STUDY OF PROTEIN-PROTEIN INTERFACES USING GAUSSIAN NETWORK MODEL

by

Seren Soner

B.S., Chemical Engineering, Boğaziçi University, 2007

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering

Boğaziçi University

2009

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor Prof. Turkan Haliloglu, who has been encouraging, understanding, and greatly supportive with me throughout my study in Polymer Research Center.

I would also like to thank my colleagues at the department and my friends at PRC; friends that can make an office as fun, lively and as sharing as possible. Special thanks to Dr. Pemra Özbek are in order, who helped me with her ideas, encouragement and perspective. Andaç Armutlulu and Şölen Ekesan, I consider myself extremely lucky to have had you as fellow friends for six years. Zeynep Kürkçüoğlu, thank you for being by my side through everything, reaching out to me whenever I need someone.

Last, but not least, I would like to offer my thanks to my family; my sister, my mother and my father, to whom I am forever indebted to for their endless care and love.

ABSTRACT

STUDY OF PROTEIN-PROTEIN INTERFACES USING GAUSSIAN NETWORK MODEL

The interface type of the proteins are useful in determining the functioning mechanism of proteins. The interface type may either be obligatory, non-obligatory or crystal (non biological). Previous studies have tried to find the interface type by various sequence and structure based methods such as residue interface propensity, conservation, hydrophobicity, shape etc.. In this thesis, a methodology based on the fluctuations of residues by the Gaussian Network Model (GNM) was developed to predict the interface type for a given protein complex structure. The scoring function identifies the interface type by analysing the domains, the associating regions across the interface of the complex structure and the plausible binding sites of the chains of the complex structure in their isolated states. The reliability of this method was tested on two datasets; PPI-Pred, and CAPRI. Out of 111 proteins in the PPI-pred dataset, correct evaluation rate was 82 percent for obligatory proteins and 76.5 percent for non-obligatory proteins. In the CAPRI experiment, on the other hand, 6 out of 10 submitted models in a pool of predicted models of 1600, were successful. The latter suggests that the method is also successful in discriminating the biological and non-biological interfaces. A web server is built for the prediction of the type of the interface for any given protein complex structure (http://www.prc.boun.edu.tr/appserv/prc/interprot/).

ÖZET

PROTEİN-PROTEİN ARAYÜZLERİNİN GAUSSIAN AĞ MODELLENMESİ

Arayüz tipleri, proteinlerin fonksiyonlarını gerçekleştirme mekanizmalarını anlamakta kullanışlıdırlar. Arayüz tipi zorunlu, zorunlu olmayan, veya kristal (biyolojik olmayan) şeklinde sınıflandırılır. Önceki araştırmalarda arayüz tipi, korunum, arayüzdeki aminoasit dağılımı, hidrofobiklik, şekil gibi çeşitli dizilim ve yapı bazlı yöntemlerle tayin edilmeye çalışılmıştır. Bu tezde, arayüz tipini tahmin etmek icin proteinlerin Gaussian Ağ Modellenmesi ile elde edilen aminoasitlerin dalgalanmasına dayanan bir metod geliştirilmistir. Puanlama fonksiyonu, arayüz tipini tanımlamak iç in alanları, kompleks yapı üzerindeki arayüzde ilişkilenen bölgeleri ve kompleksin zincirlerinin izole durumlarındaki önerilebilir bağlanma noktalarını analiz etmektedir. Metodun güvenilirliği iki veritabanı üzerinde test edilmiştir; PPI-Pred ve CAPRI. PPI-Pred veritabanındaki 111 proteinden zorunlu proteinlerde başarılı tahmin oranı yüzde 82, zorunlu olmayan proteinlerde başarılı tahmin oranı yüzde 76.5'tir. CAPRI yarışmasında ise, 1600 model arasından seçilen on modelden altısı başarılı olmuştur. Bu durum, metodun aynı zamanda biyolojik ve biyolojik olmayan arayüzleri ayırmakta başarılı olduğunu da gostermiştir. Kompleks yapıların arayüz tipinin tahmini iin bir web sunucusu kurulmuştur (http://www.prc.boun.edu.tr/appserv/prc/interprot/).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

K_D	Dissociation constant
r_c	Cut-off radius
R_i	Position vector of residue i
R_i^0	Initial position vector of residue i
ΔR_i	Change in position vector of residue i
ΔR_{ij}	Fluctuation distance between vectors R_i and R_j
U	Eigenvector matrix obtained from decomposition of Kirchoff matrix
u_k	Eigenvector of the k 'th mode
V_{GNM}	Potential function of Gaussian Network Model
γ	Force constant for the harmonic spring
Γ	Kirchoff (connectivity) matrix
λ_k	Eigenvalue of the k'th mode
Λ	Eigenvalue matrix obtained from decomposition of Kirchoff matrix

ANM	Anisotropic Network Model
ASA	Accessible Surface Area
CAPRI	Critical Assessment of Prediction of Interactions
E-I	Enzyme-Inhibitor Complex
ENM	Elastic Network Model
GNM	Gaussian Network Model
NEIT	Non-Enzyme-Inhibitor Transient
NMA	Normal Mode Analysis
PDB	Protein Data Bank
RCSB	Research Collaboratory for Structural Bioinformatics

1. INTRODUCTION

Proteins in the cell interact with other proteins and perform their molecular activities. In processes such as signal transduction and electron transport complexes, interacting proteins have a vital role (Stryer, 1995). In order to carry out their functions, transient proteins may need to come together and bind, or they can permanently act together throughout their lifetime.

Some proteins, such as trimeric G-protein which has a complex between their β and γ subunits, have a permanent interface between the subunits and the chains are always bound to each other. The interfaces and the chains within these complexes are referred to as obligatory (De *et al.*, 2005). The subunits of complexes in this type cannot survive in vivo on their own.

However, there are also complexes that are formed transiently; and the proteins can attach and detach from each other according to the necessity in the cell. These proteins are stable in both their bound state and in their unbound state. Conformation of either both or one of the subunits can change during the binding process. These complexes are referred to as non-obligatory. The oligomerisation state of these complexes in the cell depends on the protein concentration, and the dissociation constant of the complex (K_D). Any change in the concentration of the complex or chains, or the change of pH can shift the oligomeric equilibrium towards any state. So, the transient complexes can also be distinguished into "weak" transient complexes that are generally in the unbound state in vivo, or "strong" transient complexes which are generally in the bound state and dissociate only when triggered (Nooren and Thornton, 2005).

There are also non-biological complexes that can be faced in the protein complex structures from the Protein Data Bank (PDB). Obligatory and non-obligatory structures are known to associate in solution. However, some of the complexes are just artifacts of crystallization, and would not occur in the cell. These are termed as non-biological contacts, or crystal complexes (Liu *et al.*, 2006).

More than the interaction of different subunits of the proteins, these subunits are also formed of domains. Domains are individual folding units within the monomer, which are thought to be evolutionarily conserved important units (Redfern *et al.*, 2007). These domains adopt specific folds, and it is estimated that the number of the domains are up to several thousands (Grant *et al.*, 2004). SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997) are algorithms which aim to cluster the domains according to evolutionary families and fold groups. In these domains, the structure is conserved; however the sequence similarity may fall below 30 percent (Reeves *et al.*, 2006).

1.1. Classification Types Used in Literature

In order to classify the interface types, some criteria are taken into consideration:

- Size and shape: Although the absolute size is also considered, more generally, ΔASA (change of accessible surface area) on complexation is calculated. The reason is that, there is a correlation between the binding strength of the monomers and the ΔASA. Shape of the interface is also considered for classification. It was observed that the mean ΔASA values were 1685 Å for homodimers and 983 Å for hetero-complexes (Jones and Thornton, 1996).
- Residue interface propensity: The hydrophobicity of the interface can give a clue of how well the monomers bind together; and hydrophobicity value can be obtained through the relative hydrophobic residues in the interface. However, this relative hydrophobic residue number must be compared to the distribution of residues occurring on the protein surface (Jones and Thornton 1996).
- Gap Index: Gap volume estimates the volume between two monomers, and gap index is calculated as the ratio of gap volume to the interface accessible surface area. Gap index is also referred to as complementary in some sources (Jones and Thornton 1996).
- Sequence analysis: In order to distinguish the interface type, SWISS-PROT (Bairoch and Apweiler 2000) was also used. The hypothesis behind this idea is that, if a chain of any protein is found by itself, it is submitted to the database. So in order to separate the hetero-oligomers, if the two or more chains of the

oligomer are found only in one SWISS-PROT file, their association is obligatory since they are known to form complex only with the other monomers in the protein, otherwise non-obligatory. For the homo-oligomer cases, if the monomer is found as a functional monomer in the cell, the protein can be classified as transient, since it is known to dissociate and associate in the cell (Ofran and Rost, 2003).

- Interface residue conservation: It was shown that, crystal contacts are not subject to evolutionary constraints, but obligatory and non-obligatory contacts are. Functionally important interfaces are also expected to be conserved (Valdar and Thornton 2001).

1.2. Plan of Attack

There is a hierarchical organization of the domains and the monomers in complex structures of dimers and higher oligomers. Here, the analysis on dimers and high oligomers of both either single domain or multi-domain structures are studied.

With the aim of identification of different interface types, the main premise underlying the plan of the attack is as follows:

The equilibrium fluctuations of residues across the interface in a complex structure are the dynamic fingerprint of the interface type and implied biological function. Coupled to the latter, the dynamic behavior of monomers in the complex structure should somehow pre-exist in their isolated states.. For this, the equilibrium fluctuations of residues in the complex structure and in the isolated states of monomers are calculated by the Gaussian Network Model (GNM) (Haliloglu *et al.*, 1997) (Bahar *et al.*, 1997).

In the hierarchical structural organization, the identification of domains is of interest at first. The interface could be through one or more of these domains of the same chain or different chains of the complex structure with the interface type changing. Here, the interface between different single domain or multi domain chains is focused.

To this end, first the correlation between the fluctuations in the slow modes as well as considering all modes are used to identify the the mechanistically key sites of proteins, the hinge regions, and rigid structural units. The fluctuations across the interface(s) are descriptive for the interface type; transient and obligatory.

A dynamic structural unit of residues across the interface having participation from both chains, the residues of which behave as in a single domain (chain), is a sign of an obligatory interface. On the other hand, the transient interface doesn't seem to require the two domain (chains) to behave as a single chain, but let some of the residues of one of the chains to fluctuate in a coordinated fashion with the residues of the other chain. In others words, one chain should anchor the other, or anchoring could be from both chains, through some residues. Then, the relative fluctuations of residues, in both slow and fast modes in the isolated states of the monomers should suggest a plausible list of anchoring and anchoring groove residues, respectively, that could have a role in association across the interface upon complex formation.

For a given protein structure, the domains are obtained using the first two slowest modes of GNM by clustering the regions which have the same cross-correlation sign using a distance criteria. The range of the domains obtained from these two slowest modes are compared with the chain information; hence finding if all of, or part of the chains are correlated. If the chains intersect in one domain, by using slowest modes point of view, the domain is obligatory. However, other checks are made to get to a final decision about the interface type.

The associating regions at the interfaces are identified by the correlations between the fluctuations of residues in all modes of motion in the complex structure. In other words, if the residues from different chains show positively correlated fluctuations, this is reminiscent of an association between the two chains. This association could be further characterized by the anchoring residues and the residues that meet the anchoring residues on the other chain. In some cases, it is easier to identify which chain anchors to the other or, in some other cases anchoring could be from both chains. Keeping the latter in mind, the positions of these sites at the interface in both chains

are compared with the anchoring and anchoring groove residues by the fluctuations of isolated monomers in the slowest and fastest modes, respectively. This comparison is made as follows:

i. The associating sites from one chain are compared with the anchors found from the relative fluctuations in the slow modes of the same chain in the isolated state.
ii. The associating sites from one chain are compared with the anchoring groove residues found from the relative fluctuations in the fast modes of the same chain in the isolated state.
iii. If the associating sites match with the anchors found from the relative fluctuations in the slow modes, there is a third matching process employed. On the associating site, the pairs in which the cross-correlation values are the highest are found; and this pair is compared with the relative fluctuations in the fast modes of the corresponding monomer.

The overlap for the dynamic behavior in the complex and in the isolated chains is better in the case of complex structure with transient interfaces. On the other hand, the number of the associating sites across the interface is higher since the distribution of cross-correlation values are somewhat random for obligate cases. Additionally, the fluctuations across the interface doesn't show any sign for the associating of the two chains for the crystal interface, which has no biological significance.

Finally, the average of the values of the correlations between the chains are considered for the three types of interfaces. These values are close to zero for the obligatory cases and more negative for the transient cases. For the crystal cases, this value is even more negative, which decrease the possibility of any association between the chains.

Figure 1.1 is a flow chart for the interface type decision process, and summarizes the decision process generally. Six criteria are taken into consideration separately; the two slowest modes, number of associating regions, size of associating regions, the value of and the difference between chain averages of cross-correlations.

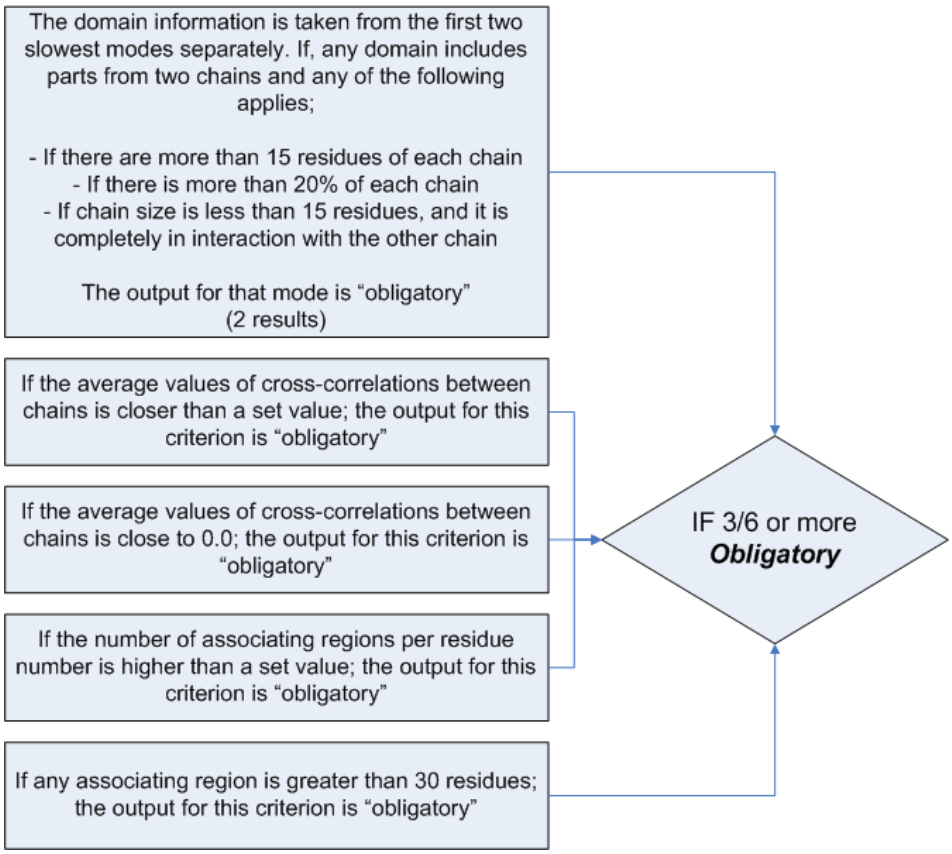

Figure 1.1: Flow chart for protein interface type decision

A server is built up where all the above information will be implemented to a scoring function to suggest the interface type for a given uploaded structure or PDB id.

2. METHODS AND MATERIALS

2.1. Gaussian Network Model (GNM)

Normal mode analysis (NMA) (Rader *et al.*, 2006) is a computational approach to predict the molecular fluctuations near a given equilibrium state. Elastic network models (ENM) based on the polymer network mechanics serve as a simpler approach for the normal mode analysis

In the Gaussian Network Model (GNM) (Haliloglu *et al.* 1997) (Bahar *et al.* 1997), which is the simplest ENM at residue level, it is assumed that the residues are connected by elastic springs, if they are located within a cutoff distance in the folded state. The fluctuations of residues are Gaussian and isotropic (Rader *et al.* 2006). Then the potential of a network of n residues can be written as follows:

$$V_{GNM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\Gamma_{ij}\left[(\Delta X_i - \Delta\ X_j)^2 + (\Delta Y_i - \Delta\ Y_j)^2 + (\Delta Z_i - \Delta\ Z_j)^2\right]\right] \qquad (2.1)$$

Here, γ is the force constant for the spring and Γ stands for the symmetric Kirchoff matrix that represents the connectivity between nodes, between i^{th} and j^{th} residues as

$$\Gamma_{ij} = \begin{cases} -\gamma^* & i \neq j \quad R_{ij} \leq r_{cutoff} \\ 0 & i \neq j \quad R_{ij} > r_{cutoff} \\ -\sum_k^N \gamma^* & i = j \neq k \end{cases} \qquad (2.2)$$

Cut-off radius (r_c) is pre-determined for each mode (7.0 Å for slow modes, 6.5 Å for fast modes, 6.8 Å for average modes) (Haliloglu *et al.*, 2008) (Haliloglu and Erman, 2009). γ^* is a scaling parameter (Bahar *et al.* 1997). A figure describing the cut-off distance and connectivity can be seen below on Figure 2.1.

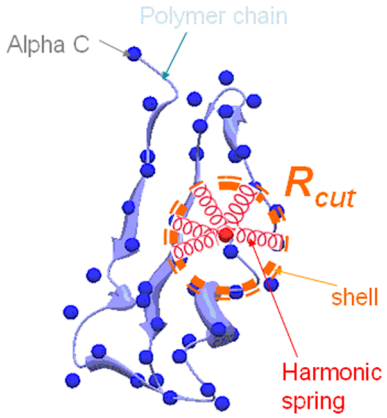

Figure 2.1: Cut-off distance shown on alpha carbons in a protein chain

In the GNM, the correlation between the fluctuations are determined as

$$\langle \Delta R_i \cdot \Delta R_j \rangle = \frac{3}{2} \left[\Gamma^{-1} \right]_{ij} = \frac{3}{2} \left[U(\Lambda^{-1}) U^T \right]_{ij} \tag{2.3}$$

$$\langle \Delta R_i \cdot \Delta R_j \rangle = \frac{3}{2} \sum_k \lambda_k^{-1} (u_k)_i (u_k)_j \tag{2.4}$$

$$\langle \Delta R_i \cdot \Delta R_j \rangle_k = \frac{3}{2} \lambda_k^{-1} (u_k)_i (u_k)_j \tag{2.5}$$

where U is an orthogonal matrix, whose columns represent the eigenvectors (u_i), and Λ is a diagonal matrix whose elements λ_i represent the eigenvalues. These eigenvalues reflect the frequencies of the motion.Due to the degrees of freedom in the system, the last one will be equal to zero; so there will be $n-1$ independent eigenvalues. These modes are organized in an ascending order, so that $\lambda_1 < \lambda_2 < \ldots < \lambda_{n-1}$. In order to minimize the computational cost, BLZPACK (Marques, 1995) is used as the eigensolver of the system.

As eigenvalues decrease, the frequency decreases, and the corresponding modes will be referred to as "slow" modes. As seen from equation 2.5, the value of the correlation between the fluctuations is inversely proportional to the eigenvalue. Slow modes yield the global cooperative motion of the protein, whereas fast modes with

higher frequencies describes more local fluctuations. Also, residues with critical high frequency values can be considered as computational hot spots.

In equation 2.5, $\langle \Delta R_i \cdot \Delta R_j \rangle_k$ is the correlation between fluctuations of i^{th} and j^{th} residues in mode k, λ_k represents the eigenvalue of the kth mode and u$_k$ represents the eigenvector associated with this eigenvalue. Equation 2.4 gives the cross-correlation of i^{th} and j^{th} residues in all modes, which take all of the modes into contribution. The correlation values vary between -1 and 1. A correlation value of 1 suggests that the two residues are auto-correlated, a value of -1 suggests that they are anti-correlated, and a value of 0 suggest that they are not correlated. However, when only one mode is considered (say first mode), the correlation will be either +1, or -1; which can come out quite handy, and will be explained in the next chapter.

On the other hand, the relative fluctuations $\langle \Delta R_{ij}^2 \rangle$ are dependent on two factors; the fluctuations of residues i and j, and the correlation between fluctuations of i and j.

$$\langle \Delta R_{ij}^2 \rangle = \langle \Delta R_i^2 \rangle - 2\langle \Delta\ R_i \cdot \Delta R_j \rangle + \langle \Delta R_j^2 \rangle \tag{2.6}$$

Equation 2.6 suggests that the relative fluctuation $\langle \Delta R_{ij}^2 \rangle$ ncreases when the mobility of residues i and j increase, and also increase when i and j move out of phase (correlated) and decrease when i and j move in phase (anti-correlated) (Haliloglu *et al.* 2008).

2.2. Scoring Functions Applied

2.2.1. Detection of Domains

The GNM has proved to be successful in the prediction of the hinge regions and the rigid parts of a given structure (Emekli *et al.*, 2008) (Bahar and Jernigan 1998). The fluctuations in the slowest modes can suggest structural domains, which are

compact and semi-independent in their motion at the same time (Kundu *et al.*, 2004).

The correlations between the fluctuations (Equation 2.4 with $k = 1$ and $k = 2$) are used to detect the sites where the change of the sign in the cross-correlation values are observed in a given complex structure (see Figure 2.2) The protein has thus been divided into segments of (-) and (+). The lengths of segments are checked to see if it is less than 15 residues. Any segment consisting of less than 15 residues is marked as a "short flexible fragment" and merged with the neighboring rigid part. Then, a clustering procedure is applied to cluster the rigid segments for the dynamic domains. (Emekli *et al.* 2008).

These rigid parts are now clustered using CAST (Ben-Dor *et al.*, 1999) clustering algorithm: Every rigid part has starting and ending points (i, j). When two rigid parts, having the same cross-correlation sign, are considered for clustering, the first one has beginning and ending points i_1 and j_1, and the second one has i_2 and j_2, respectively. The distance between ending points and starting points between rigid parts are calculated; and is clustered as one group if both of the distances are less than 12 Å. The distances between the opposite directions of the rigid part segments are also calculated, namely the distance between $i_1 - j_2$, and $i_2 - j_1$. One exception to this distance-check is that if the end point or the beginning point of the segment is a C-terminus or N-terminus of a protein chain. (Emekli *et al.* 2008)

After the rigid parts are identified in the complex structure, the chain information is compared with the clustered rigid parts of the structure. If one rigid segment does not include parts from other chains, then it is directly categorized as non-obligatory or crystal. However, if it does, then the number of residues involved is investigated. If one or more of the following criteria are met, the interface is marked as obligatory:

- If more than 15 residues of any chain is in interaction with another chain
- If more than 20 percent of any chain is in interaction with another chain
- If any chain of size less than 15 residues is completely in interaction with another chain.

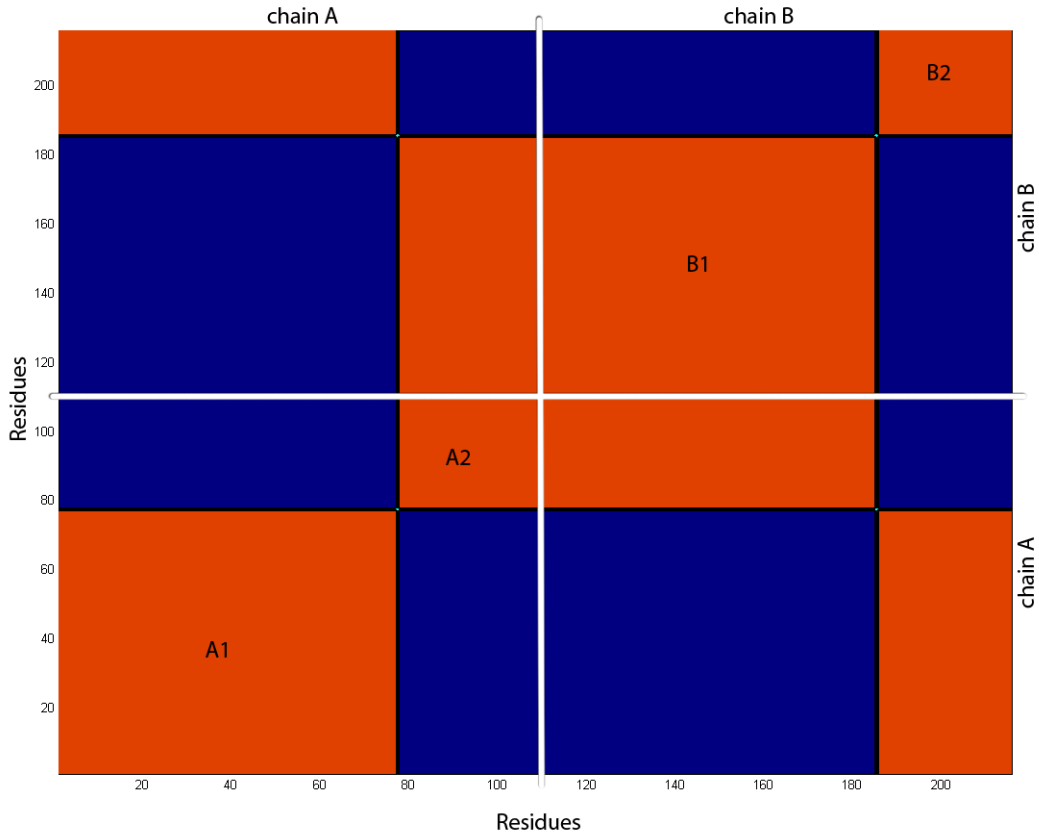

Figure 2.2: Cross-correlation graph for first slowest mode for 1hul. The numbers on both axis are not the residue numbers on the PDB file, but numbers given by the program; so any missing residues will have effect on the numbers

As an example, the case presented in Figure 2.2 is elaborated here. The latter figure displays the correlation between the fluctuations in the slowest first mode for protein complex 1hul, which is a homo dimer with 108x2 residues. The white line shows the point where the first subunit ends and the second begins. The areas marked with red represent a cross-correlation value of (+1), which means the residues in this segment move together in the same direction, and the areas marked with blue represent a cross-correlation value of (-1), which means that the associated pairs move in the opposite direction. The areas marked as A1, A2, B1 and B2 refer to the domains identified for this chain; for example residues 1-77 represent the first domain of chain A. Here, it can be observed that domains A1-B2 and A2-B1 are auto-correlated. The distances between tail points for these domain pairs are checked. Since they are less than 15 Å, they are combined as two rigid parts. The distance constraint is taken by considering

that the two parts should be in close distance to act together. Also, since the number of residues in interaction with another chain is greater than 15 residues, one of the three criteria defined above is met, the interface between these two domains are marked as obligatory.

2.2.2. Correlations Between the Fluctuations suggest the key sites across the interface for the association of the two chains.

The correlation between fluctuation in all modes of motion are obtained by the GNM. For a non-obligatory complex (if not crystal interface), it is expected that there are some residues across the interface display positively correlated fluctuations, while the same residues display negatively correlated fluctuations with the residues in the rest of its own chain. This somehow describes an anchoring behaviour across the interface, (See Figure 2.3) which could be from both chains and also require anchoring groove residues on the other chain to meet. The scoring function program reads the cross-correlation values a given complex structure and stores them into a matrix. It reads the chain information and stores also the chain lengths for each chain in the complex. The statistics of the correlation values are performed as the average, the minimum, the maximum values, and the standard deviations of these values for the chain and between different chains of the complex structure are calculated.

As an example; Figure 2.3, displays the correlations between the fluctuations of residues in all modes of motion for protein 1AVW. Protein 1AVW is a hetero-dimer, for which the first chain consists of 220 residues, whereas the second chain consists of 171 residues. (Song *et al.*, 1998)

Figures 2.5 and 2.6 display the same data on Figure 2.3 in 2-D. Figure 2.5 displays the cross-correlation values of the residues of chain A with the residues of chain B, and Figure 2.6 displays the cross-correlation values of the residues of chain B with the residues of chain A. For both graphs, the correlation values are mostly negative, which is expected for non-obligatory proteins; since the monomers should act independently. However, for Figure 2.5, it can be seen that the most positive cross-correlation can

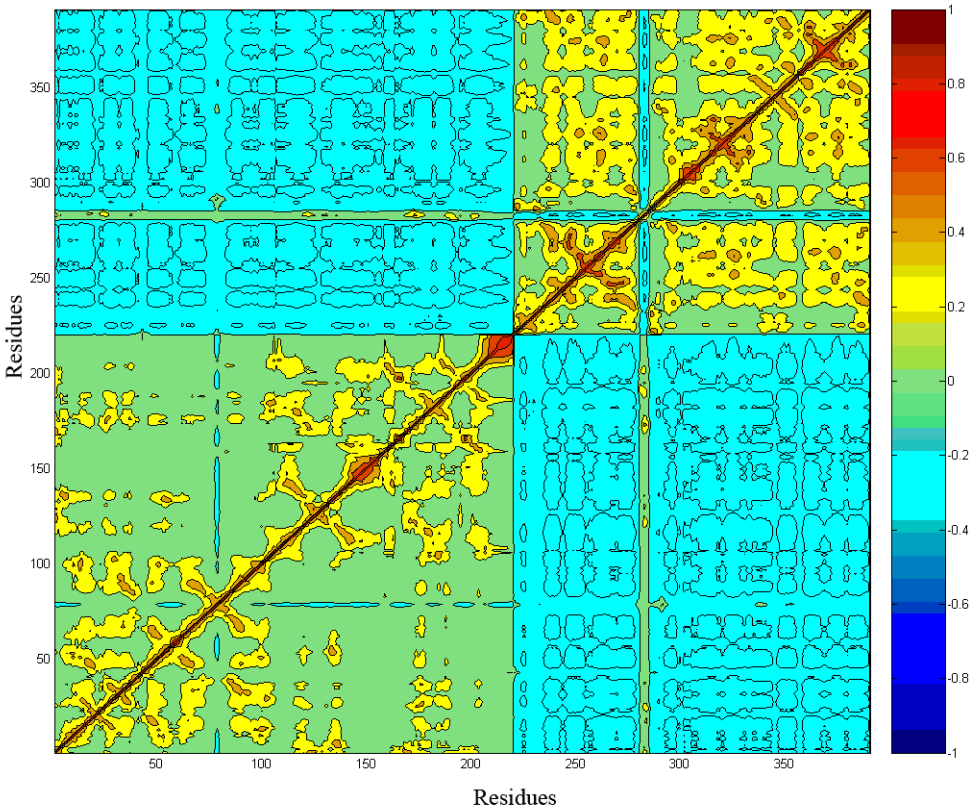

Figure 2.3: Cross-correlation graph for 1avw in all modes

be observed at residue 79 at chain A. However, the cross-correlation values are much higher than this one at residues 282-285, as can be observed from Figure 2.6.

For this protein, the average cross-correlation values between chains are calculated as in Table 2.1.

Also, average cross-correlation values of all residues in chains are calculated. After applying the algorithm defined in Figure 2.4, the possible anchors are calculated as in Table 2.2.

Table 2.1: Average cross-correlation values of chains

	A	**B**
A	0.1656	-0.1971
B	-0.1971	0.2372

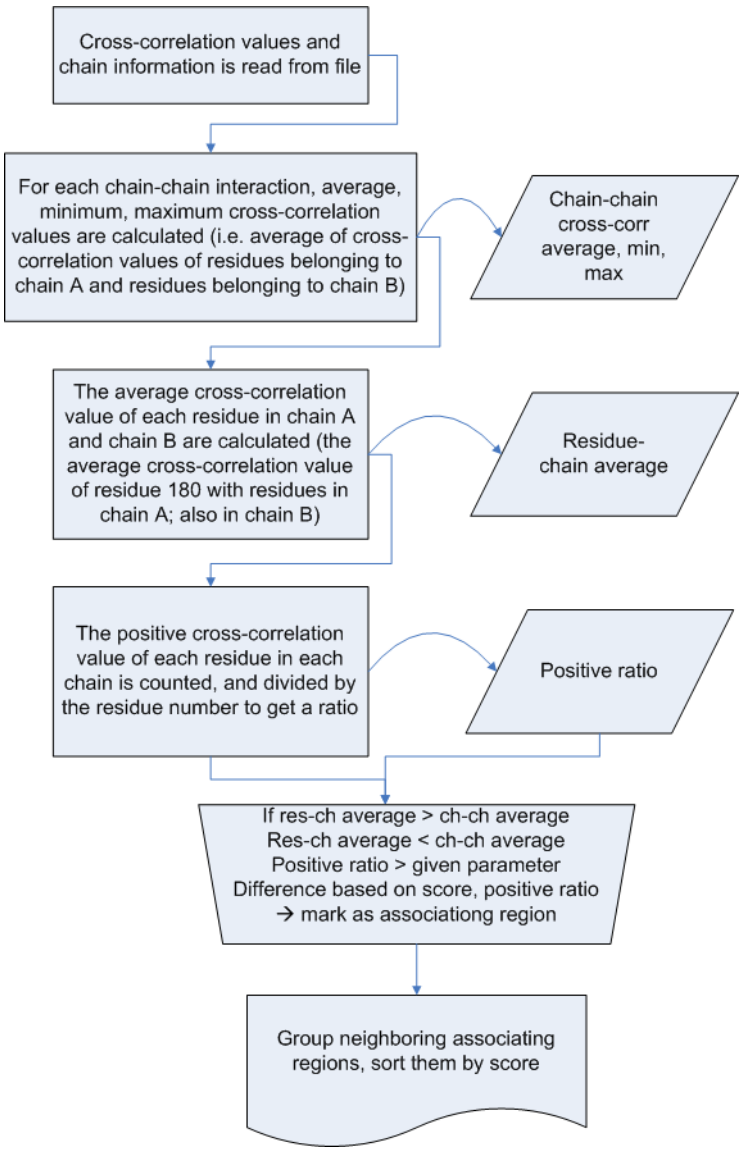

Figure 2.4: Flow chart of associating region finding algorithm

Table 2.2: Anchors from correlation between fluctuations for 1AVW

	Starting residue	**Ending residue**	**Score**
1	281 B 561	285 B 565	0.354
2	79 A 79	79 A 79	0.112

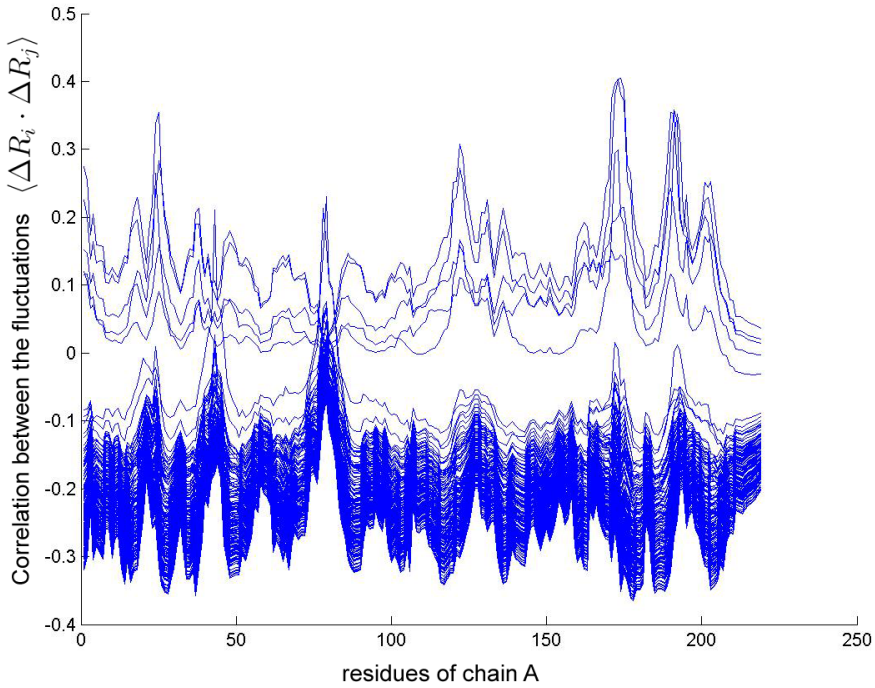

Figure 2.5: Correlation between fluctuation values of residues of chain A and residues of chain B

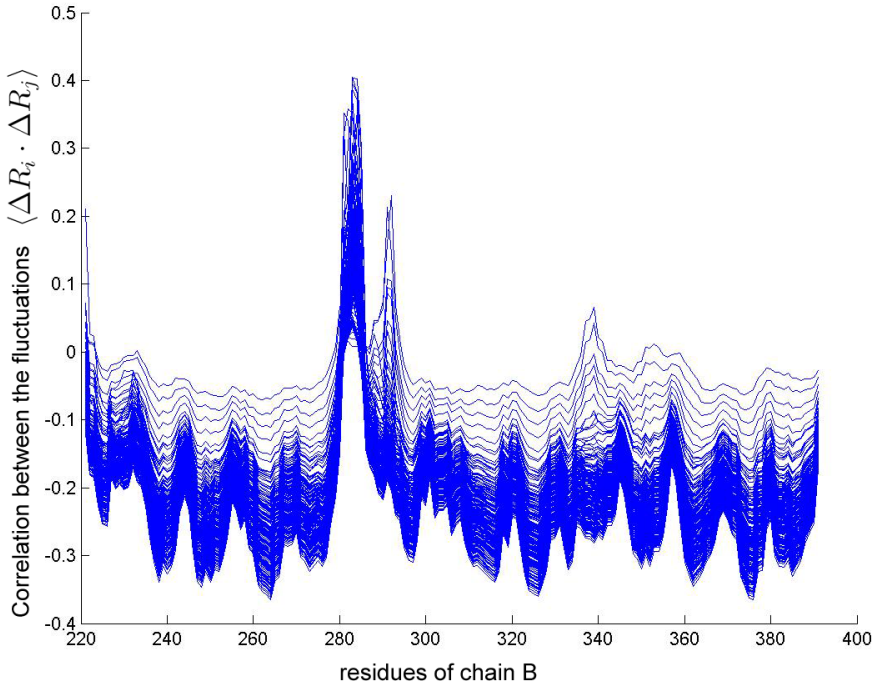

Figure 2.6: Correlation between fluctuation values of residues of chain B and residues of chain A

2.2.3. Anchors from Relative Fluctuations from Slow Modes

The relative fluctuations of residues in the slowest modes of the monomers in the isolated state suggest regions that display highly cooperative motion with the rest of residues and as well as with the residues at its N-terminus and C-terminus of the chain (ΔR_{i1}^2 and ΔR_{in}^2). The higher relative correlation for residues i and j means both high mean square fluctuations for these residues and high negative correlations between the fluctuations of these residues.

As an example, Figures 2.7 and 2.8 present the relative fluctuations of each residue with all other residues, $\langle \Delta R_{ij}^2 \rangle$, in the first slowest mode for chain A and chain B respectively.

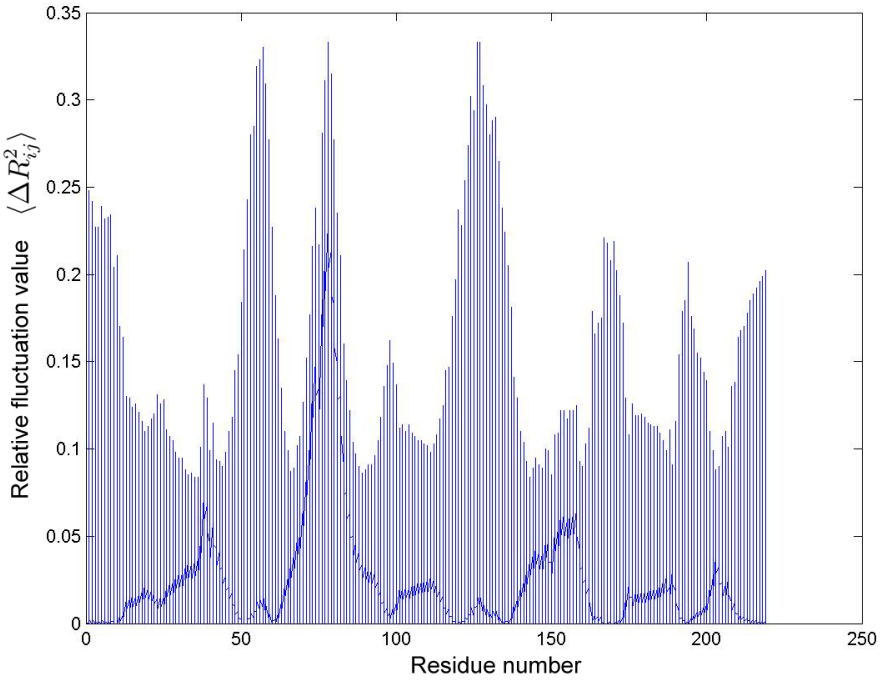

Figure 2.7: Relative fluctuations for the third slowest mode for protein 1AVW chain A

Here the peaks are identified at residues 39, 79, 154 and 159 in chain A and 246, 270, 284, 293, 321, 336, 358 and 380 in chain B. The residues at these peaks display high relative fluctuations with other residues as well as the residues at N and C termini. The relative fluctuation values of each residue with the three residues at both termini

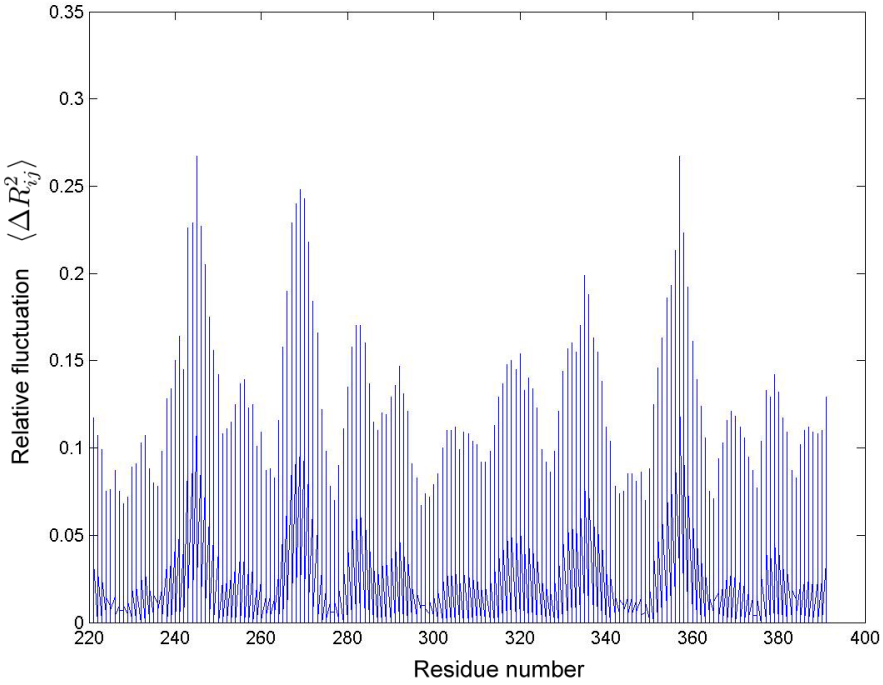

Figure 2.8: Relative fluctuations for the third slowest mode for protein 1AVW chain B

are averaged and recorded. The data in Figures 2.7 and Figure 2.8 are reprocessed to note these peak residues. Figures 2.9 and Figure 2.10 display the average relative fluctuation values of the three residues defined above.

The data in the curve of Figures 2.9 and 2.10 are then processed for the local minima and maxima by checking the derivative at a window of four residues. In identification of these points, a window of residues is considered. In order to record a point as local maxima, it should have a relative fluctuation greater than the residues in that window before and after itself. Different length of residue windows are checked and five is chosen as the most optimum length for this analysis.

The maximum points are then sorted using bubble sort. To eliminate the points with lower peak points, the anchors that have a peak value of "1/r" of the maximum peak value are eliminated. The r value has been taken as 4 after some trial and error on the proteins that are in the dataset. Tables 2.3 and Table 2.4 shows the local maxima points in Figures 2.9 and Figure 2.10 for chain A and B, respectively, sorted and filtered

by the scores.

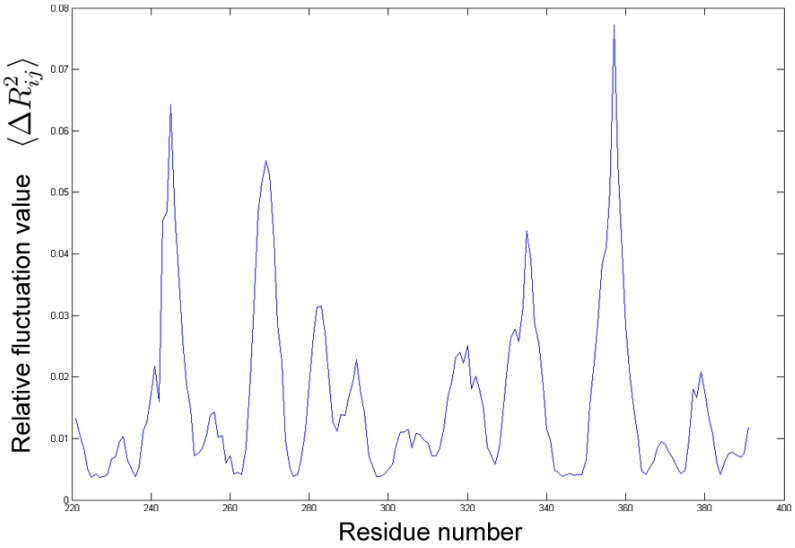

Figure 2.9: Mean relative fluctuations for third slowest mode for 1AVW chain A

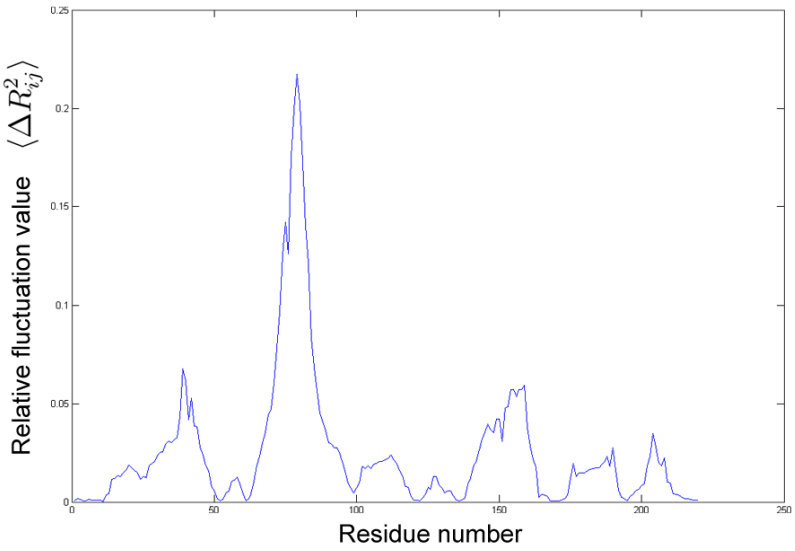

Figure 2.10: Mean relative fluctuations for third slowest mode for AVW chain B

2.2.4. Plausible key residues in binding from fast mode relative fluctuations

The fastest mode relative fluctuations of residues from the protein at a cut-off radius of 6.5 Å are evaluated. All of the values are sorted by descending value. Here, a threshold value is determined. If any score value is repeated 10 times, that value is taken as threshold. The threshold is then multiplied by 1.1, so that values only 10 percent greater than the threshold will be taken as the key residues. These residues could possibly be descriptive of the hot-spot residues at the interface.

Table 2.3: Anchors from correlation between fluctuations for 1AVW for chain A

	Residue number	**Score**
1	79	0.2173
2	39	0.0677
3	159	0.0593
4	154	0.0572

Table 2.4: Anchors from correlation between fluctuations for 1AVW for chain B

	Residue number	**Score**
1	358	0.0772
2	246	0.0643
3	270	0.0552
4	336	0.0437
5	284	0.0315
6	321	0.0252
7	293	0.0228
8	380	0.0208

As an example, the relative fluctuations in first fastest modes of chain A and B of protein complex 1AVW can be seen in Figures 2.11 and Figure 2.12. The peaks are most obvious at residues 174 and 177 at chain A and 362 and 376 at chain B, however there are also local peaks with lower values at 25, 122 and 189 at chain A; and 325, 383, 264 at chain B. We can also observe that the highest number of high relative fluctuation values are observed at 174 and 177 at chain A and 362 and 376 at chain B from the figures. Tables 2.5 and Table 2.6 are the mathematical outputs of the programs, giving these values, sorted by first the number of high relative fluctuation values, then by the value.

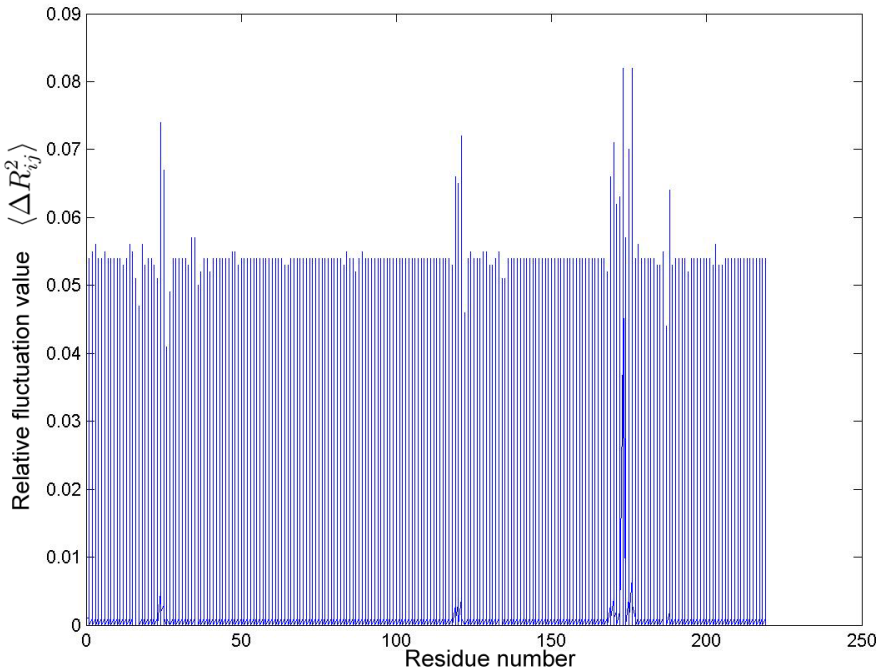

Figure 2.11: Relative fluctuations for first fastest mode for protein 1AVW chain A

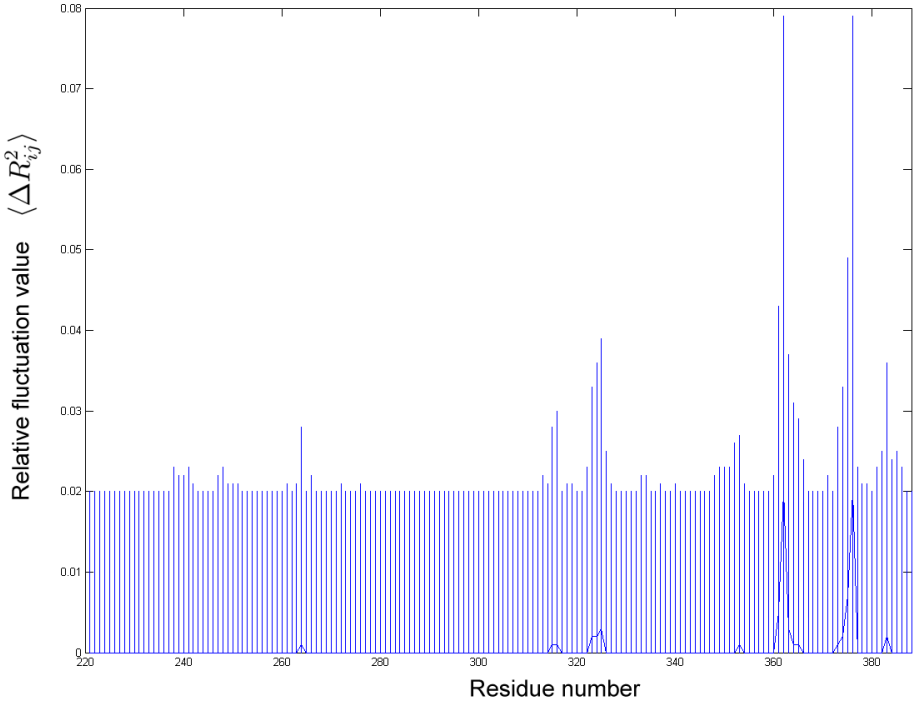

Figure 2.12: Relative fluctuations for first fastest mode for protein 1AVW chain B

Table 2.5: Hot spots obtained for first fastest modes for protein 1AVW chain A

	Residue number	**Count**	**Value**
1	174	20	0.0820
2	177	1	0.0820
3	25	1	0.0740
4	122	1	0.0720
5	171	1	0.0710
6	176	1	0.0700
7	26	1	0.0670
8	170	1	0.0660
9	120	1	0.0660
10	121	1	0.0650
11	189	1	0.0640
12	173	1	0.0630
13	172	1	0.0620

However, the most important thing in scoring the relative fluctuations from fast modes is that, the repetition of the residues which have values greater than the threshold. So after the threshold has been set, the values are then sorted by the repetition count above the threshold, and then the score. This way, both most repeating residues, and the residues with highest fluctuations are taken into consideration. Although residues with high values can be adjacent, they are not grouped to be more exact in matching calculations.

Table 2.6: Hot spots obtained for first fastest modes for protein 1AVW chain B

	Residue number	**Count**	**Value**
1	362	22	0.0790
2	376	21	0.0790
3	375	1	0.0490
4	361	1	0.0430
5	325	1	0.0390

Table 2.6: Hot spots obtained for first fastest modes for protein 1AVW chain B, continued

6	363	1	0.0370
7	324	1	0.0360
8	383	1	0.0360
9	323	1	0.0330
10	374	1	0.0330
11	364	1	0.0310
12	316	1	0.0300
13	365	1	0.0290
14	264	1	0.0280
15	315	1	0.0280
16	373	1	0.0280
17	353	1	0.0270
18	352	1	0.0260
19	382	1	0.0250
20	326	1	0.0250
21	385	1	0.0250
22	384	1	0.0240
23	366	1	0.0240

2.2.5. Matching of the fluctuations of residues in isolated chains and in the complex structure

2.2.5.1. Slow Modes. The relative fluctuations obtained from the six slowest modes for all chains, and the correlation between fluctuation values obtained for the complex are the inputs to the written program. Since the correlation value between the fluctuations are considered as groups of residues, which gives a region, it's important that the peaks of the relative fluctuation residues are in these regions. However, the relative fluctuation values are allowed to be 3 residues away from the groups of residues from correlation between fluctuations, since there may be changes during complex formation. (Haliloglu

et al. 2008) (Camacho and Vajda, 2001). It should be kept in mind that the exactness or the overlap with a window of residues is recorded as a result.

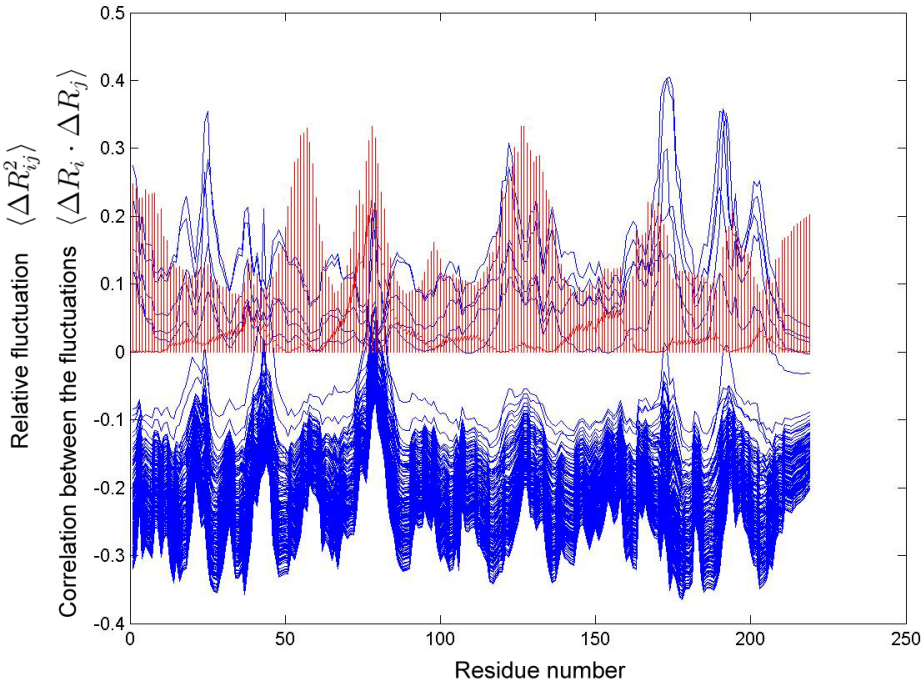

Figure 2.13: Relative fluctuations overlapped with correlation between fluctuations for third slowest mode for protein 1AVW chain A

Figures 2.13 and 2.14 show the overlap of correlation between fluctuations for complex and relative fluctuations in third slowest mode for chain A and be respectively. The two data (correlation between fluctuations for the complex and the relative fluctuations of the slow mode for the monomer) are overlapped, and it is checked if the peaks correspond. Table 2.7 shows the corresponding regions for the data in Figures 2.13 and 2.14. The correlation between fluctuation data is compared with each slow mode of the monomer likewise, and this comparison is also made on other monomers of the complex.

In the scoring of the matched anchors, the residues which match not only in one mode but multiple modes are favored by multiplying the peak value of the anchor by the reciprocal of the eigenvalue of that mode. As the result of the matching, the difference of the matches, the score of the match, and the lowest mode at which the match occurs is printed.

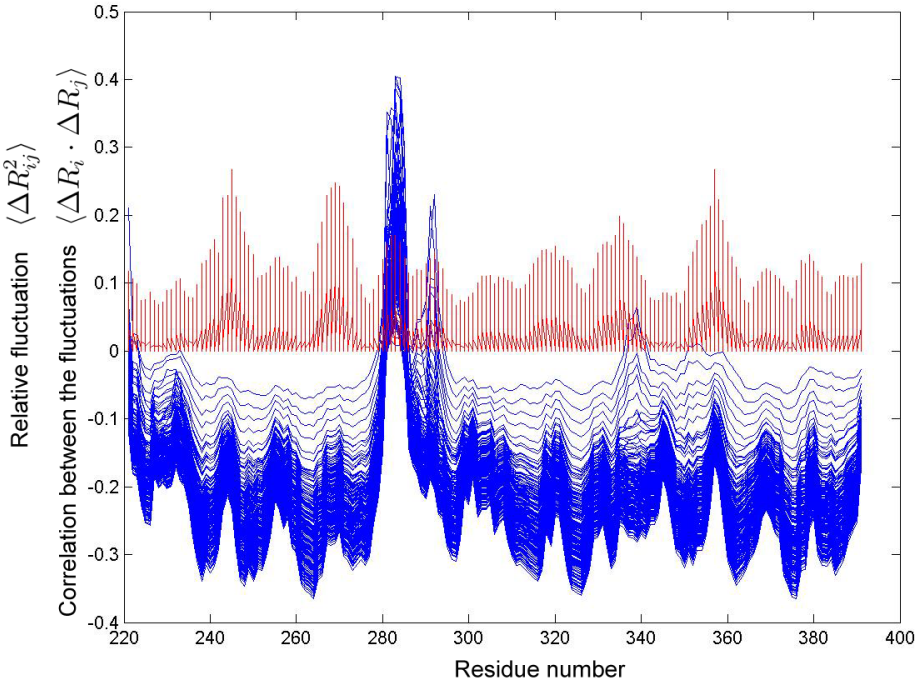

Figure 2.14: Relative fluctuations overlapped with correlation between fluctuations for third slowest mode for protein 1AVW chain B

Table 2.7: Matching of correlation between fluctuations and relative fluctuations in third slowest modes for 1AVW

Group no	**Residue number**	**Chain & PDB Res no**	**Score**	**Difference**	**Mode**
2	79	A 98	0.6814	0	3
1	283	B 564	0.4161	0	3

2.2.5.2. Direct Matching in Fast Modes. The relative fluctuations obtained from the six fastest modes for all individual chains, and the correlation between fluctuation values obtained for the complex are the inputs to the written program. Like the method in section 2.2.5.1, the relative fluctuation values are expected to cluster in space from the correlation between fluctuation values. Like the method above, all of the modes and chains are considered, however modes are now multiplied by the eigenvalue, instead of the reciprocal. This is to emphasize the first fastest mode, which has the highest eigenvalue. The difference of the matches, the score of the match, and the highest mode at which the match is encountered is printed.

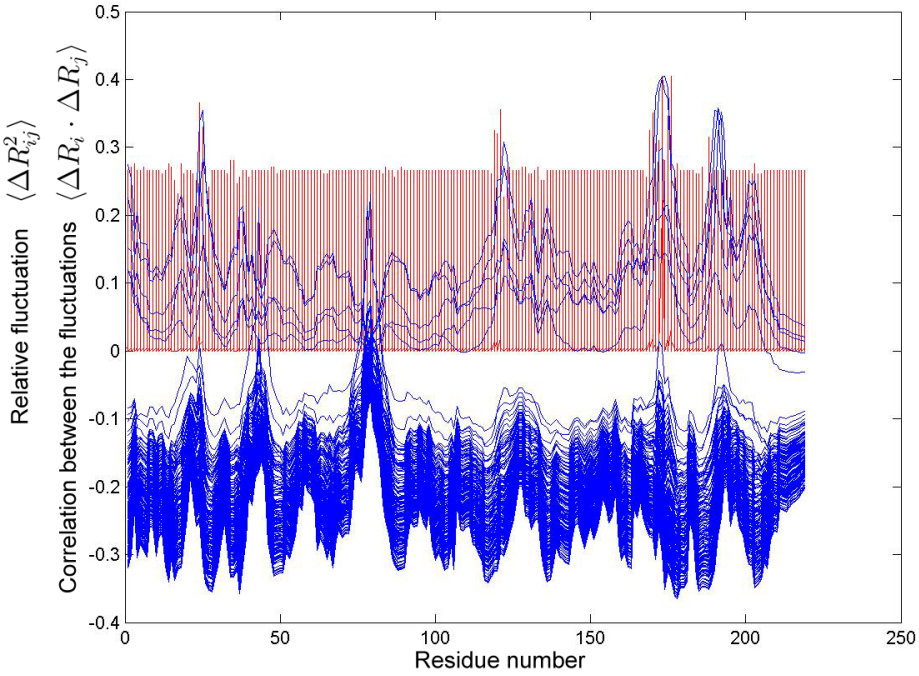

Figure 2.15: Relative fluctuations overlapped with correlation between fluctuations for first fastest mode for protein 1AVW chain A

Figures 2.15 and 2.16 are examples for this comparison for chain A and B respectively. The correlation between fluctuations and the first fast mode relative fluctuations are plotted for individual chains, and the peaks are matched. Like the comparison in 2.2.5.1, this procedure is repeated for each fast mode and each monomer in the complex.

2.2.5.3. Cross Matching in Fast Modes. When there is a match between the relative fluctuations in the slow modes and the correlation between fluctuations, the relative fluctuations in fast modes of the other monomer should be able to show the characteristics of the highly correlated residues on the associating region.

For example, Figure 2.17 is the cross-correlation figure for protein complex 1AVW. Both at residues 79 and 282, there were matches between relative fluctuations in the slow modes and the correlation between fluctuations (see Table 2.9). The marked regions are the regions with highest cross-correlation values on these associating sites. These are expected to match with relative fluctuations in fast modes.

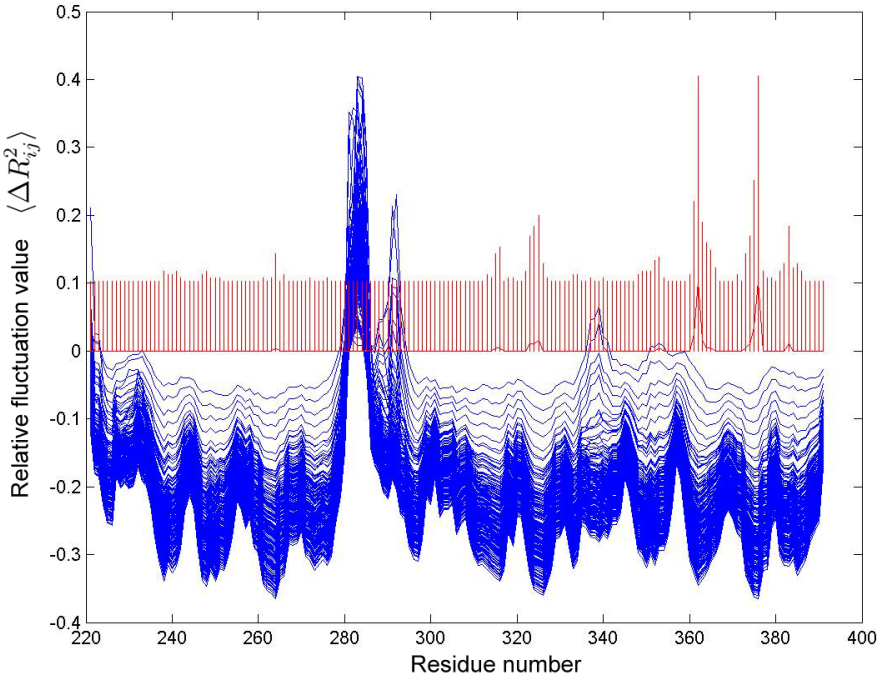

Figure 2.16: Relative fluctuations overlapped with correlation between fluctuations for first fastest mode for protein 1AVW chain B

Table 2.8: Cross matching of correlation between fluctuations and relative fluctuations in third fast mode for 1AVW

From	**To**	**Difference**	**Score**
282	121	2	54.835
282	122	1	51.6896
282	123	0	42.7776
282	124	1	39.1079
282	175	1	38.2167
282	172	2	37.378
282	176	0	35.8577
282	173	3	35.0713
282	3	2	34.7961

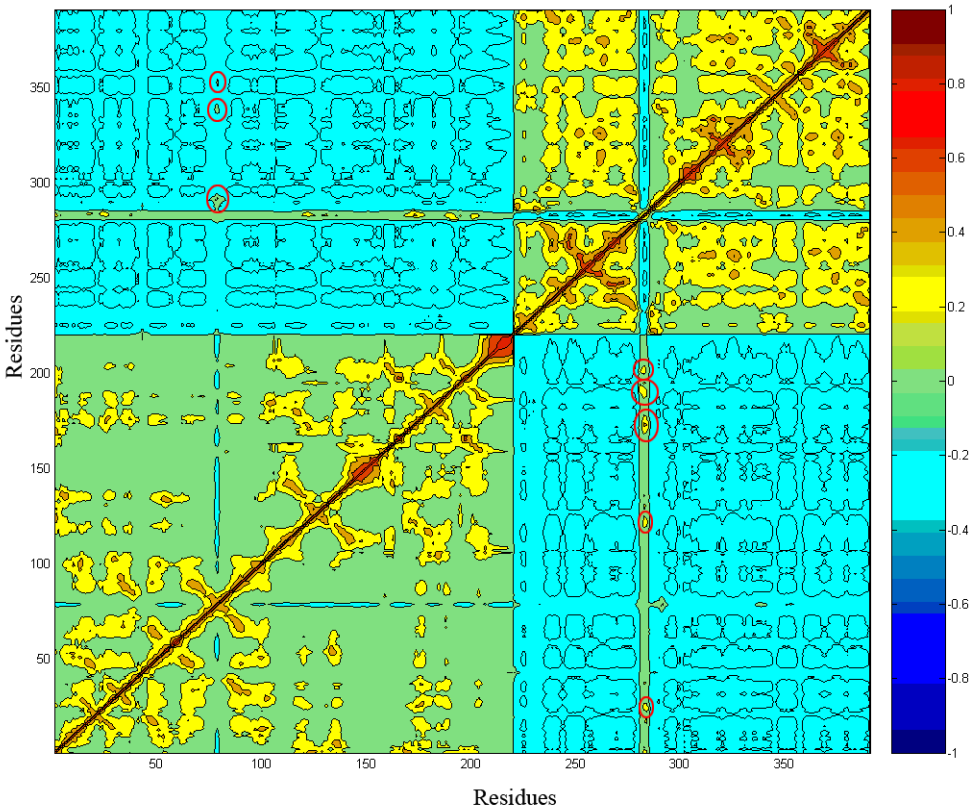

Figure 2.17: Cross-correlation graph for 1avw in all modes. The highest cross-correlated values are marked on the key sites

2.2.6. PPI-Pred Dataset

The training data for the server will be taken from PPI-Pred server (Bradford and Westhead, 2005). The data set is based on the dataset from Thornton (Jones and Thornton 1996) and Neuvirth (Neuvirth *et al.*, 2004). This dataset includes both dimers and oligomers; of which 46 are transient and 102 are obligatory. The transient cases are divided into two by their function; enzyme-inhibitor complexes (25 samples), and non-enzyme-inhibitor transients (21 samples). Obligatory cases are divided into two parts by their structures; homo-obligatory (87 samples) and hetero obligatory (15 samples).

Table 2.9 shows the names and functions for transient cases, where EI stands for Enzyme Inhibitor and NEIT stands for Non-Enzyme Inhibitor Transients and Table

2.10 shows the names and functions for obligatory cases, where Homo stands for Homo-dimers and Hetero stands for Hetero-dimers.

Table 2.9: List of transient type proteins on the training data (Bradford and Westhead, 2005)

PDB ID	**Function**	**Type**
1a4y	ribonuclease inhibitor / angiogenin	E-I
1ava	alpha amylase	E-I
1avw	trypsin / soybean trypsin inhibitor	E-I
1ay7	ribonuclease Sa / Barstar	E-I
1bvn	tendamistat	E-I
1clv	alpha amylase inhibitor	E-I
1cse	eglin C	E-I
1dpj	proteinase A / proteinase inhibitor Ia3	E-I
1dtd	carboxypeptidase A2 / metallocarboxypeptidase inhibitor	E-I
1eai	elastase/chymostrypsin inhibitor I	E-I
1f34	major pepsin inhibitor	E-I
1fss	acetylcholinesterase / fasciculin II	E-I
1gla	Glucose-Specific Factor III	E-I
1kxq	antibody Vhh fragment	E-I
1mct	squash seed inhibitor	E-I
1smp	serratia metalloproteinase / erwinia chysathemi inhibitor	E-I
1tab	bowman-birk inhibitor	E-I
1tgs	pancreatic secretory trypsin inhibitor	E-I
1udi	uracil-DNA glycosylase / UDG inhibitor	E-I
1viw	alpha amylase inhibitor	E-I
2ptc	bovine pancreatic trypsin inhibitor	E-I
2sic	subtilisin BPN / SSI	E-I
4cpa	potato carboxypeptidase A inhibitor	E-I

Table 2.9: List of transient type proteins on the training data, continued

PDB ID	Function	Type
4sgb	serine proteinase B / potato chymotrypsin inhibitor I	E-I
7cei	Im7 / DNAase	E-I
1agr	Rgs 4	NEIT
1atn	Deoxyribonuclease I	NEIT
1b6c	Fk506-Binding Protein / Tgf-B Superfamily Receptor Type I	NEIT
1bkd	Son Of Sevenless-1	NEIT
1buh	Cell cycle regulatory protein Ckshs1	NEIT
1d2z	Death Domain Of Tube	NEIT
1dow	alpha-catenin / beta-catenin A	NEIT
1eay	chey / chea	NEIT
1euv	Ulp1 Protease / Ubitquitin-Like Protein Smt3	NEIT
1f3v	traf2 2.0A	NEIT
1f5q	Cyclin Dependent Kinase 2 / Cyclin	NEIT
1he1	Exoenzyme S	NEIT
1hx1	Bag-Family Molecular Chaperone Regulator-1	NEIT
1i2m	Regulator Of Chromosome Condensation	NEIT
1i8l	Cytotoxic T-Lymphocyte Protein 4	NEIT
1kac	Adenovirus Receptor	NEIT
1pdk	Chaperone Protein Papd / Protein Papk	NEIT
1qav	alpha-1 Syntrophin	NEIT
1tx4	P50-Rhogap / Rhoa	NEIT
1xdt	Heparin-Binding Epidermal Growth Factor / Diphtheria Toxin	NEIT
3ygs	Apoptotic Protease Activating Factor / Procaspase 9	NEIT

Table 2.10: List of obligatory type proteins on training data (Bradford and Westhead, 2005)

PDB ID	**Function**	**Type**
1a0f	Glutathione S-Transferase	Homo
1a4i	Methylenetetrahydrofolate Dehydrogenase	Homo
1a4u	Alcohol Dehydrogenase	Homo
1afr	delta-9 Stearoyl-Acyl Carrier Protein	Homo
1afw	3-Ketoacetyl-Coa Thiolase	Homo
1ahj	nitrite hydratase	Hetero
1aht	alpha thrombin	Hetero
1aj8	Citrate Synthase	Homo
1ajs	Aspartate Aminotransferase	Homo
1aom	Nitrite Reductase	Homo
1aq6	L-2-Haloacid Dehalogenase	Homo
1at3	Herpes Simplex Virus Type II Protease	Homo
1az3	Ecorv Endonuclease	Homo
1b34	Small Nuclear Ribonucleoprotein Sm 2	Hetero
1b3a	Rantes	Homo
1b5e	Deoxycytidylate Hydroxymethylase	Homo
1b7b	Carbamate Kinase	Homo
1b8a	Aspartyl-tRNA Synthetase	Homo
1b8j	Alkaline Phosphatase	Homo
1b9m	Molybdate-Dependent Transcriptional Regulator	Homo
1bbh	cytochrome C	Homo
1bft	Nuclear Factor Nf-kappa-B P65	Homo
1bjn	Phosphoserine Aminotransferase	Homo
1bo1	Phosphatidylinositol Phosphate Kinase	Homo
1brm	Aspartate-Semialdehyde Dehydrogenase	Homo
1bun	beta-2-Bungarotoxin	Hetero
1bw0	Tyrosine Aminotransferase	Homo
1byf	Polyandrocarpa lectin	Homo

Table 2.10: List of obligatory type proteins on the training data, continued

PDB ID	**Function**	**Type**
1byk	Trehalose Operon Repressor	Homo
1c7n	Cystalysin	Homo
1cli	Phosphoribosyl-Aminoimidazole Synthetase	Homo
1cmb	Met Apo-Repressor (Metj)	Homo
1cnz	3-Isopropylmalate Dehydrogenase	Homo
1coz	Glycerol-3-Phosphate Cytidylyltransferase	Homo
1cp2	Nitrogenase Iron Protein	Homo
1dce	Rab Geranylgeranyltransferase	Hetero
1dj7	ferredoxin thioredoxin	Hetero
1dor	Dihydroorotate Dehydrogenase A	Homo
1e0b	Swi6 Protein	Homo
1efv	electron transfer flavoprotein	Hetero
1ete	Flt3 Ligand	Homo
1f5m	Gaf	Homo
1f6y	Iron Sulfur Protein Methyltransferase	Homo
1f8r	L-Amino Acid Oxidase	Homo
1g4y	calcium-activated potassium channel Rsk / calmodulin	Hetero
1gpe	Glucose Oxidase	Homo
1gux	Retinoblastoma Protein	Hetero
1h2a	Ferricytochrome-C3 Oxidoreductase	Hetero
1hgx	Hypoxanthine-Guanine-Xanthine Phosphoribosyl-transferase	Homo
1hjr	RuvC resolvase	Homo
1hss	0.19 alpha-Amylase Inhibitor	Homo
1hul	Interleukin-5	Homo
1isa	Iron(II) Superoxide Dismutase	Homo
1jkm	Brefeldin A Esterase	Homo
1kpe	Protein Kinase C Interacting Protein	Homo

Table 2.10: List of obligatory type proteins on the training data, continued

PDB ID	Function	Type
1luc	Bacterial Luciferase	Hetero
1mka	beta-Hydroxydecanoyl Thiol Ester Dehydrogenase	Homo
1msp	Major Sperm Protein	Homo
1nse	Nitric Oxide Synthase	Homo
1nsy	Nad Synthetase	Homo
1one	Enolase	Homo
1pnk	Penicillin Amidohydrolase	Hetero
1pp2	Phospholipase A2	Homo
1pvu	Pvuii Restriction Endonuclease	Homo
1qae	Extracellular Endonuclease	Homo
1qax	3-Hydroxy-3-Methylglutaryl-Coenzyme	Homo
1qbi	Soluble Quinoprotein Glucose Dehydrogenase	Homo
1qfe	3-Dehydroquinate Dehydratase	Homo
1qfh	Actin Binding Protein 120	Homo
1qi9	Vanadium Bromoperoxidase	Homo
1qor	Quinone Oxidoreductase	Homo
1qqj	Fumarylacetoacetate Hydrolase	Homo
1qu7	Methyl-Accepting Chemotaxis Protein	Homo
1req	Methylmalonyl-Coa Mutase	Hetero
1scf	Stem Cell Factor	Homo
1smt	Transcriptional Repressor Smtb	Homo
1sox	Sulfite Oxidase	Homo
1spu	Copper Amine Oxidase	Homo
1tco	serine/threonine phosphatase 2B	Hetero
1trk	Transketolase	Homo
1vfr	Nad(P)H: Fmn Oxidoreductase	Homo
1vhi	Epstein Barr Virus Nuclear Antigen-1	Homo
1vlt	Aspartate Receptor	Homo
1vok	Tata-Box-Binding Protein	Homo

Table 2.10: List of obligatory type proteins on the training data, continued

PDB ID	**Function**	**Type**
1vsg	Variant Surface Glycoprotein	Homo
1wgj	Inorganic Pyrophosphatase	Homo
1xik	Protein R2 Of Ribonucleotide Reductase	Homo
1xso	Cu, Zn Superoxide Dismutase	Homo
1ypi	Triose Phosphate Isomerase (TIM)	Homo
1yve	Acetohydroxy Acid Isomeroreductase	Homo
2aai	ricin	Hetero
2ae2	Tropinone Reductase-II	Homo
2arc	Arabinose Operon Regulatory Protein	Homo
2gsa	Glutamate Semialdehyde Aminotransferase	Homo
2hdh	L-3-Hydroxyacyl Coa Dehydrogenase	Homo
2hhm	Inositol Monophosphatase	Homo
2nac	Formate Dehydrogenase	Homo
2pfl	Pyruvate Formate-Lyase	Homo
2utg	Uteroglobin	Homo
3tmk	Thymidylate Kinase	Homo
4mdh	Cytoplasmic Malate Dehydrogenase	Homo
5hvp	HIV-1 Protease	Homo

However, when the proteins in this dataset are investigated by literature survey one by one, it was observed that some of the proteins had different interface types than published. Survey showed that 11 obligatory type proteins were found as crystal or transient.

- 1at3: There is a well-defined groove connecting the two monomers to each other. Also, the monomers are 29 Å apart and go under a lot of conformational change in order to assume the correct the orientation of the dimer. (Hoog *et al.*, 1997)
- 1b34: According to SCOPPI (Winter *et al.*, 2006) and Ofran's dataset (Ofran and Rost 2003) this protein complex has two interfaces, one of them permanent

and one of them transient.

- 1b3a: The complex has two active forms: monomer and dimer. However, at some conditions (high concentration, presence of glycosaminoglycans), it prefers the dimer form; but it is not known which is the active form. (Wilken *et al.*, 1999)
- 1bbh: The structure suggests that the dimer is disrupted upon ligand binding, also that there is a lock and key relationship between the monomers. (Ren *et al.*, 1993)
- 1bun: beta-2-Bungarotoxin, hetero-dimer: According to Ofran's dataset (Ofran and Rost 2003), this protein's interface has transient interaction.
- 1dj7: ferredoxin thioredoxin, hetero-dimer: According to Ofran's dataset (Ofran and Rost 2003), this protein's interface has transient interaction.
- 1dor: Dihydroorotate Dehydrogenase, homo-dimer. From the key reference of the protein, it can be seen that the two monomers of this protein are crystallographically independent, and are related by a non-crystallographic twofold axis. Hence, this protein is chosen as crystal when testing the server. (Rowland *et al.*, 1996)
- 1e0b: Swi6 protein, homo-dimer. The key reference of the protein mentions that the monomers are independently stable, however dimerisation may be favoured in vivo. This suggests that the protein can be considered as non-obligatory, since the monomers can bind and unbind when necessary. (Cowieson *et al.*, 2000).
- 1jkm: The biological unit of this protein is tetramer (Wei *et al.*, 1999) and when the tetramer is studied as the complex, the obligatory interaction can be observed. However, when in dimer form, the complex is non-obligatory.
- 1qfe: In Chapter 1, a method that allowed was introduced which allowed us to check the interface type of the protein. According to this method, for a homo-dimer, if one of the monomers can be found seperately in a submitted PDB file, this protein complex is said to be transient. And 1QFE's monomer can be found separately (PDB ID: 1gqn).
- 1ypi: The association constant for the two monomers has been evaluated; which suggests that both of the forms, dimer and monomer can exist (Lolis *et al.*, 1990)

3. RESULTS & DISCUSSION

3.1. Obligatory Cases

The obligatory cases are investigated with the methods defined in Chapter 2, and the GNM output has been studied by using the server and the visual evaluation of the graphical outputs.

The visual results, and the server results for obligatory cases suggested by the dataset are given in Table 3.1.

Table 3.1: Prediction results of obligatory cases on training data (Bradford and Westhead, 2005)

PDB ID	**Dataset**	**Visual result**	**Server result**
1a0f	Obligatory	Obligatory	Obligatory
1a4i	Obligatory	Obligatory	Non-obligatory
1a4u	Obligatory	Non-obligatory	Non-obligatory
1afw	Obligatory	Obligatory	Obligatory
1aj8	Obligatory	Obligatory	Obligatory
1ajs	Obligatory	Obligatory	Obligatory
1aom	Obligatory	Obligatory	Obligatory
1aq6	Obligatory	Obligatory	Obligatory
1az3	Obligatory	Non-obligatory	Non-obligatory
1b5e	Obligatory	Obligatory	Non-obligatory
1b8j	Obligatory	Obligatory	Obligatory
1b9m	Obligatory	Obligatory	Obligatory
1bjn	Obligatory	Obligatory	Obligatory
1bo1	Obligatory	Obligatory	Non-obligatory
1bw0	Obligatory	Non-obligatory	Non-obligatory
1byf	Obligatory	Obligatory	Obligatory
1byk	Obligatory	Obligatory	Obligatory

Table 3.1: Prediction results of obligatory cases on training data, continued

PDB ID	**Dataset**	**Visual result**	**Server result**
1cmb	Obligatory	Obligatory	Obligatory
1cnz	Obligatory	Obligatory	Obligatory
1coz	Obligatory	Non-obligatory	Obligatory
1cp2	Obligatory	Non-obligatory	Non-obligatory
1dpj	Obligatory	Obligatory	Obligatory
1efv	Obligatory	Obligatory	Obligatory
1f34	Obligatory	Obligatory	Obligatory
1f5m	Obligatory	Non-obligatory	Non-obligatory
1f6y	Obligatory	Obligatory	Obligatory
1g4y	Obligatory	Obligatory	Obligatory
1gpe	Obligatory	Crystal	Crystal
1h2a	Obligatory	Obligatory	Non-obligatory
1hgx	Obligatory	Non-obligatory	Non-obligatory
1hul	Obligatory	Obligatory	Obligatory
1isa	Obligatory	Non-obligatory	Non-obligatory
1kpe	Obligatory	Obligatory	Obligatory
1luc	Obligatory	Obligatory	Obligatory
1mct	Obligatory	Obligatory	Obligatory
1mka	Obligatory	Obligatory	Non-obligatory
1msp	Obligatory	Non-obligatory	Non-obligatory
1nse	Obligatory	Obligatory	Obligatory
1nsy	Obligatory	Obligatory	Obligatory
1one	Obligatory	Obligatory	Non-obligatory
1pnk	Obligatory	Obligatory	Obligatory
1pp2	Obligatory	Non-obligatory	Non-obligatory
1pvu	Obligatory	Obligatory	Obligatory
1qae	Obligatory	Non-obligatory	Non-obligatory
1qax	Obligatory	Obligatory	Obligatory
1qbi	Obligatory	Non-obligatory	Non-obligatory

Table 3.1: Prediction results of obligatory cases on training data, continued

PDB ID	**Dataset**	**Visual result**	**Server result**
1qfh	Obligatory	Obligatory	Obligatory
1qi9	Obligatory	Obligatory	Obligatory
1qor	Obligatory	Obligatory	Non-obligatory
1qqj	Obligatory	Non-obligatory	Non-obligatory
1qu7	Obligatory	Obligatory	Obligatory
1smt	Obligatory	Obligatory	Obligatory
1sox	Obligatory	Non-obligatory	Non-obligatory
1spu	Obligatory	Obligatory	Obligatory
1trk	Obligatory	Non-obligatory	Non-obligatory
1vfr	Obligatory	Obligatory	Obligatory
1vhi	Obligatory	Obligatory	Non-obligatory
1vlt	Obligatory	Obligatory	Non-obligatory
1vok	Obligatory	Non-obligatory	Obligatory
1vsg	Obligatory	Obligatory	Obligatory
1wgj	Obligatory	Non-obligatory	Non-obligatory
1xik	Obligatory	Obligatory	Obligatory
1xso	Obligatory	Non-obligatory	Non-obligatory
2aai	Obligatory	Obligatory	Obligatory
2ae2	Obligatory	Non-obligatory	Obligatory
2arc	Obligatory	Non-obligatory	Non-obligatory
2gsa	Obligatory	Obligatory	Obligatory
2hdh	Obligatory	Obligatory	Obligatory
2hhm	Obligatory	Non-obligatory	Non-obligatory
2nac	Obligatory	Obligatory	Obligatory
2pfl	Obligatory	Obligatory	Obligatory
2utg	Obligatory	Obligatory	Obligatory
4mdh	Obligatory	Non-obligatory	Non-obligatory

From the 73 obligatory proteins in the dataset, 23 are found to be non-obligatory

by using visual evaluation. Of the remaining 50, 41 are evaluated as obligatory by using the server. Since the server results should be based on not the dataset, but our evaluation of the dataset, the correct result ratio is 82 percent.Three obligatory cases, 1hul, 1qfh and 1vsg are examined here in detail. 1hul is a homo-dimer with 108x2

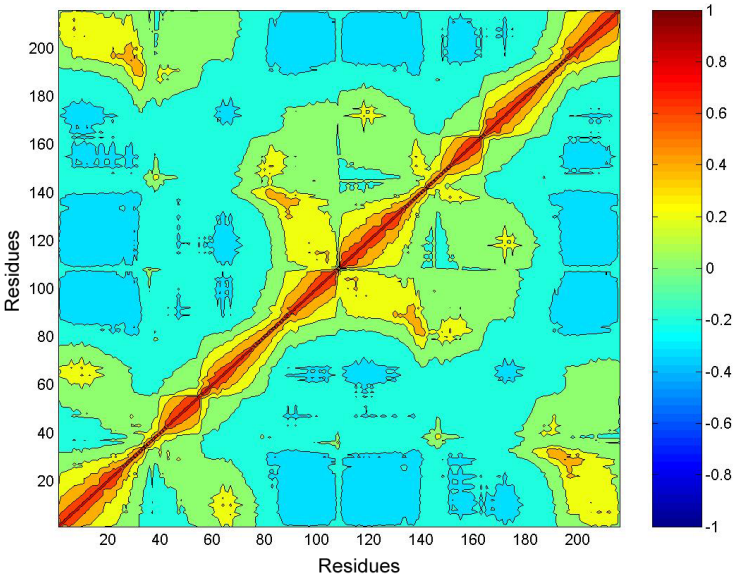

Figure 3.1: Cross-correlation graph for 1hul in all modes

residues (Milburn *et al.*, 1993). However, when the cross-correlation graph is studied (Figures 2.2 and 3.1), the first domain and the second domain of the chains have negative cross-correlation, while the first domain of chain A and the second domain of chain B have positive cross-correlation. This fact leads the protein to be classified as obligatory.

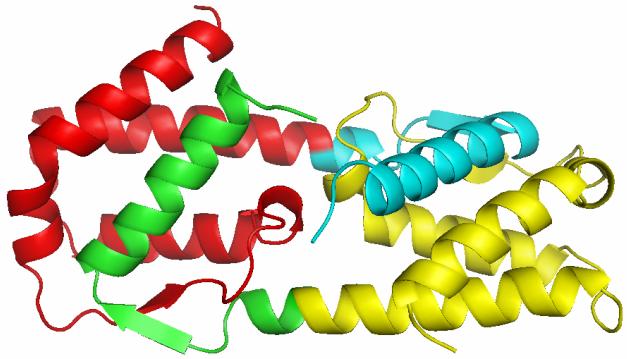

Figure 3.2: Cartoon view for 1hul. Parts colored as yellow represents domain A1 (1-77), green represents A2 (78-108), red represents B1 (109-185), cyan represents B2 (186-216).

1qfh is a homo-dimer with 212x2 residues (McCoy *et al.*, 1999). When the cross-correlation graph (Figure 3.3) is studied, it can be observed that A2 domain's interaction with B2 is obligatory, and these domains have an associating region with A1 and B1. However, A1 and B1 domains show crystal interface property, which is in line with the fact that these domains are far in their crystallographic state as seen in Figure 3.4.

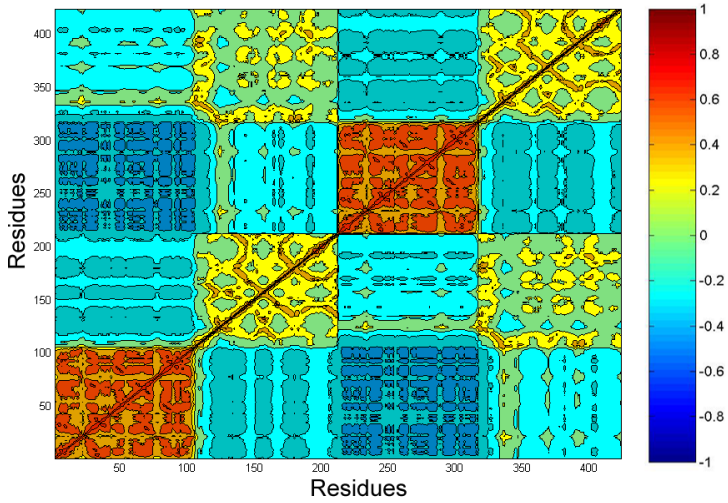

Figure 3.3: Cross-correlation graph for 1qfh in all modes

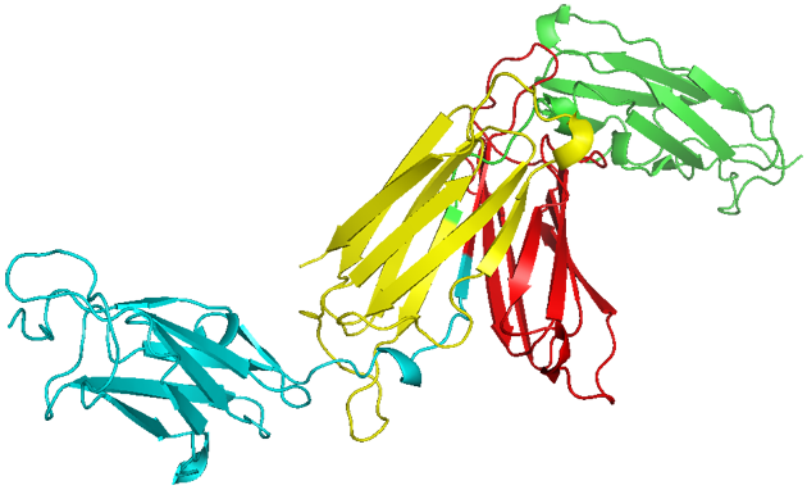

Figure 3.4: Cartoon view for 1qfh. Parts colored as green represents domain A1 (1-116), yellow represents A2 (117-212), cyan represents B1 (213-328), red represents B2 (329-424).

1vsg is a homo-dimer with 362x2 residues (Feymann *et al.*, 1990). When the cross-correlation graph is studied, it can be observed that there are basically two domains; and these domains are in obligatory interaction with each other; and have a crystal interface with the other domain. As mentioned in Chapter 1, this holds an example for

the cases where the chains behave as a single chain. In such cases, end of the chains can not be identified easily and hence there is no sharp distinction between the chains.

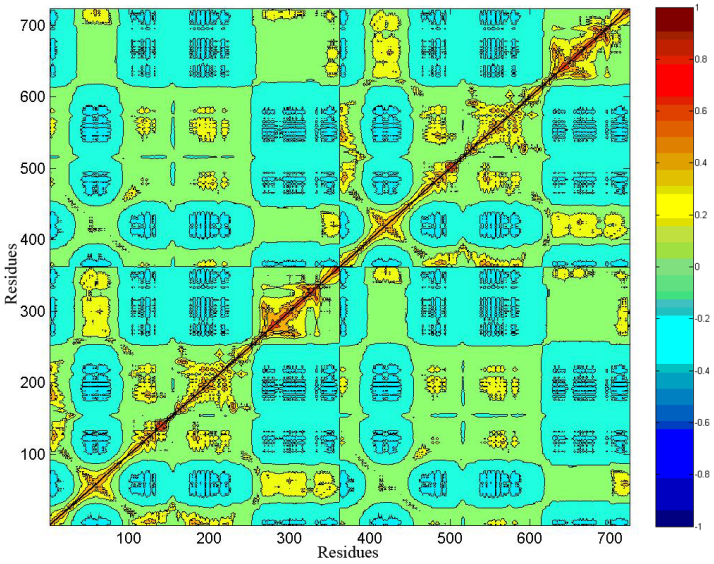

Figure 3.5: Cross-correlation graph for 1vsg in all modes

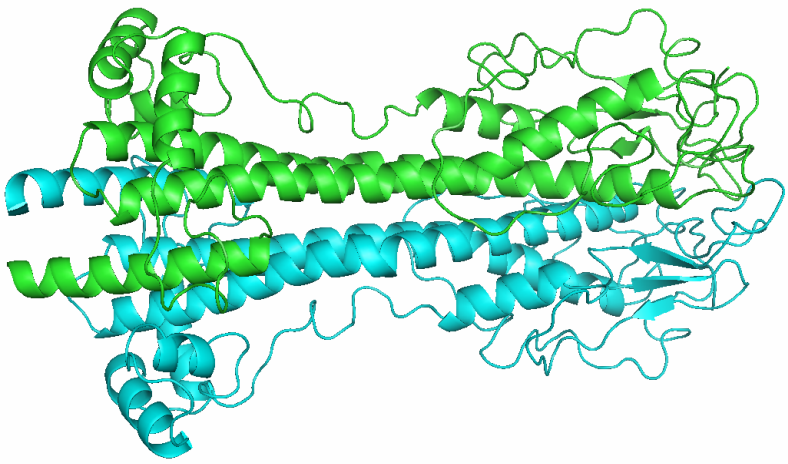

Figure 3.6: Cartoon view for 1vsg. Chain A is represented in green, chain B is represented in cyan.

3.2. Non-obligatory Cases

Non-obligatory cases are investigated using the same method in section 3.1. The visual results, and the server results for all non-obligatory cases suggested by the dataset are given in Table 3.2.

Table 3.2: Prediction results of non-obligatory cases on training data (Bradford and Westhead, 2005)

PDB ID	**Dataset**	**Visual result**	**Server result**
1at3	Non-obligatory	Non-obligatory	Non-obligatory
1atn	Non-obligatory	Non-obligatory	Non-obligatory
1avw	Non-obligatory	Non-obligatory	Non-obligatory
1ay7	Non-obligatory	Non-obligatory	Non-obligatory
1b34	Non-obligatory	Non-obligatory	Non-obligatory
1b3a	Non-obligatory	Non-obligatory	Non-obligatory
1bbh	Non-obligatory	Non-obligatory	Non-obligatory
1bkd	Non-obligatory	Non-obligatory	Non-obligatory
1buh	Non-obligatory	Non-obligatory	Obligatory
1bun	Non-obligatory	Non-obligatory	Non-obligatory
1bvn	Non-obligatory	Non-obligatory	Obligatory
1cse	Non-obligatory	Non-obligatory	Non-obligatory
1dj7	Non-obligatory	Non-obligatory	Non-obligatory
1dow	Non-obligatory	Obligatory	Obligatory
1dtd	Non-obligatory	Non-obligatory	Non-obligatory
1e0b	Non-obligatory	Crystal	Crystal
1euv	Non-obligatory	Non-obligatory	Non-obligatory
1f3v	Non-obligatory	Non-obligatory	Non-obligatory
1fss	Non-obligatory	Non-obligatory	Obligatory
1hx1	Non-obligatory	Non-obligatory	Obligatory
1jkm	Non-obligatory	Non-obligatory	Non-obligatory
1kac	Non-obligatory	Non-obligatory	Non-obligatory
1pdk	Non-obligatory	Non-obligatory	Obligatory
1qav	Non-obligatory	Non-obligatory	Obligatory
1qfe	Non-obligatory	Non-obligatory	Non-obligatory
1smp	Non-obligatory	Non-obligatory	Non-obligatory
1tab	Non-obligatory	Non-obligatory	Non-obligatory
1tgs	Non-obligatory	Non-obligatory	Non-obligatory

Table 3.2: Prediction results of non-obligatory cases on training data, continued

PDB ID	Dataset	Visual result	Server result
1tx4	Non-obligatory	Non-obligatory	Obligatory
1udi	Non-obligatory	Non-obligatory	Non-obligatory
1viw	Non-obligatory	Non-obligatory	Non-obligatory
1xdt	Non-obligatory	Obligatory	Obligatory
1ypi	Non-obligatory	Non-obligatory	Non-obligatory
2sic	Non-obligatory	Non-obligatory	Non-obligatory
3ygs	Non-obligatory	Non-obligatory	Non-obligatory
4sgb	Non-obligatory	Non-obligatory	Non-obligatory
7cei	Non-obligatory	Non-obligatory	Obligatory

Table 3.2 shows that, from the 37 non-obligatory proteins in dataset, two are found to be obligatory, and one is found to be crystal based on visual evaluation. Of the remaining 34, 26 are evaluated as non-obligatory by using the server. Since the server results should be based on not the dataset, but visual evaluation of the dataset, the correct result ratio is 76.5 percent.

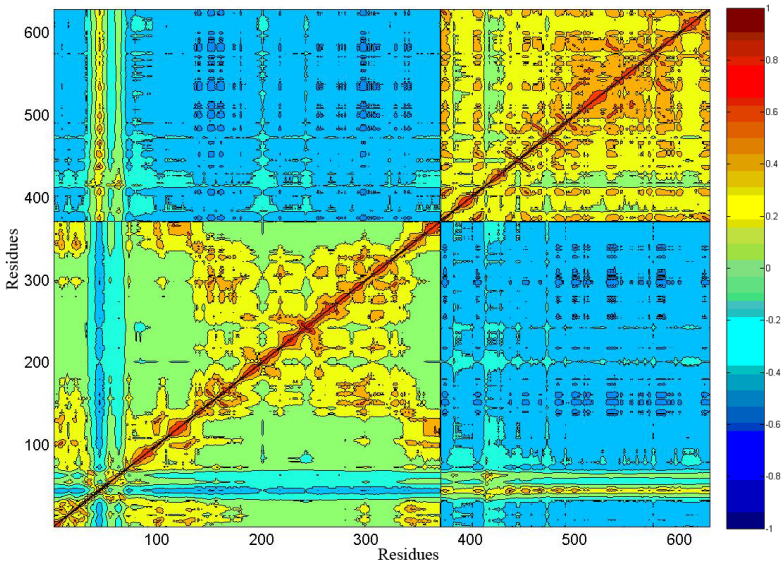

Figure 3.7: Cross-correlation graph for 1atn in all modes

In addition to 1avw, which was discussed in detail in Chapter 2, the following

cases (1atn and 1cse) are also representative structures for non-obligatory cases.

1atn, which is a heterodimer with chain A having 371 residues and chain D having 258 residues, is also a good example for non-obligatory cases (Kabsch *et al.*, 1990). The main associating region is between residues 31-71 in chain A as can be observed from the cross-correlation graph (Figure 3.7), and this region overlaps well with the relative fluctuations in the second slow mode of chain A (see Figure 3.8). As observed from Figure 3.9, the cartoon representation of the protein complex shows that the associating region is buried in the interface.

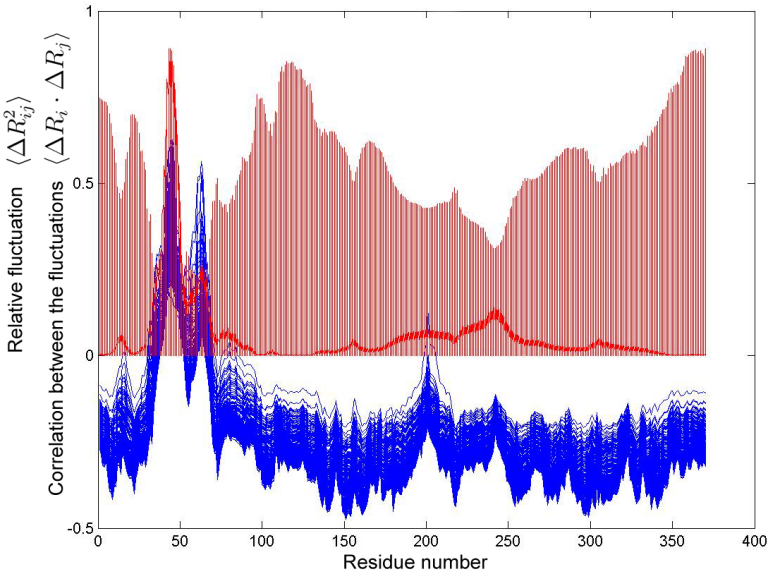

Figure 3.8: The cross-correlation in all modes for the complex and relative fluctuations in second slowest mode of the chain A for 1atn, overlapped.

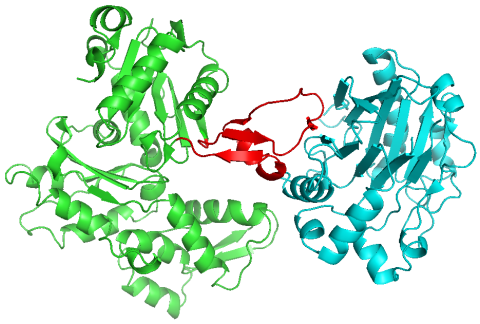

Figure 3.9: Cartoon view for 1atn. Chain A is represented in green, chain B is represented in cyan. The main associating region; residues 31-71 in chain A is colored as red.

1cse is an enzyme-inhibitor complex (Bode *et al.*, 1987) with the enzyme subunit of 274 residues and the inhibitor subunit of 63 residues. Figure 3.10 suggests that the main associating region is at inhibitor subunit, at residues 307-315; moreover there is another associating region at residues 126-130 at the enzyme subunit. These associating regions overlap well with the relative fluctuations in slow modes as can be seen in Figures 3.11 and 3.12.

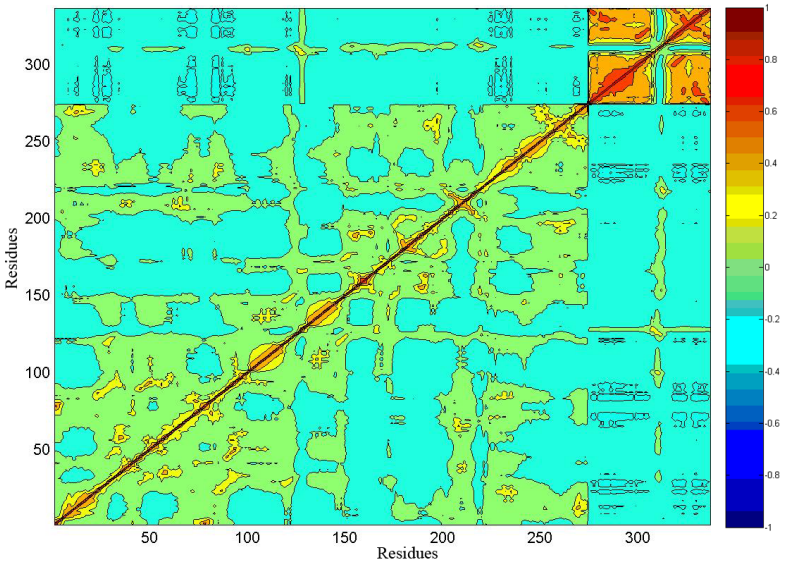

Figure 3.10: Cross-correlation graph for 1cse in all modes

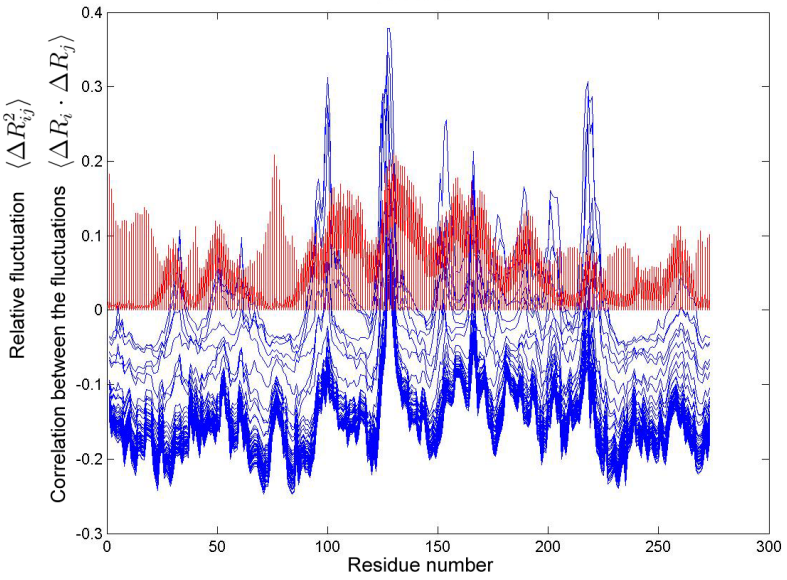

Figure 3.11: The cross-correlation in all modes for the complex and relative fluctuations in second slowest mode of the chain E for 1cse, overlapped.

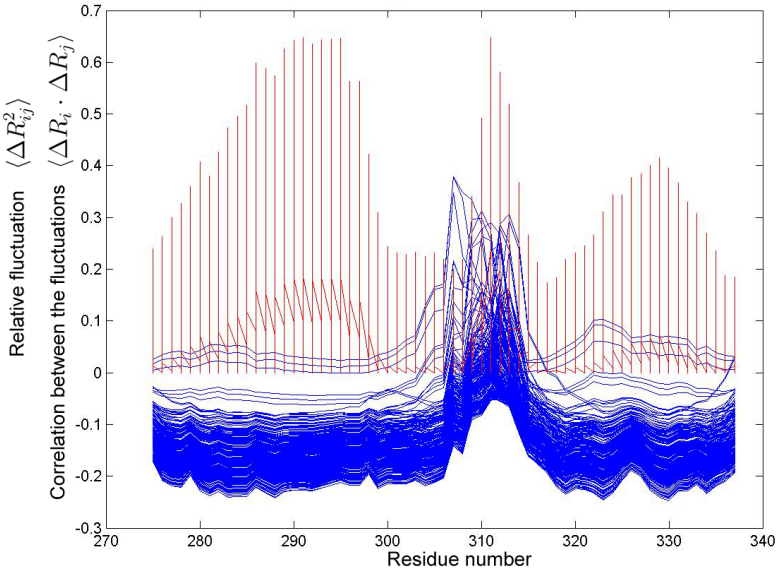

Figure 3.12: The cross-correlation in all modes for the complex and relative fluctuations in first slowest mode of the chain I for 1cse, overlapped.

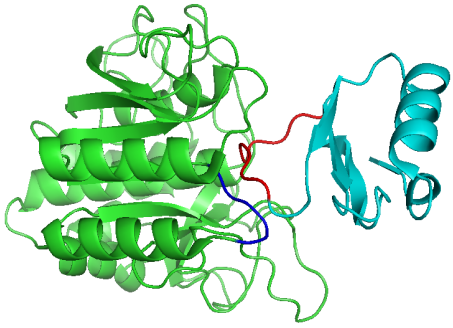

Figure 3.13: Cartoon view for 1cse. Chain A is represented in green, chain B is represented in cyan. The main associating region; residues 307-315 in chain I is colored as red, residues 126-130 in chain E is colored in blue.

3.3. Misdefined Cases

In some cases the results obtained with the method is not in agreement with the literature data. Table 3.3 holds a list of these 24 cases. These cases will be further discussed in Subsections 3.3.1, 3.3.2 and 3.3.3.

Table 3.3: Prediction results of misdefined cases on training data (Bradford and Westhead, 2005)

PDB ID	**Dataset**	**Visual result**	**Server result**
1a4u	Obligatory	Non-obligatory	Non-obligatory
1az3	Obligatory	Non-obligatory	Non-obligatory
1bw0	Obligatory	Non-obligatory	Non-obligatory
1coz	Obligatory	Non-obligatory	Obligatory
1cp2	Obligatory	Non-obligatory	Non-obligatory
1dow	Non-obligatory	Obligatory	Obligatory
1f5m	Obligatory	Non-obligatory	Non-obligatory
1gpe	Obligatory	Crystal	Crystal
1hgx	Obligatory	Non-obligatory	Non-obligatory
1isa	Obligatory	Non-obligatory	Non-obligatory
1msp	Obligatory	Non-obligatory	Non-obligatory
1pp2	Obligatory	Non-obligatory	Non-obligatory
1qae	Obligatory	Non-obligatory	Non-obligatory
1qbi	Obligatory	Non-obligatory	Non-obligatory
1qqj	Obligatory	Non-obligatory	Non-obligatory
1sox	Obligatory	Non-obligatory	Non-obligatory
1trk	Obligatory	Non-obligatory	Non-obligatory
1vok	Obligatory	Non-obligatory	Obligatory
1wgj	Obligatory	Non-obligatory	Non-obligatory
1xdt	Non-obligatory	Obligatory	Obligatory
1xso	Obligatory	Non-obligatory	Non-obligatory
2ae2	Obligatory	Non-obligatory	Obligatory
2arc	Obligatory	Non-obligatory	Non-obligatory
2hhm	Obligatory	Non-obligatory	Non-obligatory
4mdh	Obligatory	Non-obligatory	Non-obligatory

3.3.1. Obligatory cases defined as Non-obligatory

There are a total of 21 cases which are defined as obligatory in the literature, while the method developed in this thesis classifies them as non-obligatory. 1msp (Bullock *et al.*, 1996) holds an example for these cases. When the cross-correlation graph is studied, it can be observed that the two chains have negative cross-correlation values between them (Figure 3.14) and there seems to be an associating region between the residues 10 and 23. This site is also marked in the cartoon representation of the complex (Figure 3.15), and overlaps well with the fourth slowest mode of chain A as can be observed in Figure 3.16. This protein complex is identified as non-obligatory since an anchoring behaviour among different chains is observed.

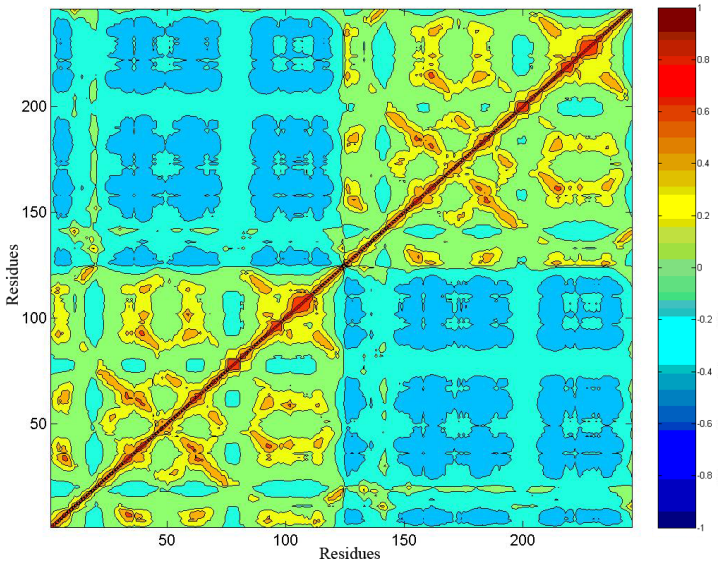

Figure 3.14: Cross-correlation graph for 1msp in all modes

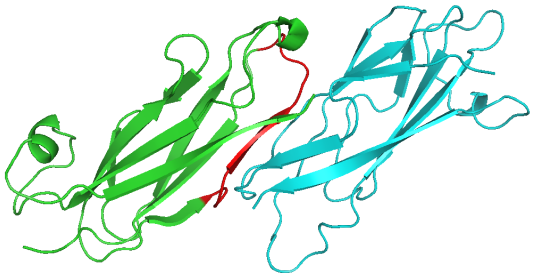

Figure 3.15: Cartoon view for 1msp. Chain A is represented in green, chain B is represented in cyan. The main associating region; residues 10-23 in chain A is colored as red.

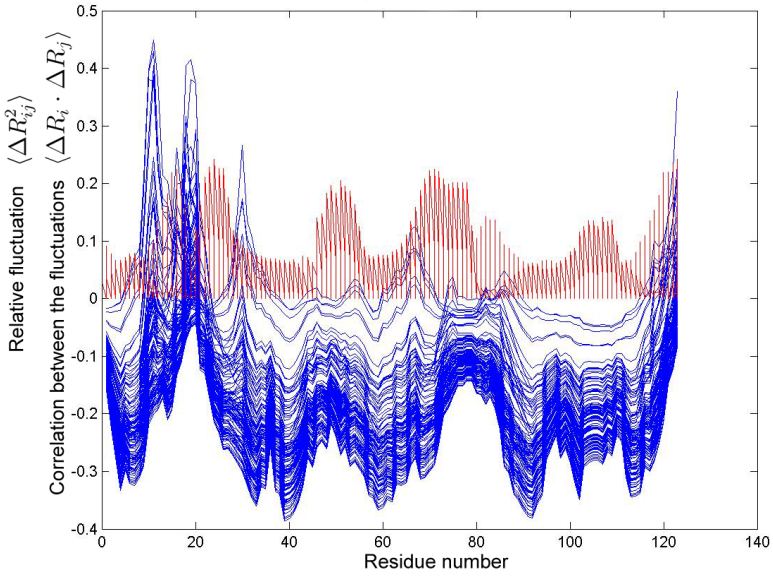

Figure 3.16: The cross-correlation in all modes for the complex and relative fluctuations in fourth slowest mode of the chain A for 1msp, overlapped.

3.3.2. Obligatory cases defined as crystal

1gpe is the only case that is given to be obligatory in the literature, while identified as crystal by the method developed in this thesis. 1gpe (Wohlfahrt *et al.*, 1999) is a homo-dimer with 587x2 residues. When the cross-correlation graph (Figure 3.17) is studied, there is a sharp distinction between the chains, and there are no positive cross-correlations between the two chains. Not having observed any interaction results in this case to be identified as crystal.

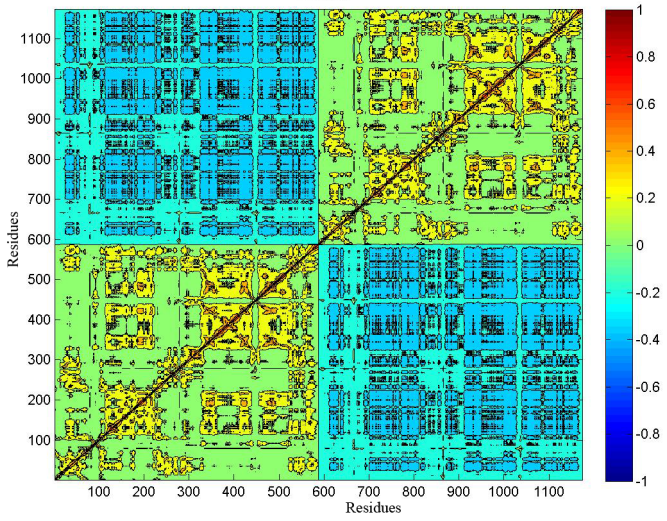

Figure 3.17: Cross-correlation graph for 1gpe in all modes

3.3.3. Non-obligatory cases defined as obligatory

There are two cases that are identified as non-obligatory by the literature, but are defined as obligatory with the method used in this thesis: 1dow and 1xdt.

As to give an example, 1dow will be examined in detail in this section. 1dow is a hetero-dimer with chain A consisting of 200 residues and chain B of 31 (Pokutta *et al.*, 2000). Both chains have only alpha-helices as secondary structure. The cross-correlation graph (Figure 3.18) suggests that there are two domains in chain A. Residues 1-87 are in obligatory interaction with chain B, whereas the other domain which consists of residues 88-200, has no interaction with the other chain at all.

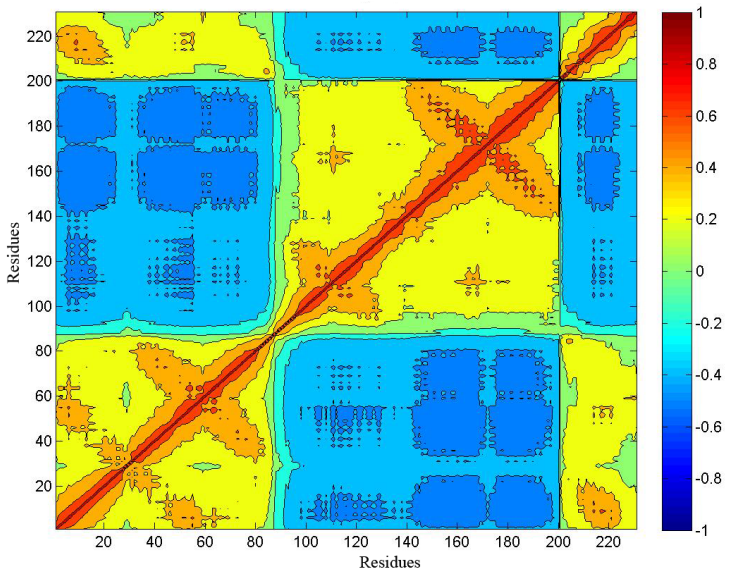

Figure 3.18: Cross-correlation graph for 1dow in all modes

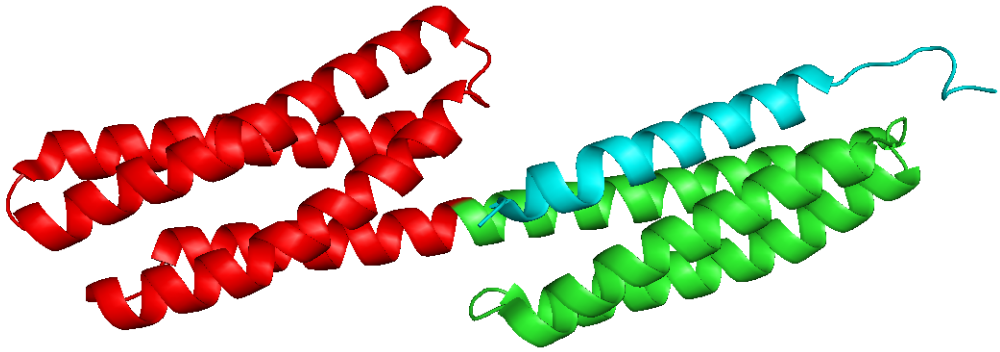

Figure 3.19: Cartoon view for 1dow. Residues 1-87 in chain A are represented in green, 88-200 are represented as red; chain B is represented in cyan.

3.4. CAPRI

CAPRI (Critical Assessment of Prediction of Interactions) is a communitywide protein-protein interaction prediction experiment. For any CAPRI prediction round, the target, which is known only by the author, is aimed to be predicted in its bound form. To start with, the predictors are given the monomers in their unbound states. Since there is a lock-and-key fit at the right docking conformation, the target is expected to be in line with the criteria of non-obligatory classification discussed in this thesis (Janin *et al.*, 2003).

There are two phases of each round of CAPRI: prediction and scoring. PRC has participated in CAPRI in 16^{th} and 17^{th} rounds in the scoring phase. The predictors, with the given unbound conformation, make a prediction of the target, and upload up to 100 complexes for scorers to score. In the second phase, scorers score these complexes and submit ten models as their scoring result. So basically, the scorer's success is dependent on the predictors'.

In the 16^{th} round, a total of 1600 complexes were submitted for the scorers. However, of these complexes, only 34 were correct; and 18 groups of scorers could identify only 29 of them. Using the underlying idea behind this thesis, 6 of the 10 submitted models were scored successfully.

However, in the 17^{th} round, there were no correct predictions to choose from; so none of the scorer groups could score any complex correctly.

For round 16, the main idea was that the associating regions found from the correlation between fluctuations would match with either the relative fluctuations in the slow modes or the relative fluctuations in the fast modes. It was assumed that the relative fluctuations in the slow mode would match with the associating region from the receptor chain, and the relative fluctuations in the fast mode would match with the associating region from the ligand chain.

The cross-correlation graph of T37.S11.M03 (Target 37, Scorer group 11, Model 3) is given in Figure 3.20. There are associating regions at 34-35, 58 and 62. And these regions overlap well with the relative fluctuations in the third slowest mode of chain A (see Figure 3.21). Additionally, in line with the expectations, these residues also match well with the relative fluctuations in the second fast mode of the ligand chain (see Figure 3.22).

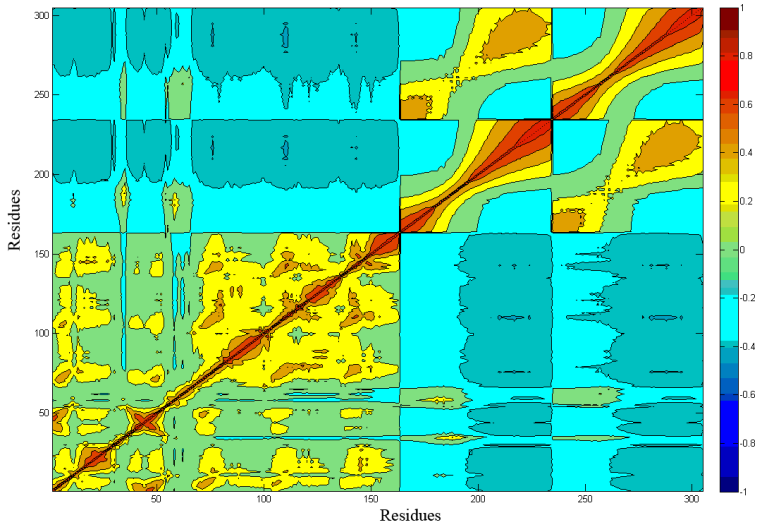

Figure 3.20: Cross-correlation graph for T37.S11.M03 in all modes

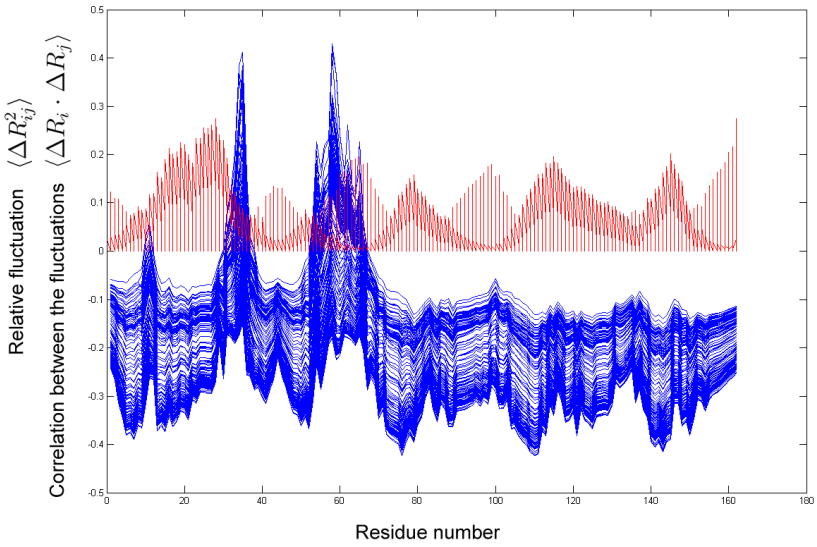

Figure 3.21: The cross-correlation in all modes for the complex and relative fluctuations in third slowest mode of the chain A for T37.S11.M03, overlapped.

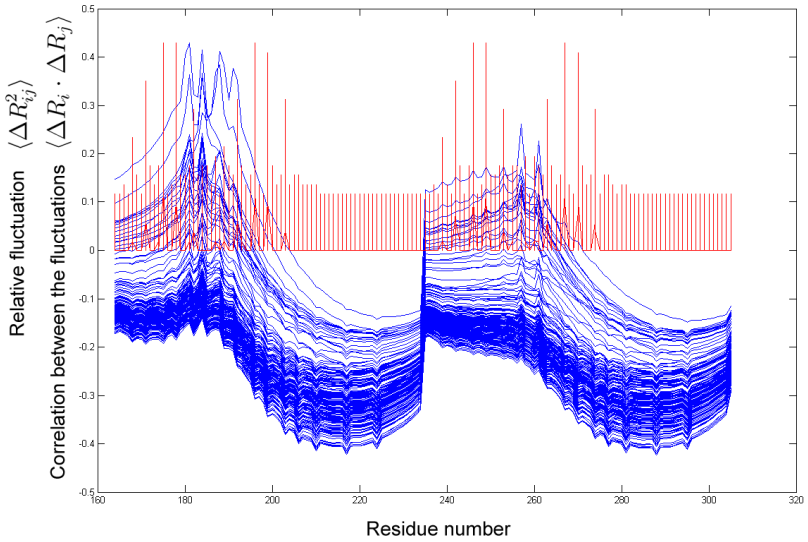

Figure 3.22: The cross-correlation in all modes for the complex and relative fluctuations in second fastest mode of the chain B for T37.S11.M03, overlapped.

4. WEB SERVER

In order to allow the science community to easily access the work done on this thesis, a web server is built. This web server is aimed to be user-friendly, giving results as compact and clear as possible, and being updated constantly according to the feedbacks coming from the users.

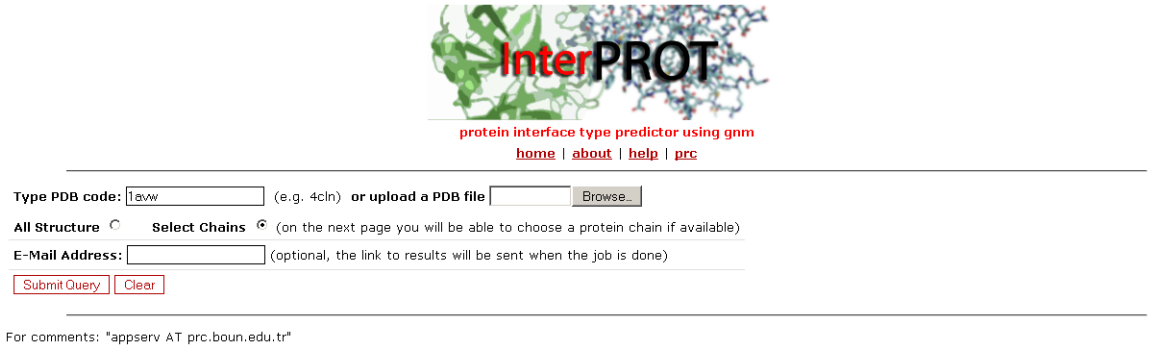

Figure 4.1: Home page for the built web server

The web server's design is kept similar to a previously developed server, HingeProt (Emekli *et al.* 2008), for consistency. The user can upload either their own PDB file, or enter a PDB ID, which will be downloaded from RCSB (Research Collaboratory for Structural Bioinformatics) (See Figure 4.1). The results can either be accessed immediately upon the job's completion, or a notification e-mail can be sent to the user if an e-mail address is supplied. For the calculations, users are able to select chains; so any dimer can be selected from a higher oligomer for the prediction of the interface (See Figure 4.2). Afterwards, making all the necessary calculations, the server's output consists of the type of the interface, the cartoon representation of the given protein and the associating regions. Additionally, if asked, the user can easily access the plausible key regions, anchoring residues, and matching residues along the interface.

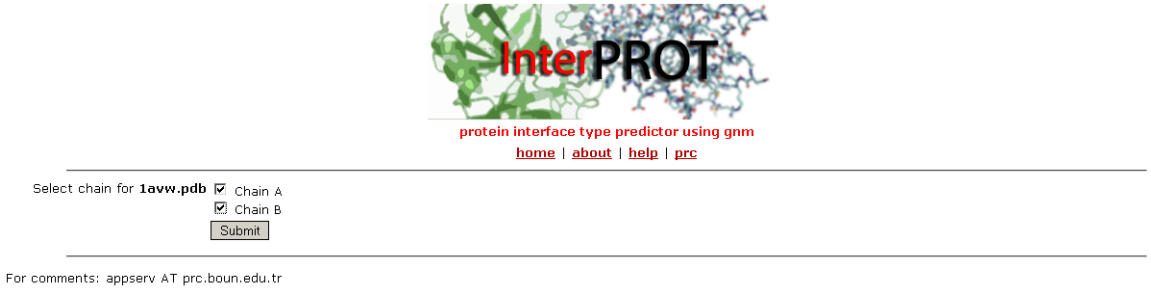

Figure 4.2: Chain selection page for the built web server

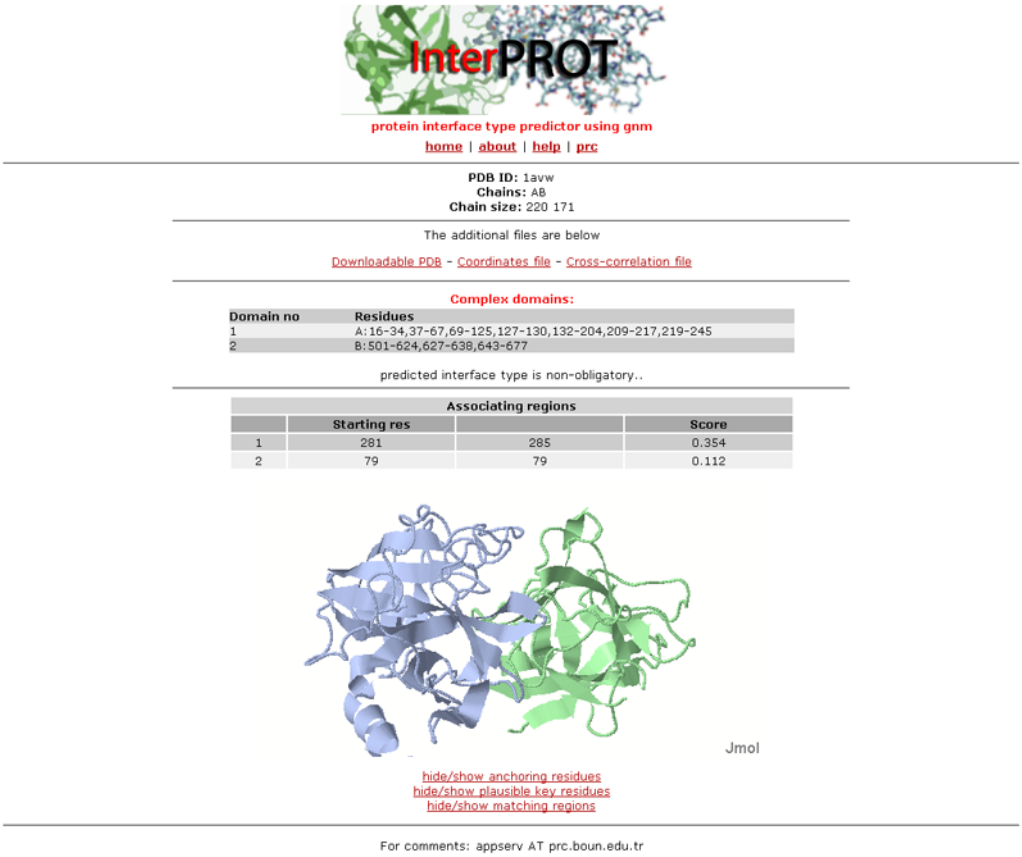

Figure 4.3: Results page for the built web server

The web server can be reached through on http://www.prc.boun.edu.tr/.

5. CONCLUSIONS

The fluctuations across the interface is a dynamic fingerprint of the interface type. The fluctuations of residues in a complex structure and in the isolated states of its chains obtained from the GNM, successfully identify the interface type as non-obligatory, obligatory, or crystal.

The dataset on which the proposed method is tested is PPI-Pred consisting of 111 proteins. The successful prediction rate is 82 percent for the obligatory and 76.5 percent for the non-obligatory; which stands as a powerful tool for scoring the predicted structures.

Moreover, the success of the present approach is proved once more in one of the latest CAPRI (Critical Assessment of Prediction of Interactions) competitions (round 16). The successful evaluation rate for the modelled complex structures was 6/10, which also displays its potential use in future CAPRI predictions.

The success in the prediction supports the premise where the associating regions across the interface in a complex, are hidden intrinsically in the relative fluctuations of its monomers in their isolated states for transient complexes.

A built web server will provide a user-friendly web interface to predict the interface type of any given protein, and the plausible key regions, the anchoring regions across the interface.

APPENDIX A: COMBINED RESULTS

Table A.1: Detailed results for dataset (Bradford and Westhead, 2005). Dataset: Dataset result, Visual: Visual result, Server: Server result, Groups: number of associating groups, Min: minimum chain-chain cross-correlation value, Max: maximum chain-chain cross-correlation value, Ob: Obligatory, Non: Non-obligatory

PDB ID	Dataset	Visual	Server	Groups	Min	Max
1a0f	Ob	Ob	Ob	3	-0.248	0.380
1a4i	Ob	Ob	Non	16	-0.255	0.450
1a4u	Ob	Non	Non	16	-0.255	0.520
1afw	Ob	Ob	Ob	11	-0.178	0.459
1aj8	Ob	Ob	Ob	9	-0.208	0.642
1ajs	Ob	Ob	Ob	13	-0.256	0.539
1aom	Ob	Ob	Ob	18	-0.295	0.452
1aq6	Ob	Ob	Ob	10	-0.231	0.506
1at3	Non	Non	Non	5	-0.423	0.509
1atn	Non	Non	Non	2	-0.474	0.641
1avw	Non	Non	Non	2	-0.365	0.405
1ay7	Non	Non	Non	4	-0.479	0.407
1az3	Ob	Non	Non	4	-0.610	0.723
1b34	Non	Non	Non	5	-0.350	0.287
1b3a	Non	Non	Non	2	-0.469	0.625
1b5e	Ob	Ob	Non	12	-0.174	0.457
1b8j	Ob	Ob	Ob	29	-0.180	0.468
1b9m	Ob	Ob	Ob	0	-0.014	1.000
1bbh	Non	Non	Non	14	-0.252	0.435
1bjn	Ob	Ob	Ob	14	-0.195	0.432
1bkd	Non	Non	Non	4	-0.291	0.517
1bo1	Ob	Ob	Non	8	-0.456	0.604
1buh	Non	Non	Ob	2	-0.420	0.496
1bun	Non	Non	Non	2	-0.517	0.386

Table A.1: Detailed results for dataset, continued

PDB ID	Dataset	Visual	Server	Groups	Min	Max
1bvn	Non	Non	Ob	4	-0.206	0.516
1bw0	Ob	Non	Non	20	-0.248	0.482
1byf	Ob	Ob	Ob	6	-0.318	0.518
1byk	Ob	Ob	Ob	18	-0.330	0.411
1cmb	Ob	Ob	Ob	7	-0.258	0.599
1cnz	Ob	Ob	Ob	12	-0.343	0.627
1coz	Ob	Non	Ob	9	-0.358	0.631
1cp2	Ob	Non	Non	4	-0.357	0.468
1cse	Non	Non	Non	3	-0.247	0.378
1dj7	Non	Non	Non	3	-0.435	0.494
1dor	C	Non	Non	16	-0.237	0.539
1dow	Non	Ob	Ob	2	-0.498	0.662
1dpj	Ob	Ob	Ob	0	-0.140	0.501
1dtd	Non	Non	Non	5	-0.225	0.422
1e0b	Non	C	Non	8	-0.705	-0.256
1efv	Ob	Ob	Ob	10	-0.275	0.644
1euv	Non	Non	Non	4	-0.287	0.484
1f34	Ob	Ob	Ob	7	-0.212	0.457
1f3v	Non	Non	Non	9	-0.291	0.424
1f5m	Ob	Non	Non	12	-0.220	0.469
1f6y	Ob	Ob	Ob	6	-0.327	0.430
1fss	Non	Non	Ob	2	-0.180	0.516
1g4y	Ob	Ob	Ob	9	-0.534	0.650
1gpe	Ob	C	Non	46	-0.354	0.377
1h2a	Ob	Ob	Non	9	-0.138	0.501
1hgx	Ob	Non	Non	9	-0.270	0.442
1hul	Ob	Ob	Ob	5	-0.327	0.608
1hx1	Non	Non	Ob	3	-0.368	0.511
1isa	Ob	Non	Non	19	-0.411	0.282

Table A.1: Detailed results for dataset, continued

PDB ID	Dataset	Visual	Server	Groups	Min	Max
1jkm	Non	Non	Non	26	-0.202	0.407
1kac	Non	Non	Non	7	-0.372	0.427
1kpe	Ob	Ob	Ob	10	-0.208	0.626
1luc	Ob	Ob	Ob	19	-0.272	0.492
1mct	Ob	Ob	Ob	2	-0.136	0.388
1mka	Ob	Ob	Non	7	-0.172	0.513
1msp	Ob	Non	Non	6	-0.386	0.449
1nse	Ob	Ob	Ob	11	-0.271	0.508
1nsy	Ob	Ob	Ob	6	-0.248	0.493
1one	Ob	Ob	Non	24	-0.182	0.414
1pdk	Non	Non	Ob	9	-0.332	0.547
1pnk	Ob	Ob	Ob	5	-0.208	0.743
1pp2	Ob	Non	Non	13	-0.223	0.420
1pvu	Ob	Ob	Ob	3	-0.571	0.691
1qae	Ob	Non	Non	8	-0.386	0.383
1qav	Non	Non	Ob	3	-0.344	0.615
1qax	Ob	Ob	Ob	9	-0.148	0.566
1qbi	Ob	Non	Non	20	-0.275	0.454
1qfe	Non	Non	Non	10	-0.524	0.291
1qfh	Ob	Ob	Ob	3	-0.478	0.605
1qi9	Ob	Ob	Ob	12	-0.178	0.567
1qor	Ob	Ob	Non	9	-0.497	0.623
1qqj	Ob	Non	Non	17	-0.209	0.405
1qu7	Ob	Ob	Ob	0	-0.446	0.684
1smp	Non	Non	Non	5	-0.262	0.489
1smt	Ob	Ob	Ob	5	-0.422	0.657
1sox	Ob	Non	Non	17	-0.271	0.391
1spu	Ob	Ob	Ob	15	-0.139	0.739
1tab	Non	Non	Non	3	-0.178	0.391

Table A.1: Detailed results for dataset, continued

PDB ID	Dataset	Visual	Server	Groups	Min	Max
1tgs	Non	Non	Non	3	-0.188	0.486
1trk	Ob	Non	Non	19	-0.125	0.608
1tx4	Non	Non	Ob	6	-0.279	0.408
1udi	Non	Non	Non	8	-0.282	0.418
1vfr	Ob	Ob	Ob	6	-0.250	0.604
1vhi	Ob	Ob	Non	6	-0.257	0.479
1viw	Non	Non	Non	8	-0.257	0.415
1vlt	Ob	Ob	Non	3	-0.567	0.688
1vok	Ob	Non	Ob	7	-0.333	0.324
1vsg	Ob	Ob	Ob	4	-0.287	0.596
1wgj	Ob	Non	Non	12	-0.385	0.483
1xdt	Non	Ob	Ob	1	-0.661	0.877
1xik	Ob	Ob	Ob	13	-0.243	0.715
1xso	Ob	Non	Non	13	-0.297	0.331
1ypi	Non	Non	Non	6	-0.336	0.495
2aai	Ob	Ob	Ob	7	-0.350	0.679
2ae2	Ob	Non	Ob	8	-0.311	0.387
2arc	Ob	Non	Non	8	-0.529	0.406
2gsa	Ob	Ob	Ob	10	-0.144	0.463
2hdh	Ob	Ob	Ob	2	-0.432	0.446
2hhm	Ob	Non	Non	19	-0.256	0.317
2nac	Ob	Ob	Ob	9	-0.258	0.488
2pfl	Ob	Ob	Ob	25	-0.348	0.410
2sic	Non	Non	Non	4	-0.287	0.372
2utg	Ob	Ob	Ob	4	-0.289	0.412
3ygs	Non	Non	Non	3	-0.474	0.467
4mdh	Ob	Non	Non	6	-0.296	0.529
4sgb	Non	Non	Non	6	-0.208	0.345
7cei	Non	Non	Ob	7	-0.371	0.345

REFERENCES

Bahar, I., A. Atilgan and B. Erman, 1997. "Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential". *Fold. Des.*, Vol. 2, pp. 173-181.

Bahar, I., A. R. Atilgan, M. C. Demirel and B. Erman, 1998, "Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability", *Physical Review Letters*, Vol. 80, pp. 2733-2736.

Bahar, I. and R. L. Jernigan, 1998, "Vibrational dynamics of transfer RNAs. Comparison of the free and enzyme-bound forms" *J. Mol. Biol.* Vol. 281, pp. 871-884.

Bairoch, A. and R. Apweiler, 2000, "The SWISS-PROT protein sequence data bank and its supplement TrEMBL" *Nucl. Acids Res.* Vol. 28, pp. 45-48.

Ben-Dor, A., R. Shamir and Z. Yakhini, 1999 "Clustering gene expression patterns", *J. Comput. Biol.* Vol. 6(3/4) pp. 281297.

Benach J., S. Atrian, R. Gonzalez-Duarte and R. Ladenstein, 1998, "The refined crystal structure of Drosophila lebanonensis alcohol dehydrogenase at 1.9 Å resolution", *J. Mol. Biol.*, Vol. 282, pp. 383-399.

Bode, W., E. Papamokos and D. Musil, 1987. "The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech Hirudo medicinalis. Structural analysis, subtilisin structure and interface geometry.", *Eur. J. Biochem.*, Vol. 166, pp. 673-692.

Bradford J.R and D. R. Westhead, 2005, "Improved prediction of proteinprotein binding sites using a support vector machines approach", *Bioinformatics* Vol. 21, pp. 1487-1494.

Bullock T. L., T. M. Roberts and M. Stewart, 1996, "2.5 Å resolution crystal structure of the motile major sperm protein (MSP) of Ascaris suum.", *J. Mol. Biol.*, Vol. 263, pp. 284-296.

Camacho C. J. and S. Vajda, 2001, "Protein docking along smooth association pathways", *PNAS*, Vol. 98, pp. 10636-10641.

Cowieson N. P., J. F. Partridge, R. C. Allshire, P. J. McLaughlin, 2000, "Dimerisation of a chromo shadow domain and distinctions from the chromodomain as revealed by structural analysis" *Curr. Biol.*, Vol. 10, pp. 517-525.

Dai, S., C. Schwendtmayer, P. Schrmann, S. Ramaswamy, and H. Eklund , 2000, "Redox Signaling in Chloroplasts: Cleavage of Disulfides by an Iron-Sulfur Cluster", *Science*, Vol. 287, pp. 655-668.

De S., O. Krishnadev, N. Srinivasan and N. Rekha, 2005, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different", *BMC Structural Biology*, Vol. 5, pp. 15-30.

Emekli U., D. Scheidman-Duhovny, H. J. Wolfson, R. Nussinov and T. Haliloglu, "HingeProt: Automated Prediction of Hinges in Protein Structures?", *Proteins* Vol. 70, pp. 1219-1227.

Feymann D., J. Down, M. Carrington, I. Roditi, M. Turner and D. Wiley, 1990, "2.9 Å resolution structure of the N-terminal domain of a variant surface glycoprotein from Trypanosoma brucei.", *J. Mol. Biol.*, Vol. 216, pp. 141-160.

Grant, A., D. Lee and C. A. Orengo, 2004, "Progress towards mapping the universe of protein folds", *Genome Biol.*, Vol. 5, pp. 107

Haliloglu, T., I. Bahar and B. Erman, 1997, "Gaussian dynamics of folded proteins", *Physical Review Letters*, Vol. 79, pp. 3090-3093.

Haliloglu, T., O. Keskin, B. Erman and R. Nussinov, 2005, "How similar are protein

folding and protein binding nuclei? Examination of vibrational motions of energy hotspots and conserved residues", *Biophys. J.*, Vol. 88, pp. 1552-1559.

Haliloglu, T., E. Seyrek and B. Erman, 2008, "Prediction of Binding Sites in Receptor-Ligand Complexes with the Gaussian Network Model", *Physical Review Letters*, Vol. 100, 228102.

Haliloglu, T. and B. Erman, 2009, "Analysis of Correlations between Energy and Residue Fluctuations in Native Proteins and Determination of Specific Sites for Binding", *Physical Review Letters*, Vol. 102, pp. 088103.

Hoog S. S., W. W. Smith, X. Qiu, C. A. Janson, B. Hellmig, M. S. McQueney, K. O'Donnell, D. O'Shannessy, A. G. DiLella, C. Debouck and S. S. Abdel-Meguid, 1997, "Active Site Cavity of Herpesvirus Proteases Revealed by the Crystal Structure of Herpes Simplex Virus Protease/Inhibitor Complex", *Biochemistry*, Vol. 36, pp. 14023-14029.

Janin J., K. Henrick, J. Moult, L. T. Eyck, M. J. E. Sternberg, S. Vajda, I. Vakser and S. J. Wodak, 2003, "CAPRI: A Critical Assessment of PRedicted Interactions", *Proteins*, Vol. 52, pp. 2-9.

Jones, S. and J. M. Thornton, 1996, "Principles of protein-protein interactions", *Proc Natl Acad Sci USA*, Vol. 93, pp. 13-20.

Kabsch W., H. G. Mannherz, D. Suck, E. Pai and K. C. Holmes, 1990, "Atomic structure of the actin: DNase I complex.", *Nature*, Vol. 347, pp. 37-44.

Knuth, D. 1998, "The Art of Computer Programming: Sorting and Searching", Ed. 2, Vol. 3. *Addison-Wesley.*

Kundu, S., D. C. Sorensen, G. N. Phillips, 2004, "Automatic Domain Decomposition of Proteins by a Gaussian Network Model", *Proteins*, Vol. 57, pp. 725-733.

Liu, S., Q. Li and L. Lai, 2006, "A combinatorial score to distinguish biological and

nonbiological protein-protein interfaces", *Proteins* vol. 64, pp.68-78.

Lolis, E., T. Alber, R. C. Davenport, D. Rose, F. C. Hartman and G. A. Petsko, 1990, "Structure of yeast triosephosphate isomerase at 1.9-.ANG. resolution", *Biochemistry*, Vol. 29, pp. 6609-6618.

Marques, O. A., 1995, "BLZPACK: Description and Users guide", Technical Report TR/PA/95/30, CERFACS.

McCoy A. J., P. Fucini, A. A. Noegel and M. Stewart, 1999, "Structural basis for dimerization of the dictyostelium gelation factor (ABP120) rod", *Nat. Struct. Biol*, Vol. 6, pp. 836-841.

Milburn, M. V., A. M. Hassell, M. H. Lambert, S. R. Jordan, A. E. Proudfoot, P. Graber and T. N. Wells, 1993 "A novel dimer configuration revealed by the crystal structure at 2.4 A resolution of human interleukin-5.", *Nature*, Vol. 363, pp. 172-176.

Murzin, A.G., S.E. Brenner, T. Hubbard and C. Chothia, 1995, "SCOP: A structural classification of proteins database for the investigation of sequences and structures", *J. Mol. Biol.*, Vol. 247, pp. 536-540.

Neuvirth, H., R. Raz and G. Schreiber, 2004, "ProMate: A structure based prediction program to identify the location of protein-protein binding sites", *J. Mol. Biol.*, Vol. 338, pp. 181-199.

Nooren, I. M. and J. M. Thornton, 2003, "Diversity of protein-protein interactions", *EMBO J.*, Vol. 22, pp. 3486-3492.

Ofran, Y. and B. Rost, 2003, "Analysing six types of protein-protein interfaces", *J. Mol. Biol.*, Vol. 1325, pp. 377-387.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, 1997, "CATHA hierarchic classification of protein domain structures", *Structure*,

Vol. 5, pp. 10931108.

Pokutta S. and W. I. Weiss, 2000, "Structure of the dimerization and beta-catenin-binding region of alpha-catenin.", *Mol. Cell.*, Vol. 5, pp. 533-543.

Rader, A. C., C. Chennubhotla, L. W. Yang, I. Bahar, ""The Gaussian Network Model: Theory and Applications" in "Normal Mode Analysis. Theory and Applications to Biological and Chemical Systems" *Chapman Hall CRC Press* pp. 41-64.

Redfern, O. C., A. Harrison, T. Dallman, F. M. G. Pearl and C. A. Orengo, 2007, "CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures", *PLoS Computational Biology*, 3(11).

Reeves, G.A., T. J. Dallman, O. C. Redfern, A. Akpor and C. A. Orengo, 2006, "Structural diversity of domain superfamilies in the CATH database", *J. Mol. Biol.*, Vol. 360, pp. 725-741.

Ren Z., T. Meyer, D. E. McRee, 1993, "Atomic structure of a cytochrome c' with an unusual ligand-controlled dimer dissociation at 1.8 A resolution.", *J. Mol. Biol.*, Vol. 234, pp. 433-445.

Song, H.K. and S. W. Suh, 1998, "Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from Erythrina caffra and tissue-type plasminogen activator", *J. Mol. Biol.*, Vol. 275, pp. 347-363.

Rowland P., F. s. Nielsen, K. F. Jensen and S. Larsen, 1997, "The crystal structure of the flavin containing enzyme dihydroorotate dehydrogenase A from Lactococcus lactis" *Structure*, Vol 5, pp. 239-252.

Stryer, L. 1995, "Signal transduction cascades", Biochemistry *Freeman and Company* ed. 4, pp. 325-360

Valdar W.S. and J. M. Thornton, 2001, "Protein-protein interfaces: analysis of amino acid conservation in homodimers", *Proteins*, Vol. 42, pp. 108-124.

Wilken J., D. Hoover, D. A. Thompson, P. N. Barlow, H. McSparron, L. Picard, A. Wlodawer and J. Lubkowski, 1999, "Total chemical synthesis and high-resolution crystal structure of the potent anti-HIV protein AOP-RANTES", *J. Mol. Biol.*, Vol. 6, pp. 43-51.

Wei Y., J. A. contreras, P. Sheffield, T. Osterlund, U. Derewanda, R. E. Kneusel, U. Matern, C. Holm, Z. S. Derewanda, 1999, "Crystal structure of brefeldin A esterase, a bacterial homolog of the mammalian hormone-sensitive lipase.", *Nat. Struct. Biol.*, Vol. 6, pp. 340-345.

Winter C., A. Henschel, W. K. Kim and M. Schroeder, 2006, "SCOPPI: a structural classification of protein-protein interfaces.", *Nucleic Acids Res.*, Vol. 34, pp. 310-314.

Wohlfahrt G., S.Witt, J. Hendle, D. Schomburg, H. M. Kalisz and H. J. Hecht, 1999, "1.8 and 1.9 A resolution structures of the Penicillium amagasakiense and Aspergillus niger glucose oxidases as a basis for modelling substrate complexes.", *Acta Crystallogr D Biol Crystallogr*, Vol. 55, pp. 969-977.

Zhu, H., F. S. Domingues, I. Sommer and T. Lengauer, 2006, "NOXclass: prediction of protein-protein interaction types", *BMC Bioinf.*, Vol. 7, pp. 27-41.