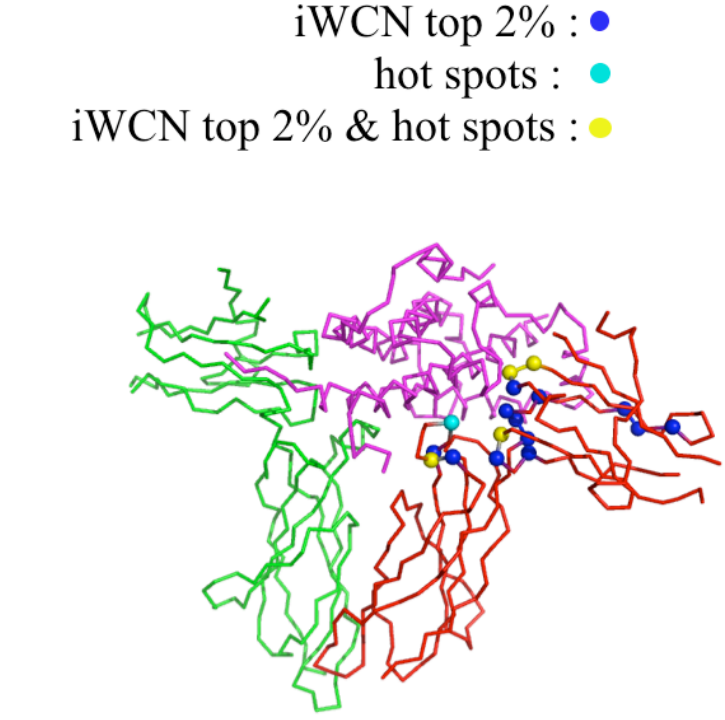

iWCN top 2% :
hot spots :
iWCN top 2% & hot spots :

Table I : The total number residues of conservation > 7 and the percentage of conservation > 7 in iWCN and COS for ATP binding site in protein of PDB code 1A1V .

The total number residues of conservation > 7 :	247
The percentage of conservation > 7 in iWCN top 5% :	89/124 = 71.8%
The percentage of conservation > 7 in iWCN top 2% :	52/66 = 78.8%
The percentage of conservation > 7 in COS top 5% :	112/169 = 66.3%
The percentage of conservation > 7 in COS top 2% :	70/111 = 63.1%

Table II : The percentage of conservation > 7 in iWCN and COS for DNA binding site of PDB code 1A1V.

The percentage of conservation > 7 in iWCN top 5% :	98/167 = 58.7%
The percentage of conservation > 7 in iWCN top 2% :	62/98 = 63.3%
The percentage of conservation > 7 in COS top 5% :	97/178 = 48.9%
The percentage of conservation > 7 in COS top 2% :	65/125 = 52.0%

Table III : The total number residues of conservation > 7 and the percentage of conservation > 7 in iWCN and COS for human hormone receptor binding site of PDB code 3HHR b chain .

The total number residues of conservation > 7 :	110
The percentage of conservation > 7 in iWCN top 5% :	28/39 = 71.8%
The percentage of conservation > 7 in iWCN top 2% :	11/16 = 68.8%
The percentage of conservation > 7 in COS top 5% :	25/40 = 62.5%
The percentage of conservation > 7 in COS top 2% :	7/16 = 43.8%

Table IV : A. The comparison of experimental data and COS to identify hot spot residues in protein of PDB code 3HHR . B. The accuracy of COS .

A.

	experimental data	cos Top 5%	cos Top 2%
R43	X	X	X
E44	-	X	X
W76	-	-	-
T77	-	-	-
S102	-	X	-
I103	-	X	X
W104	X	X	X
I105	X	X	-
C108	-	-	-
E120	-	X	X
K121	-	X	X
C122	-	-	-
D126	-	X	X
E127	-	X	X
D164	-	X	-
I165	X	X	X
Q166	-	X	X
K167	-	X	X
W169	X	-	-
R217	-	-	-
N218	-	-	-

X : Hot spots identified by experimental alanine scanning , and residues predicted to be hot spots .

B.

	TP	FP	FN	TN	Accu
cos Top 5%	4	10	1	6	10/21
cos Top 2%	3	8	2	8	11/21

Accu : Accuracy

Table V : A. The comparison of experimental data and iWCN to identify hot spot residues in protein of PDB code 3HHR . B. The accuracy of iWCN .

A.

	experimental data	iWcn Top 5%	iWcn Top 2%
R43	X	X	X
E44	-	X	-
W76	-	-	-
T77	-	-	-
S102	-	X	-
I103	-	X	X
W104	X	X	X
I105	X	X	X
C108	-	-	-
E120	-	X	X
K121	-	X	-
C122	-	-	-
D126	-	X	X
E127	-	X	X
D164	-	X	X
I165	X	X	X
Q166	-	X	-
K167	-	X	-
W169	X	X	-
R217	-	-	-
N218	-	X	-

X : Hot spots identified by experimental alanine scanning , and residues predicted to be hot spots .

B.

	TP	FP	FN	TN	Accu
iWcn Top 5%	5	11	0	5	10/21
iWcn Top 2%	4	5	1	11	15/21

Accu : Accuracy