united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR/1499 - 14

INTERNATIONAL WORKSHOP ON PROTEOMICS: PROTEIN STRUCTURE, FUNCTION AND INTERACTIONS

(5 - 16 May 2003)

"Characterization of protein interfaces: a graph theoretic approach"

presented by:

S. Vishveshwara
Indian Institute of Science, Bangalore
India

Characterization of Protein Interfaces : A Graph Theoretic Approach

Saraswathi Vishveshwara

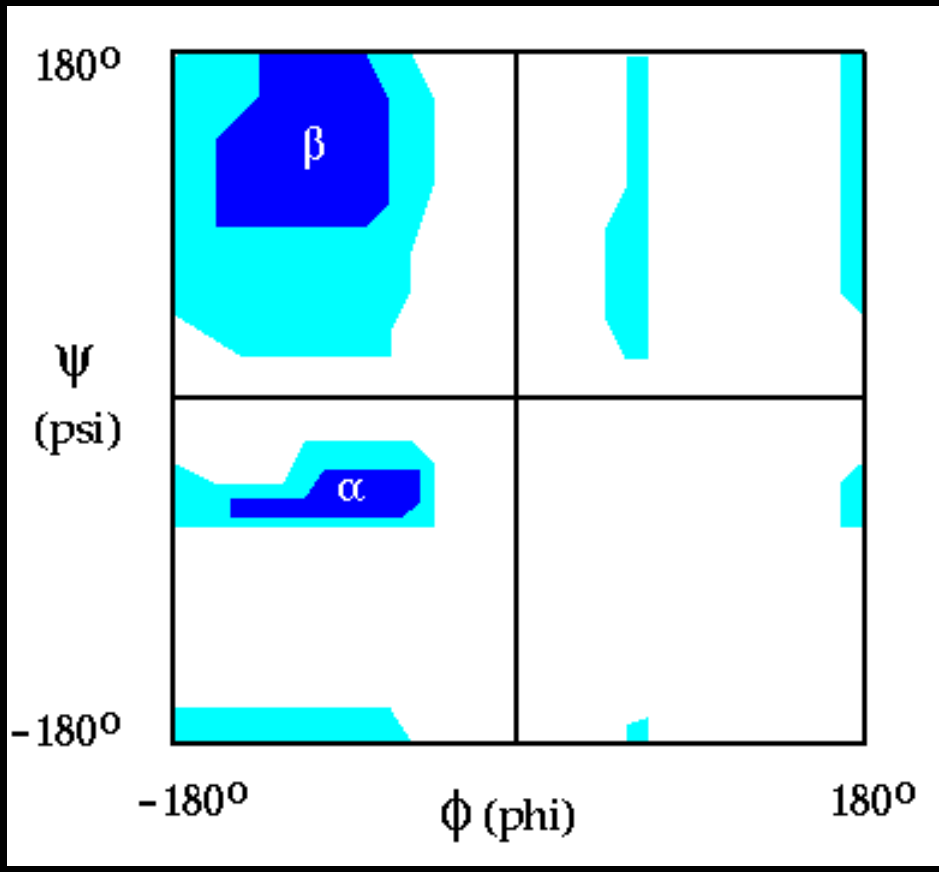
Indian Institute of Science

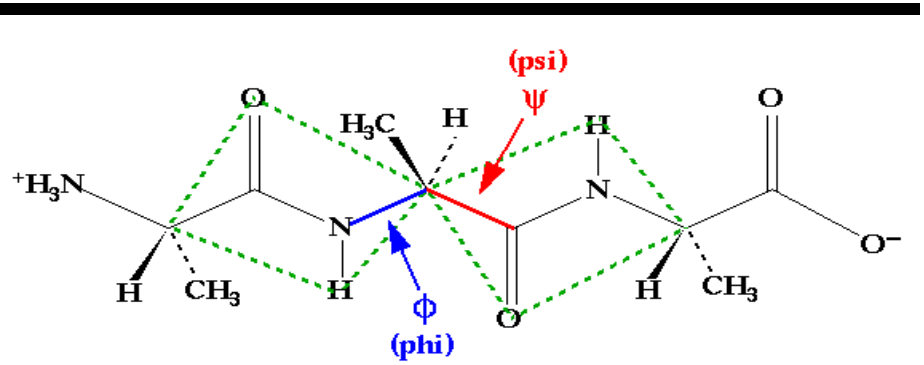
Bangalore, India

G. N. Ramachandran

Molecular Biophysics Unit

….and his Φ-ψ Map





Outline:

- Introduction to protein graphs
- Side chain cluster detection in proteins from graph spectral analysis
- Characterization of protein-protein interfaces in:

  a set of homodimers

  RNA polymerase

- Summary

Investigations on Protein-Protein Interactions

- Multi-subunit proteins, protein-receptor complexes, antigen-antibody complexes, signal transduction proteins etc.

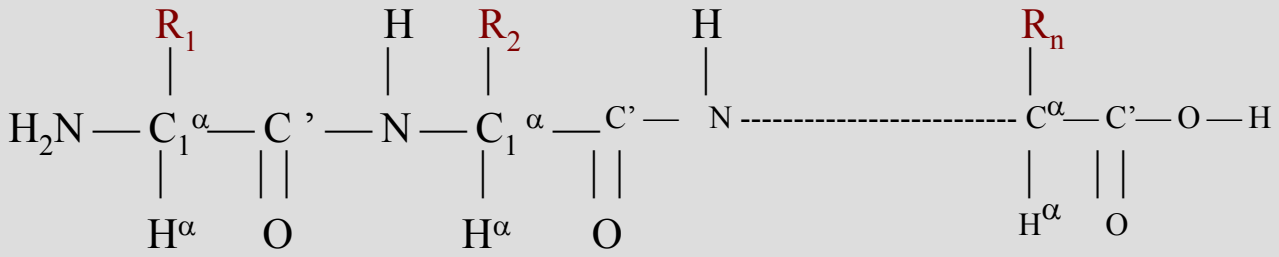
- Chothia & Janin : Difference in accessible surface area

- Thornton's Group : Surface patch analysis (surface and charge complementarities, hydrophobic patches & geometric properties)

- Glaser et al., Bogan & Thorn, Valdar & Thornton : Amino acid preferences at interfaces

- Sternberg et al., : Protein-protein docking using energy considerations.

Levels of Interactions in Proteins



Primary Structure

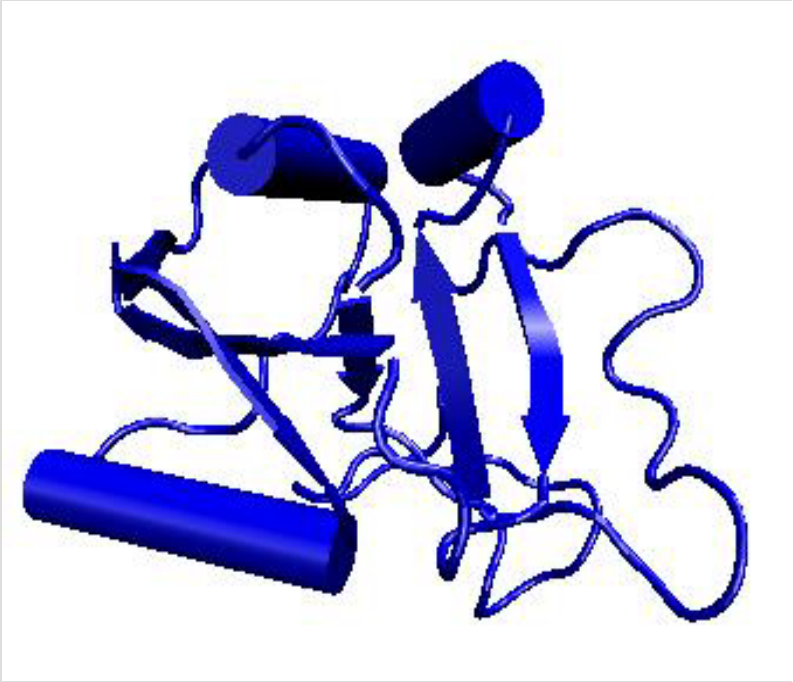
Secondary Structure

Types of interactions :

- Covalent
- Non-covalent
- Hydrogen bond

Types of Neighbours :

- Sequence
- Spatial



Tertiary Structure

Concepts of Graph Theory for Protein Structure Analysis

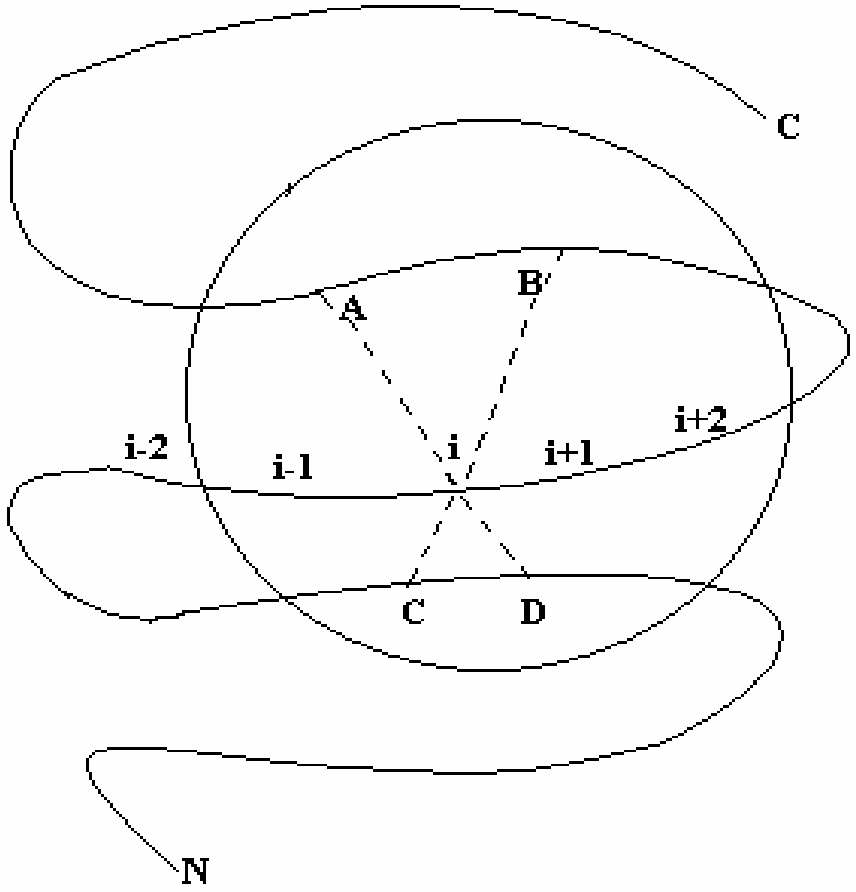
Nodes	Edges	Graph Operation	Purpose	References
Secondary structure (α-helix, β-strand)	Spatially Close Secondary Structures	Identification of Subgraph Isomorphism	Fold & Pattern identification	Mitchell et al., 1989; Grindley et al., 1993;
Secondary structure (α-helix, β-strand)	Spatially Close Secondary Structures (dynamically arrived at)	Dynamical Matrix construction	Testing Folding Rules	Przytycka et al.,2002
Side Chain	Spatial Proximity	Identification of Subgraph Isomorphism	Functionally & Structurally important motif recognition	Artymiuk et al.,1994
Side Chain	Spatial Proximity decided by overlap cut off criterion (Weighted edge)	Graph Spectra, Identification of Clusters and Cluster centers	Identification of clusters important for function, structure and folding	Kannan & Vishveshwara, 1999
Backbone	Spatial neighbours within radius cut off (6.5 - 7.0 Å)	Graph Spectra, Identification of Clusters and Cluster centers	Identification of Proteins with similar folds	Patra & Vishveshwara., 2000
Backbone	Spatial neighbours within radius cut off (7.0 Å)	Graph Spectra	Protein Dynamics	Bahar, 1999
All Atoms	Defined based on Constraints (Weighted Edge)	Graph Spectra	Protein Dynamics	Jacobs et al., 2001

Graph Spectral Method:

- Definition of nodes and edges
- Construction of Adjacency and Laplacian (Kirchoff's) Matrix
- Solution to the Matrix
- Graph Spectra: Eigenvalues and Eigenvector components
- Clusters – From Second lowest Eigenvalue and its vector components.
- Cluster centres – Vector components from the highest eigenvalue corresponding to the chosen cluster

Protein Graphs:

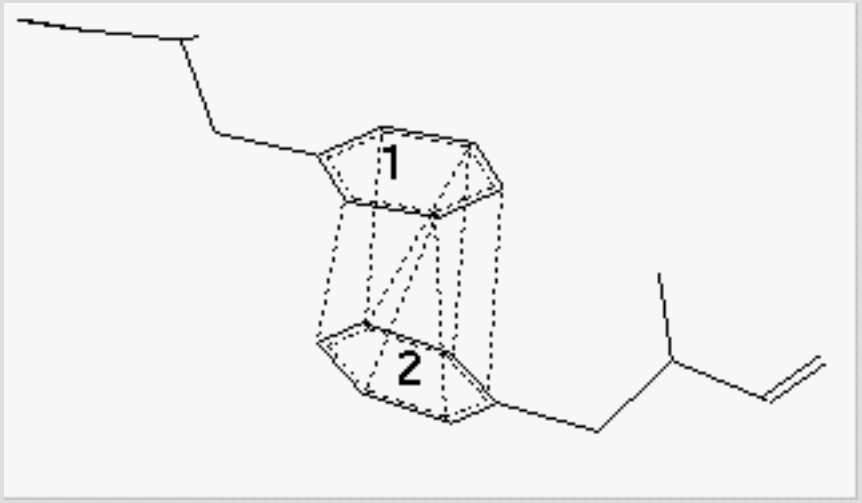
Main Chain Interaction (back bone level)



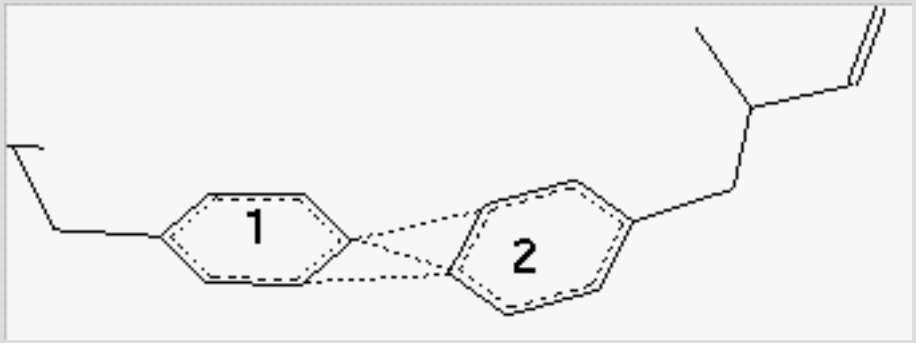
- Nodes : Residues
- Edges : Based on interaction criterion

Patra & Vishveshwara, Biophys. Chem 84, 13 (2000)

Side Chain Interaction

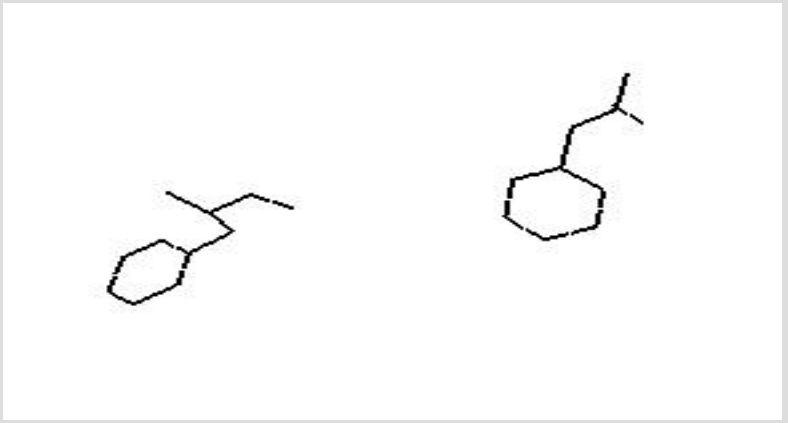


a) High Contact (8.57%)



b) Low Contact (3.21%)

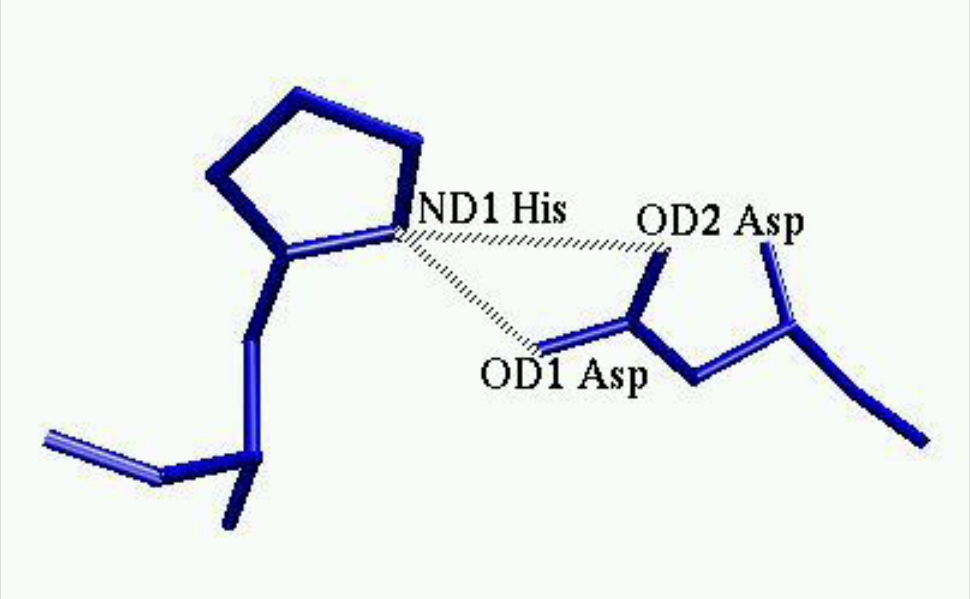
c) No interaction



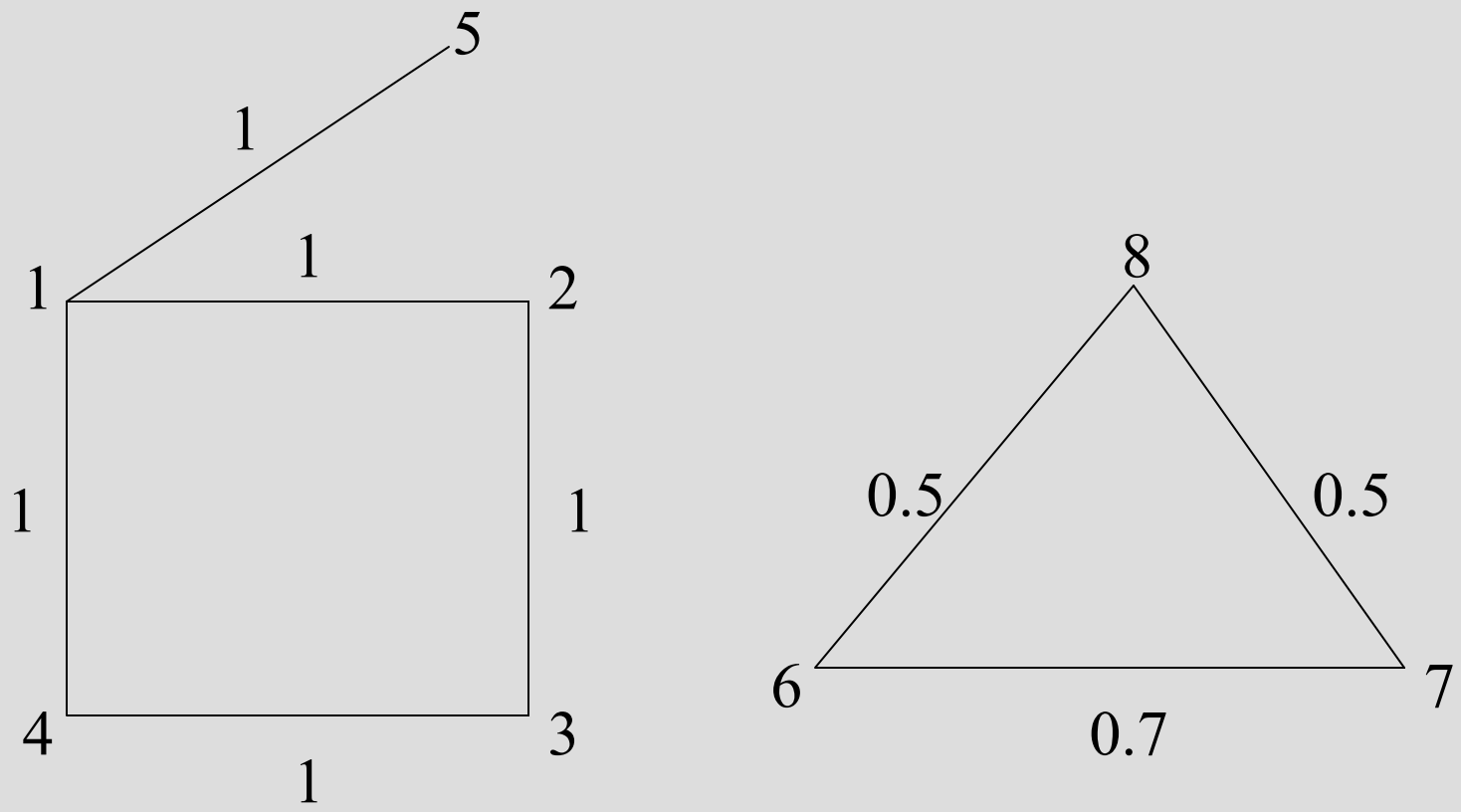
Contact Criterion for obtaining side chain clusters

High and low contact criteria. Two pairs of phenylalanine rings interacting with each other are shown. The dotted lines between the phenylalanines indicate the atoms that are within a distance of 4.5Å. a) high contact (8.57%) ; b) low contact (3.21%); c) No interaction.

Hydrogen Bond Interaction



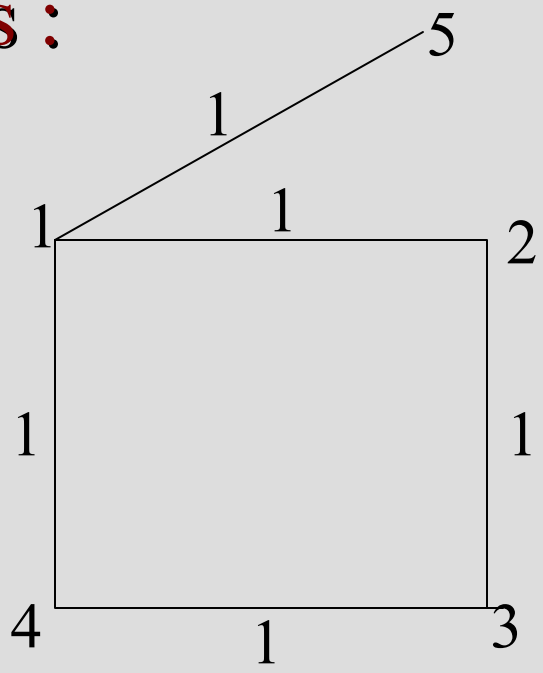
Hydrogen bond between the side chains of amino acids Histidine and Aspartate. Hydrogen bonds can be formed between side-chain or backbone polar atoms, which come within a distance cutoff of about 3Å.



a) Unweighted graph b) Weighted Graph

A schematic representation of a Graph

Matrix representation of graphs :



Adjacency Matrix of Graph 1a :

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Degree matrix :

$$D = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Laplacian matrix :

$$L = D - A$$

$$L = \begin{bmatrix} 3 & -1 & 0 & -1 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & -1 & 2 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Graph Spectra :

The Laplacian Matrix is diagonalized to yield Eigenvalues and Eigenvector components.

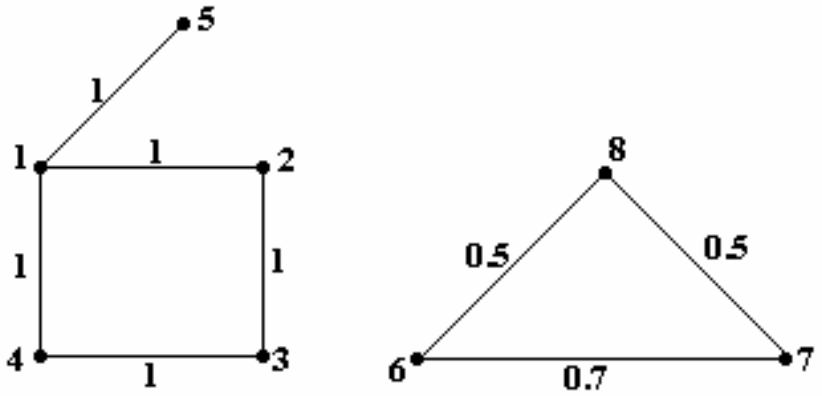
The second smallest eigenvalue yields the cluster information.

Hall K.M, *Manag Sci*, **17**, 219 (1970)

The top eigenvalues give information regarding cluster centres.

Kannan& Vishveshwara, J.Mol.Biol **292**, 441 (1999)

Eigen Spectra of the Laplacian Matrix of graph

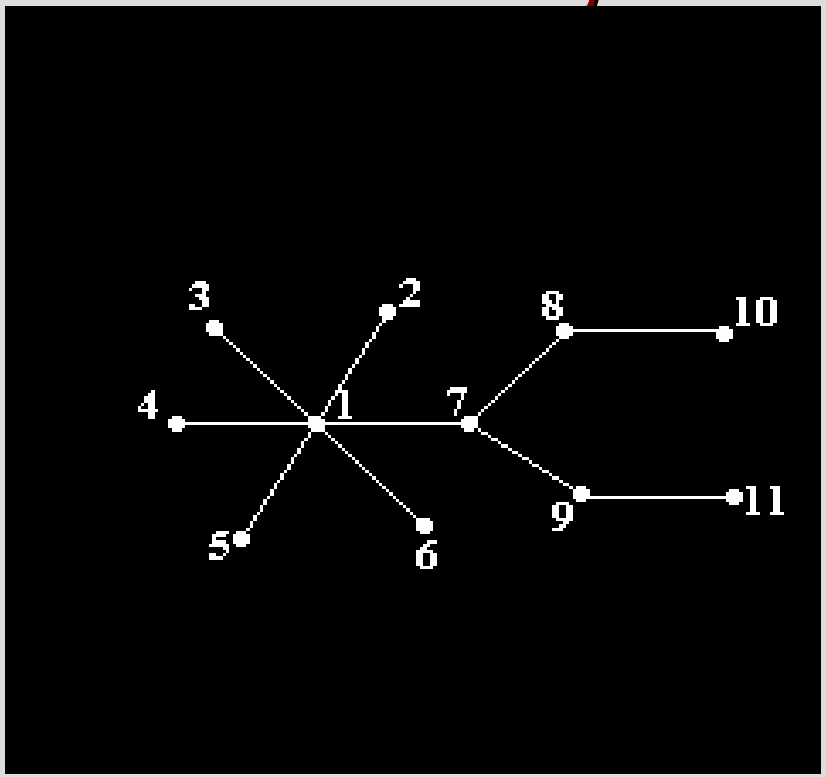


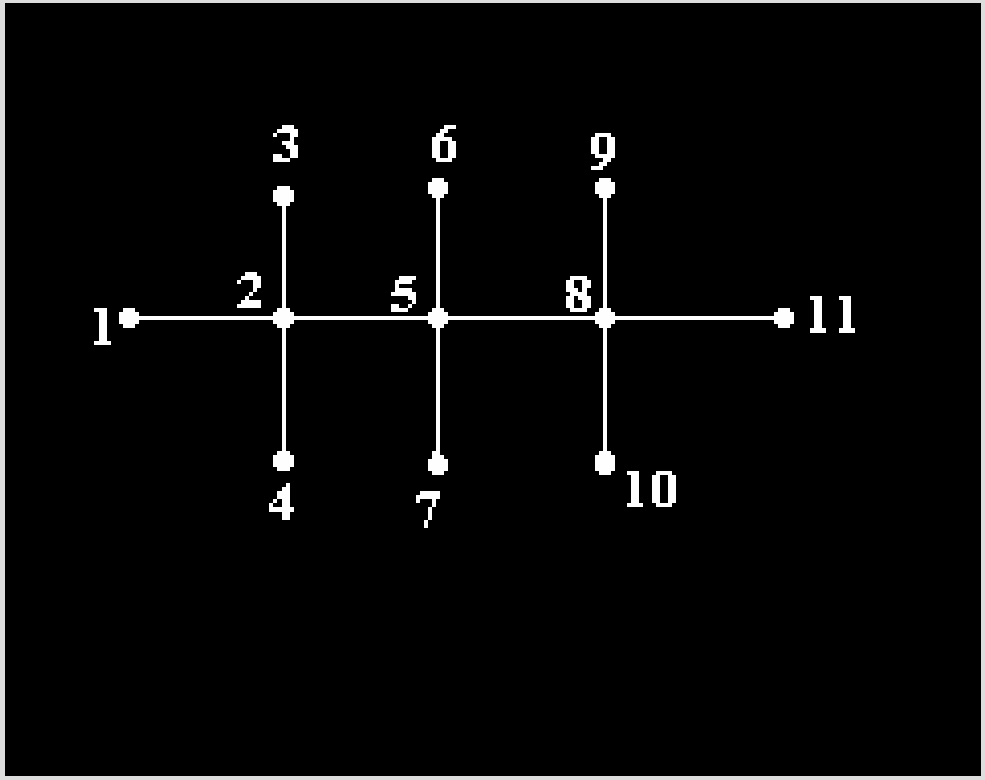
Eigen-values	0.0000	0.0800	0.9016	1.5500	1.9500	2.0600	2.7420	4.5164
Node	EVC[1]	EVC	EVC	EVC	EVC	EVC	EVC	EVC
1	0.3536	0.2739	0.1380	-0.0000	0.0000	0.0000	0.5362	**0.7024**[2]
2	0.3536	0.2739	-0.2560	-0.0000	-0.0000	0.7071	0.2422	-0.4193
3	0.3536	0.2739	-0.4375	-0.0000	-0.0000	-0.0000	-0.7031	0.3380
4	0.3536	0.2739	-0.2560	-0.0000	-0.0000	-0.7071	0.2422	-0.4193
5	0.3536	0.2739	0.8115	0.0000	-0.0000	-0.0000	-0.3175	-0.2018
6	0.3536	-0.4564	-0.0000	-0.4082	**-0.7071**	0.0000	0.0000	0.0000
7	0.3536	-0.4564	0.0000	-0.4082	**0.7071**	-0.0000	-0.0000	-0.0000
8	0.3536	-0.4564	-0.0000	0.8165	-0.0000	0.0000	0.0000	0.0000

[1] : Eigenvector Components

[2] : Largest vector components of top eigenvalues of each cluster are shown in **blue**

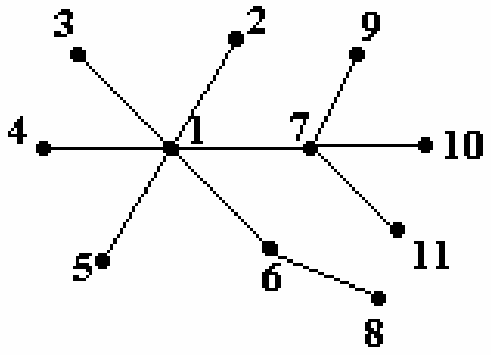
Vector component & Cluster Centre :

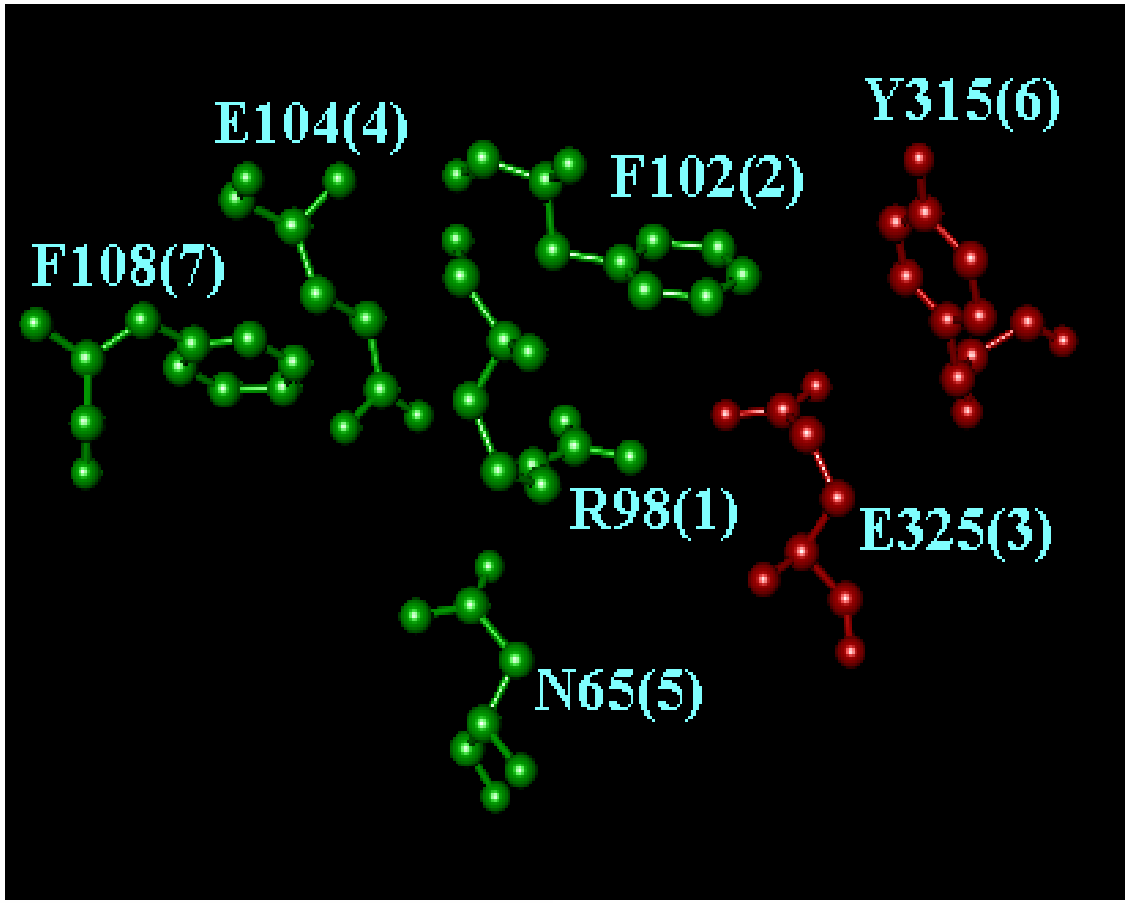




- The node with the highest degree has the highest vector component magnitude in the top eigenvalue.
- In case there are many nodes of the same degree, then the degrees of the nodes adjacent to these nodes are considered.
- Further, the one closest to the geometric centre of the graph gets the preference.

Vishveshwara, Brinda, Kannan JTCC ,**1**, 187 (2002)



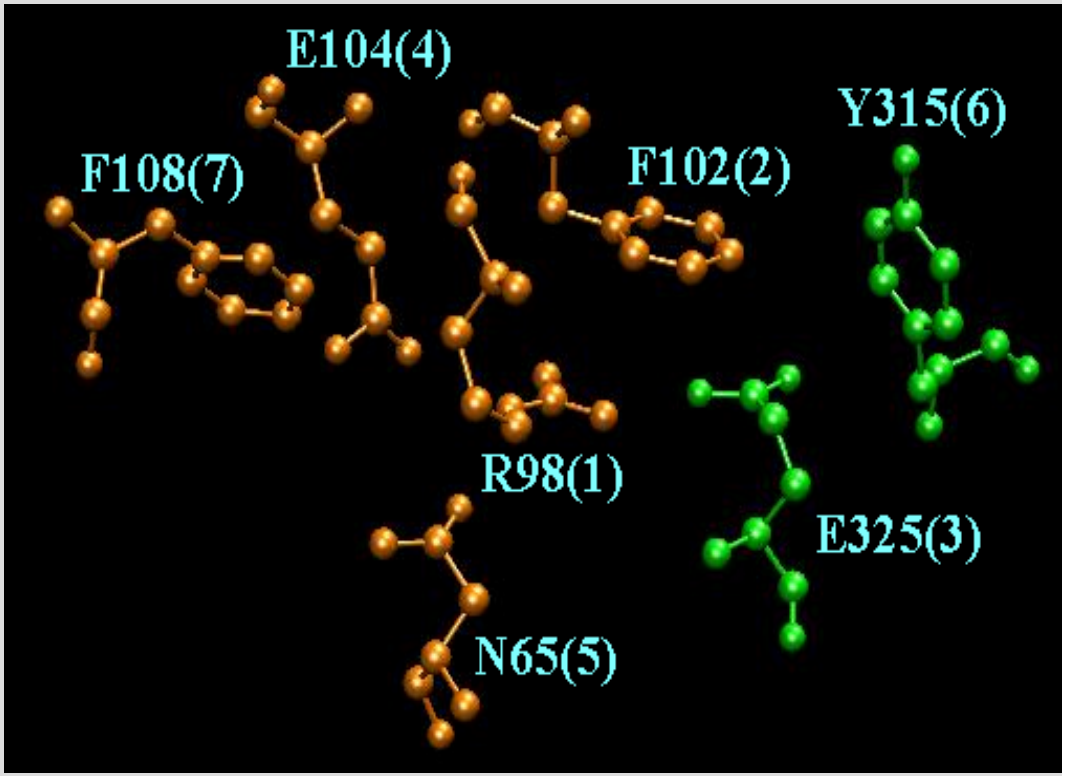


$$E(v) = \max_{v_i \in G} d(v, v_i)$$

Centre of a Graph

Role of Vector Components :

Residue	Vector Component	
	HEV	2nd LEV
F108(A)	0.055	0.146
Y315(B)	0.069	0.146
N65(A)	0.143	0.146
E104(A)	0.196	0.146
E325(B)	0.400	0.146
F102(A)	0.598	0.146
R98(A)	0.645	0.146



•The vector component of a node can be interpreted as a direct measure of the contribution of that node to the stability of the cluster and to the overall connectivity of the graph.

• The contributions of the nodes in the cluster decreases as we move away from the centre of any cluster and this is reflected in the magnitude of the vector components in largest eigenvalue of the cluster.

Interaction Criteria : (for edge formation)

$$INT(R_i,R_j) = N(R_i,R_j) / Norm(Restype(R_i)) \times 100$$

where $N(R_i,R_j)$ is the number of distinct interacting pairs of side chains atoms between the residues R_i and R_j. A distance cutoff of 4.5 Å is used.

$$Norm(Restype(R_i)) = \left(\sum_{K=1}^{p} Maxm\ (Type(R_{ik}))\right) / p$$

S.No.	Residue Type	Norm
1.	Alanine	55.7551
2.	Arginine	93.7891
3.	Asparagine	73.4097
4.	Aspartic acid	75.1507
5.	Cystine	54.9528
6.	Glutamine	78.1301
7.	Glutamic acid	81.8288
8.	Glycine	47.3129
9.	Histidine	83.7357
10.	Isoleucine	67.9452
11.	Leucine	72.2517
12.	Lysine	69.6096
13.	Methionine	69.2569
14.	Phenyl Alanine	93.3082
15.	Proline	51.3310
16.	Serine	61.3946
17.	Threonine	63.7075
18.	Trptophan	106.703
19.	Tyrosine	100.719
20.	Valine	62.3673

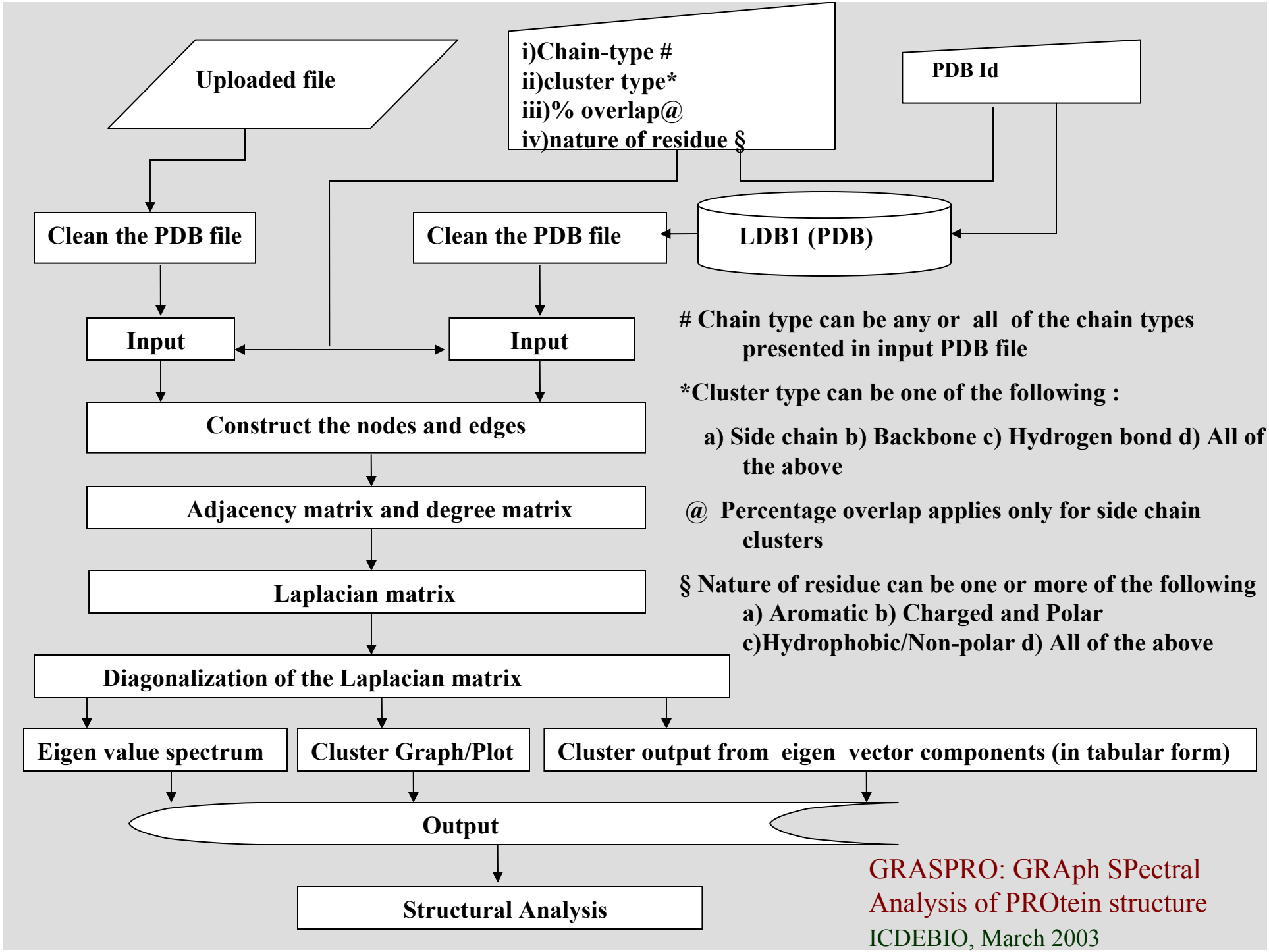
Normalisation values for amino acid residue types

Matrix Construction :

$a_{ij} = 1/d_{ij}$ if i and j are connected

$= 1/100$, otherwise

d_{ij} = distance between the C^{β} (C^{α} in case of Glycine) of i & j


Uploaded file
i)Chain-type #
ii)cluster type*
iii)% overlap@
iv)nature of residue §
PDB Id
Clean the PDB file
Clean the PDB file
LDB1 (PDB)
Input
Input
Construct the nodes and edges
Adjacency matrix and degree matrix
Laplacian matrix
Diagonalization of the Laplacian matrix
Eigen value spectrum
Cluster Graph/Plot
Cluster output from eigen vector components (in tabular form)
Output
Structural Analysis
Chain type can be any or all of the chain types presented in input PDB file
*Cluster type can be one of the following :
a) Side chain b) Backbone c) Hydrogen bond d) All of the above
@ Percentage overlap applies only for side chain clusters
§ Nature of residue can be one or more of the following
a) Aromatic b) Charged and Polar
c)Hydrophobic/Non-polar d) All of the above
GRASPRO: GRAph SPectral Analysis of PROtein structure
ICDEBIO, March 2003

Applications

A) Clusters of Importance:

a) Active/Binding site

b) Domain identification

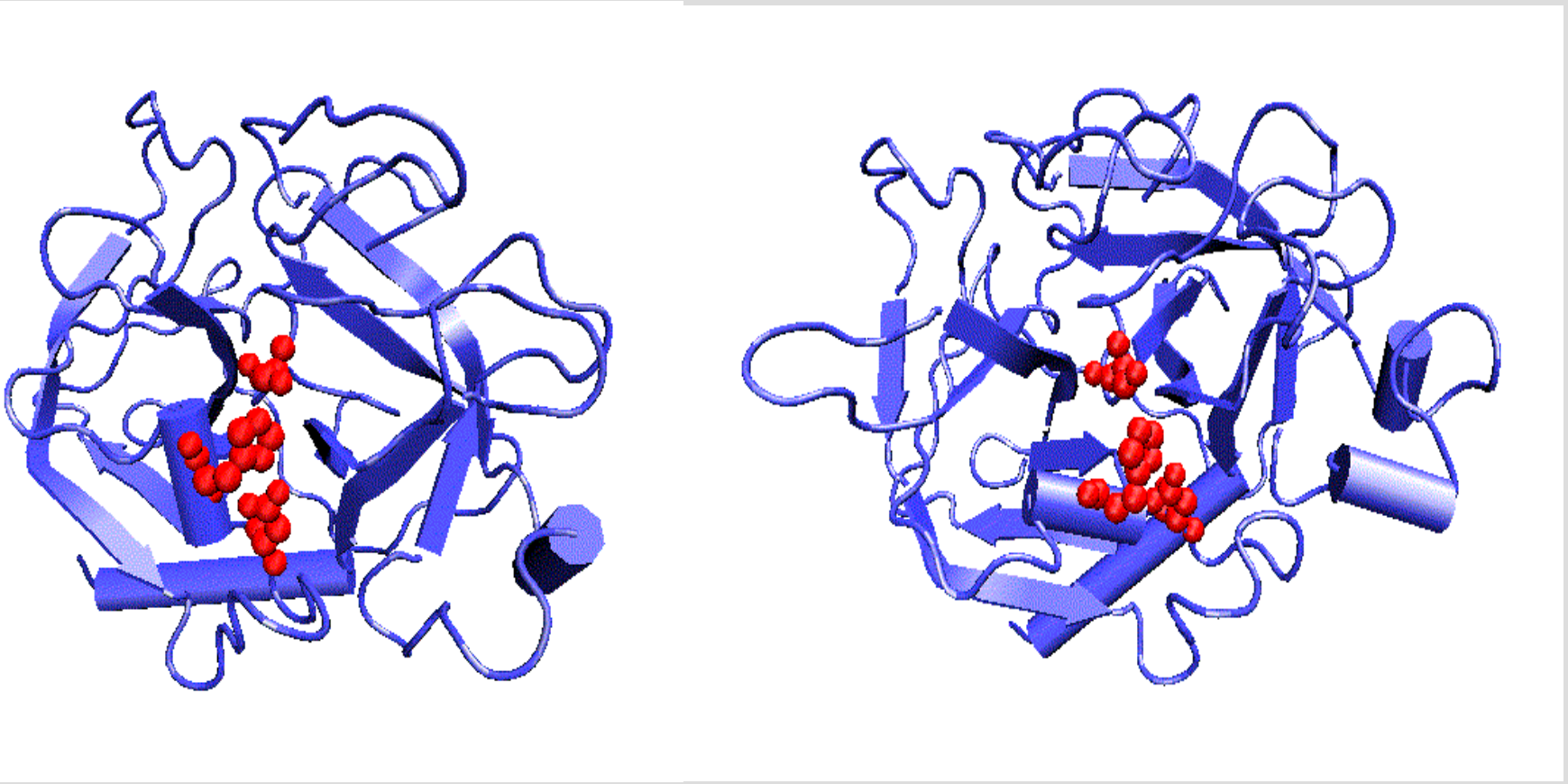
c) Determinant of thermal stability-Aromatic clusters

d) Protein-Protein interaction surfaces

B) Protein Structure Comparison

a) Clusters in topologically similar proteins

b) Eigen value and Vector component Spectra

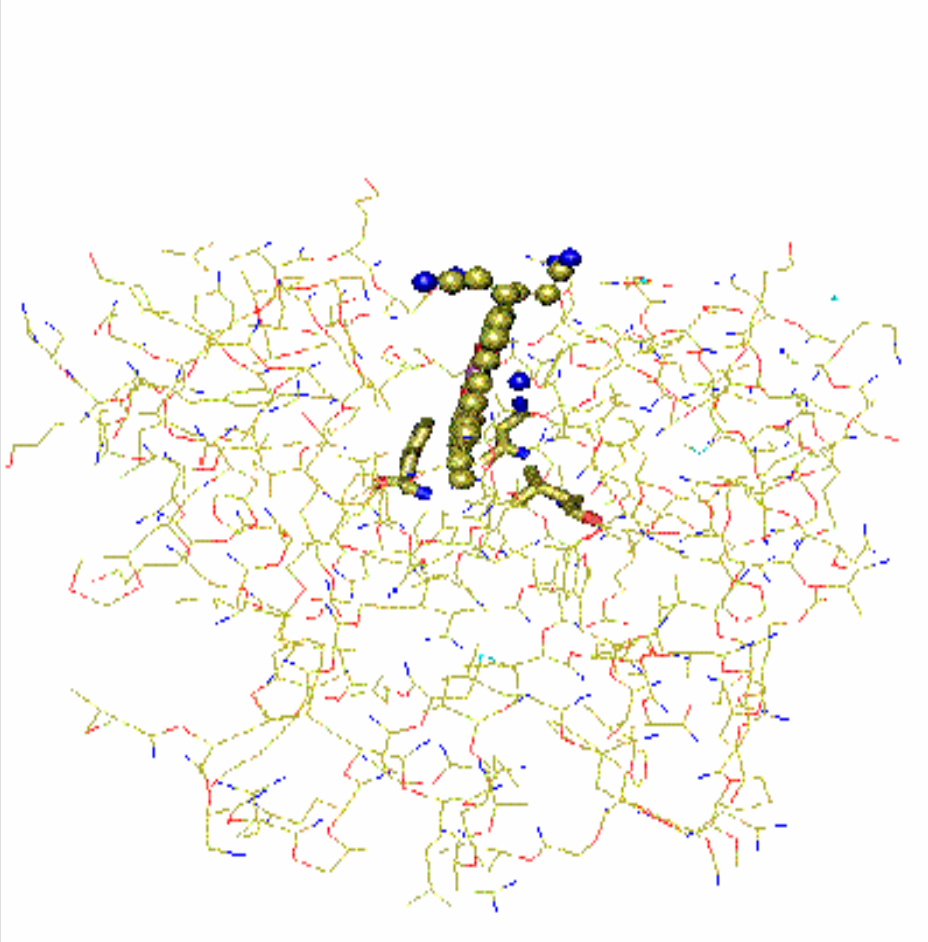


Trypsin

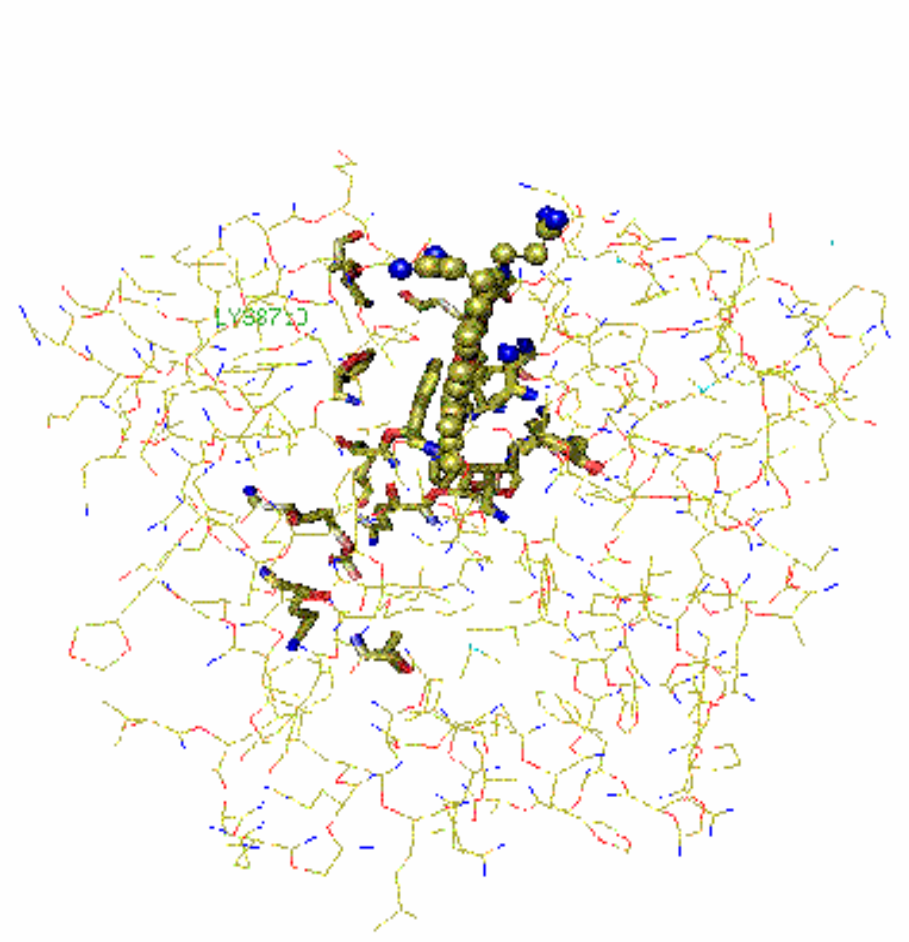
Elastin

Active site cluster in Serine Proteases. The active site cluster consisting of Ser-His-Asp triad in Trypsin and Elastin are shown using van der Waal's representation. This catalytic triad pattern is characteristic of the Serine Proteases and is known to be conserved.

Myoglobin (4mbn) at high contact criterion

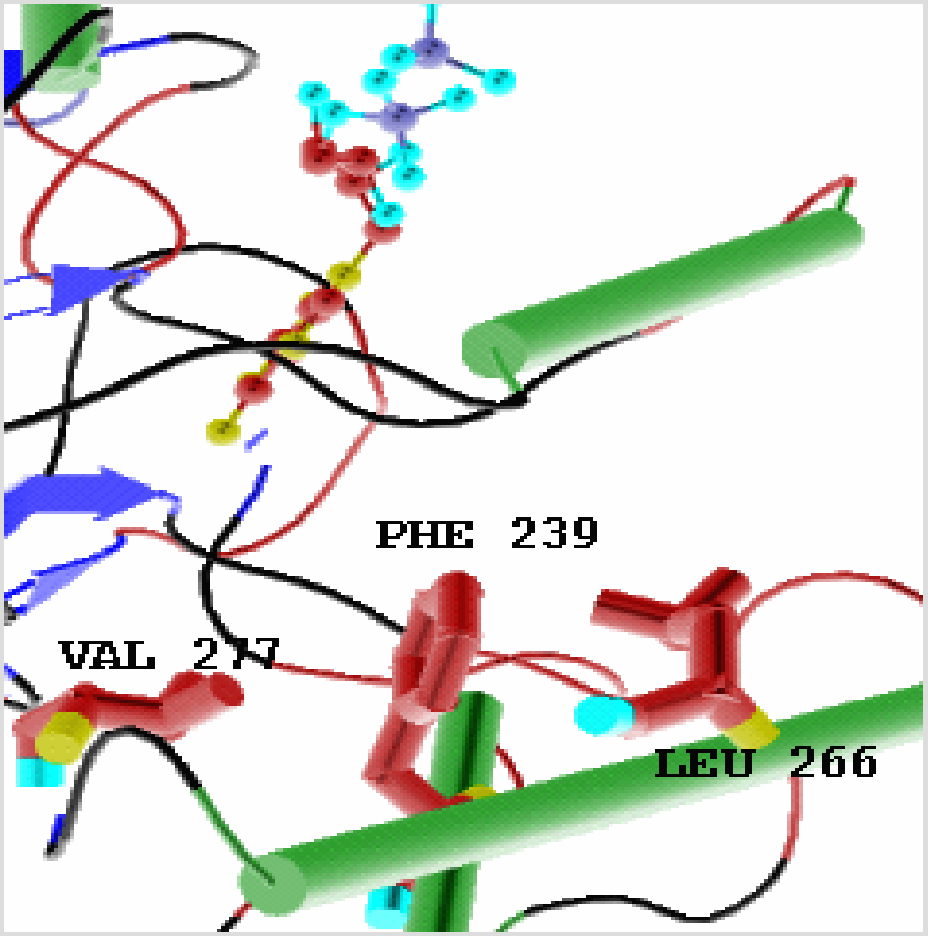


Myoglobin (4mbn) at low contact criterion

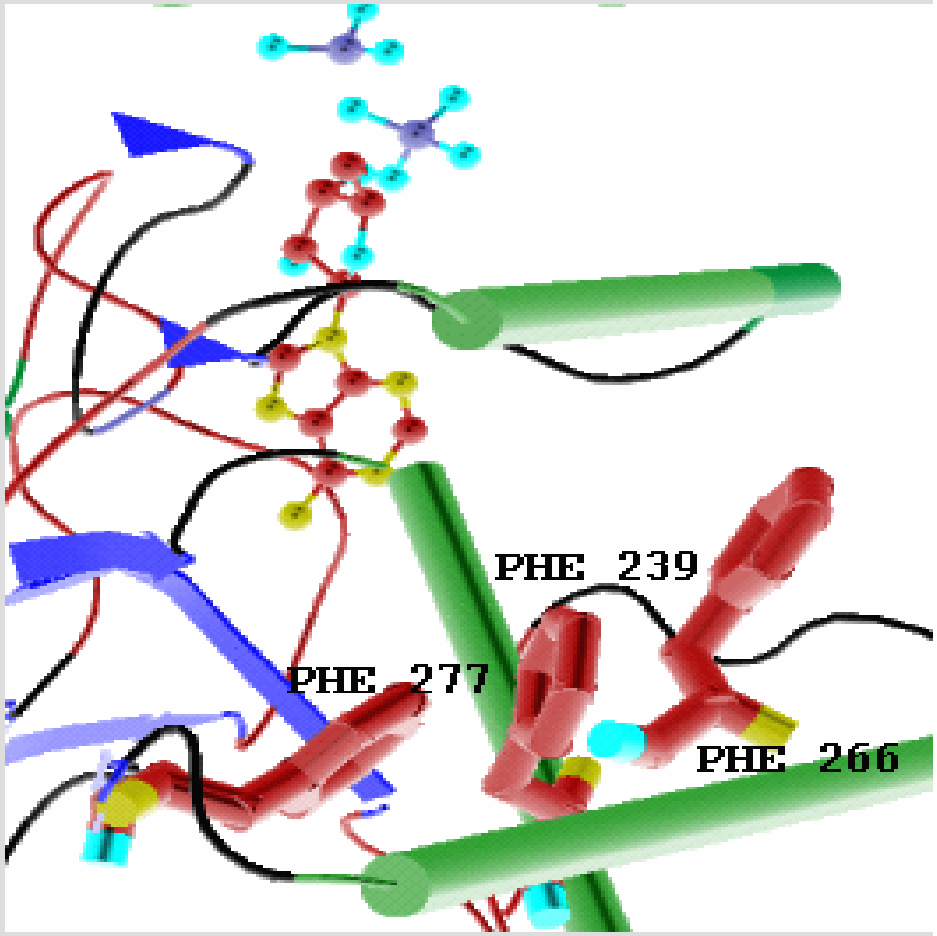


Expansion of the active site cluster in Myoglobin at high and low cut off. The cluster residues along with the porphyrin ring at the active site are shown in bold.

Mesophilic Phosphoglycerate kinase (3pgk)



Thermophilic Phosphoglycerate kinase (1php)



Aromatic clusters in mesophilic (3pgk) and thermophilic (1php) Phosphoglycerate kinase. Residues 239F, 266L and 277V form a cluster in the mesophile. The non-aromatic Leucine and Valine are replaced by aromatic phenylalanines in positions 266 and 277 respectively, in the Thermophilic protein, thus forming a network of aromatic interactions.

Analysis of Homodimers :

I : Characterization of Protein Interfaces

II : Identification of Hot spots on protein interfaces

III : Identification of dimerization sites on monomers

Brinda,Kannan, Vishveshwara, Prot.Eng, **15**, 265 (2002)

Application to Homodimers

Dataset : 20 Proteins (2.5Å resolution or better)

Functional Dimers and non-homologous

Cardiotoxin(1cdt),

Spo0B(1ixm),

PhospholipaseA2(1pp2),

Triose phosphate isomerase(1ypi),

Citrate Synthase(2cts),

Tyrosyl tRNA Synthetase(2ts1),

Aspartate amino transferase (3aat),

Rubisco(2Rus),

Glutathione reductase(3grs),

Iron Superoxide Dismutase(3sdp),

Interleukin8(1il8),

Mannose Binding Protein(1msb),

Uteroglobin(1utg),

CytochromeC(2ccy),

Gene5 DNA Binding Protein(2gn5),

Thymidylate Synthase(2tsc),

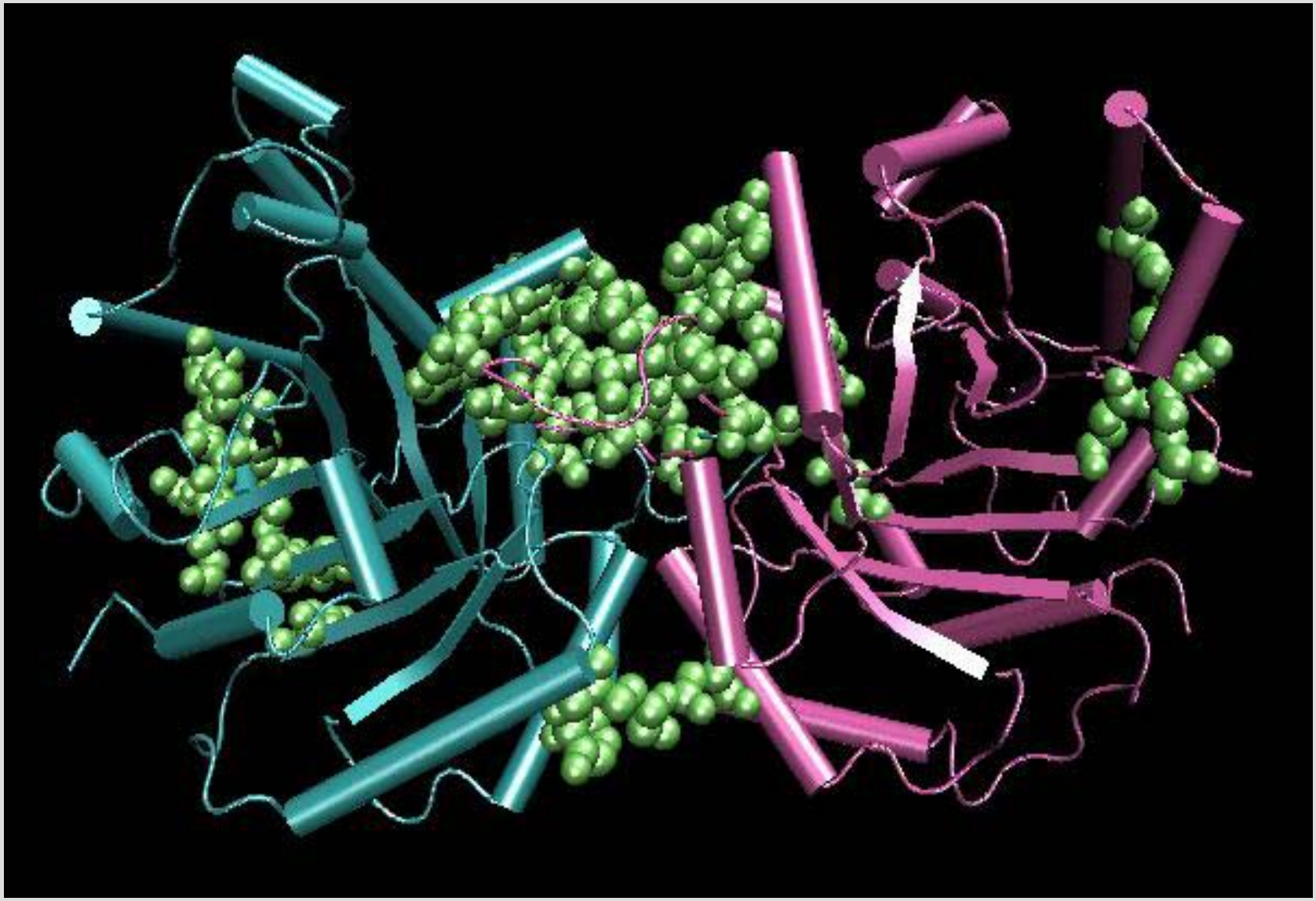
Tryptophan repressor(2wrp),

Catabolic gene activator Protein (3gap),

Isocitrate dehydrogenase(3icd),

Malate Dehydrogenase(4mdh)

Side Chain Clusters in Triose Phosphate Isomerase with 12% cut off

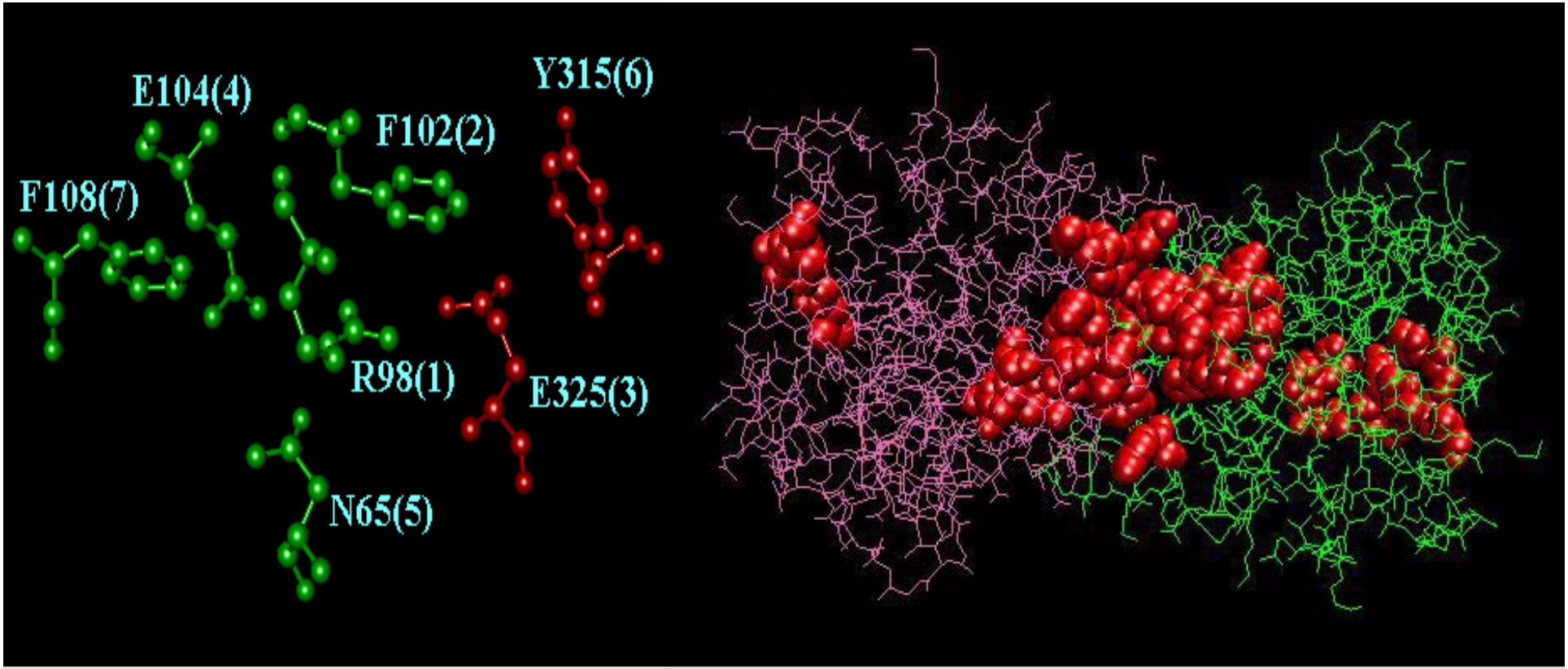


Residue	Vector Component	
	HEV	2nd LEV
F108(A)	0.055	0.146
Y315(B)	0.069	0.146
N65(A)	0.143	0.146
E104(A)	0.196	0.146
E325(B)	0.400	0.146
F102(A)	0.598	0.146
R98(A)	**0.645**	0.146
R3(A)	**0.753**	0.087
R189(A)	0.649	0.087
D227(A)	0.104	0.087
H95(A)	0.338	0.262
N10(A)	0.475	0.262
T75(B)	**0.813**	0.262
E97(B)	0.334	0.083
H95(B)	0.479	0.083
T75(A)	**0.812**	0.083
Y49(B)	0.214	0.007
D48(A)	0.576	0.007
K17(B)	**0.789**	0.007
E37(B)	0.195	0.392
R205(B)	0.589	0.392
F6(B)	**0.784**	0.392
H185(B)	0.213	0.001
D227(B)	0.293	0.001
I206(B)	0.349	0.001
R189(B)	0.528	0.001
R3(B)	**0.684**	0.001
Y67(A)	0.062	0.152
E104(B)	0.125	0.152
N65(B)	0.203	0.152
E77(A)	0.390	0.152
F102(B)	**0.627**	0.152
R98(B)	**0.627**	0.152

Triose Phosphate Isomerase : 12% Cut off

2nd LEV : Cluster forming residues

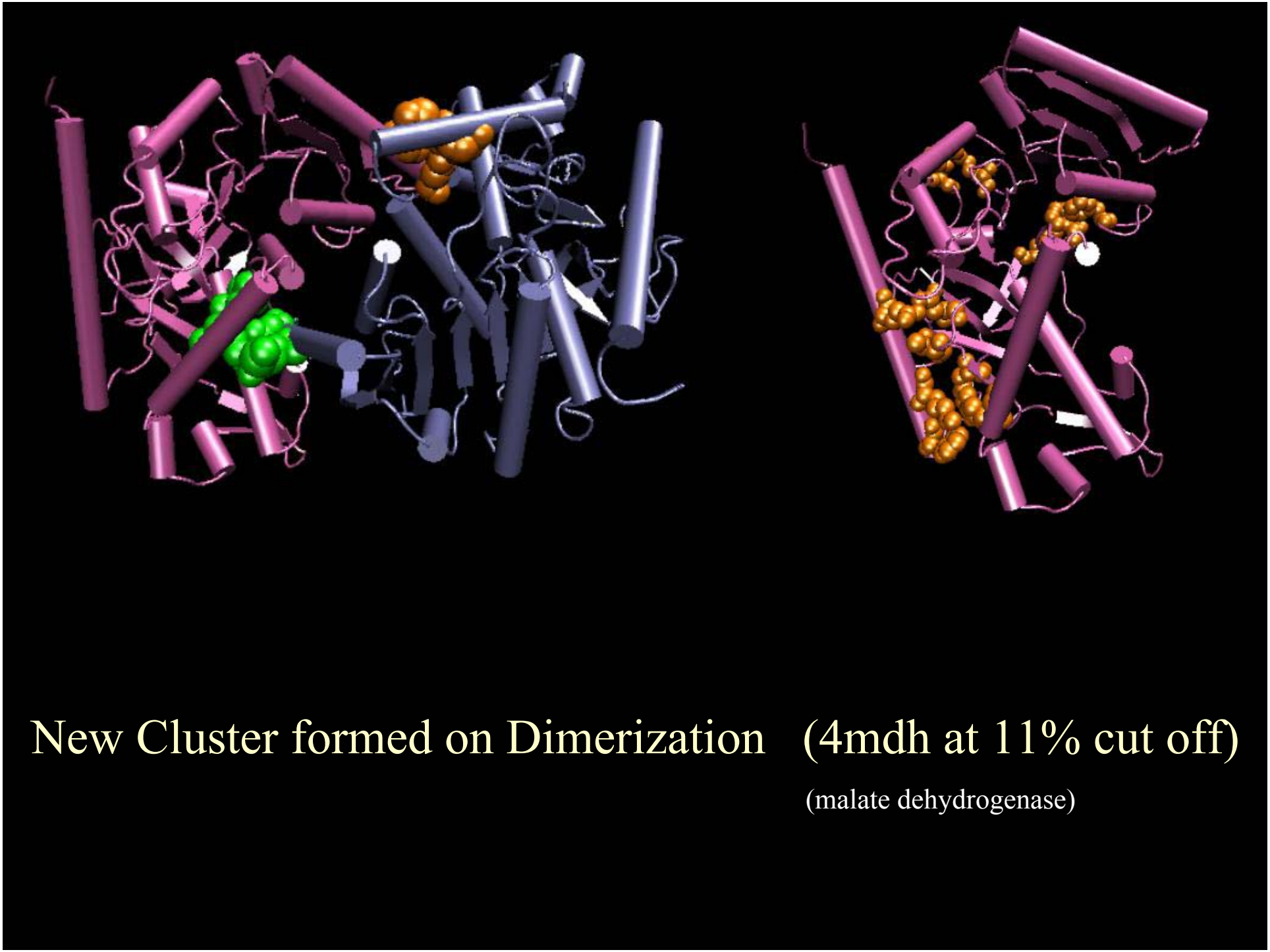
HEV : Cluster Centres



An interface cluster in Triose Phosphate Isomerase

R98 and F102 have been identified as hot spots

The predicted hot spots have been compared with the experimentally available results.



New Cluster formed on Dimerization (4mdh at 11% cut off)

(malate dehydrogenase)

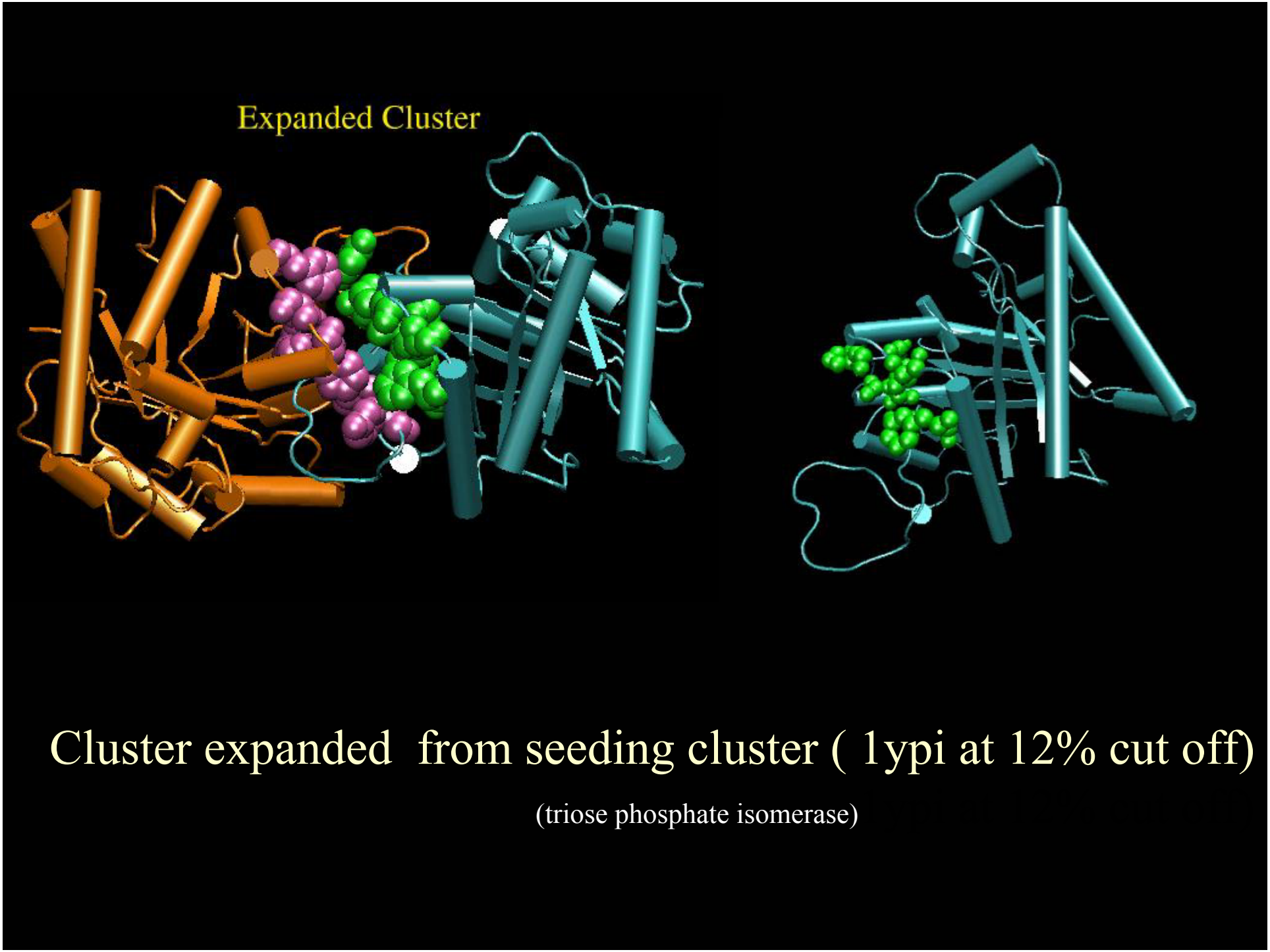
I. Characterization of Dimer Interface :

1. Identification of interface clusters :

- Clusters determined in all monomers and dimers
- 14% to 6% contact criteria
- Same cutoff used for a monomer-dimer pair.
- Interface clusters have contributions from both chains of the dimer.

2. Two different types of interface clusters :

- New interface clusters
- Expanded interface cluster

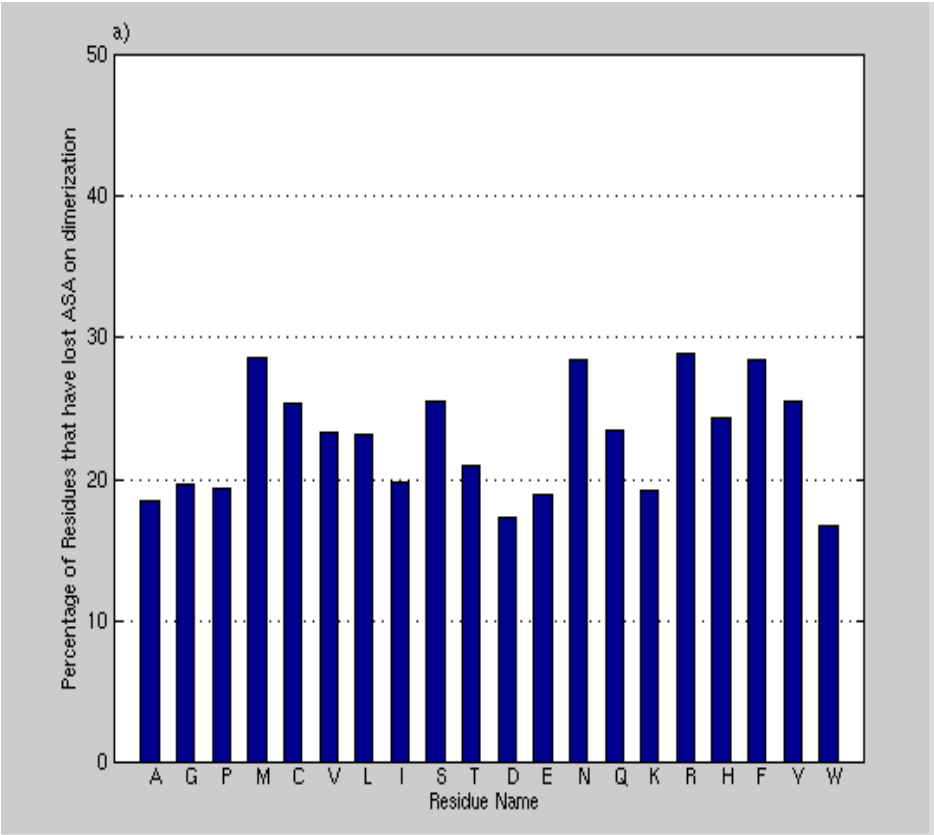


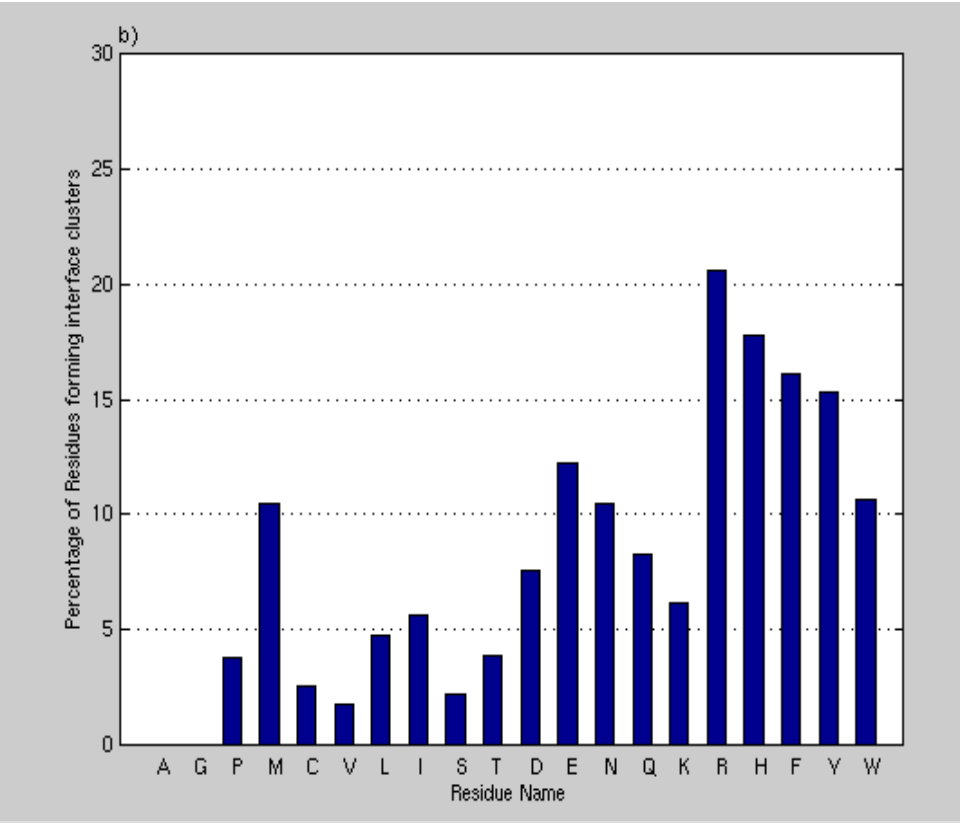
Cluster expanded from seeding cluster (1ypi at 12% cut off)

(triose phosphate isomerase)

Amino Acid	Total No. of residues in the data set	No. of residues that have lost accessible surface area*	No. of residues which are part of interface clusters
Ala	369	68	0
Gly	366	66	0
Pro	186	36	7
Met	105	30	11
Cys	79	20	2
Val	278	65	5
Leu	397	92	19
Ile	247	49	14
Ser	231	59	5
Thr	234	49	9
Asp	237	41	18
Glu	286	54	35
Asn	162	46	17
Gln	145	34	12
Lys	291	56	18
Arg	180	52	37
His	107	26	19
Phe	155	44	25
Tyr	137	35	21
Trp	66	11	7

Amino acid composition at the dimer interface





Composition of Interface Clusters :

- Both hydrophobic and polar residues are found in interface clusters
- Oppositely charged residues stabilize the interface clusters.
- Arginine, Histidine, Phenylalanine, Tyrosine and Glutamic acid (aromatic and charged) are more frequently found in interfaces.
- Clusters are more discriminatory than δASA.

II : Identification of Hot Spots at dimer interfaces:

Cluster Centre (Using Graph Spectra):

- Eigenvector component of higher eigenvalues
- First and second highest eigenvector components

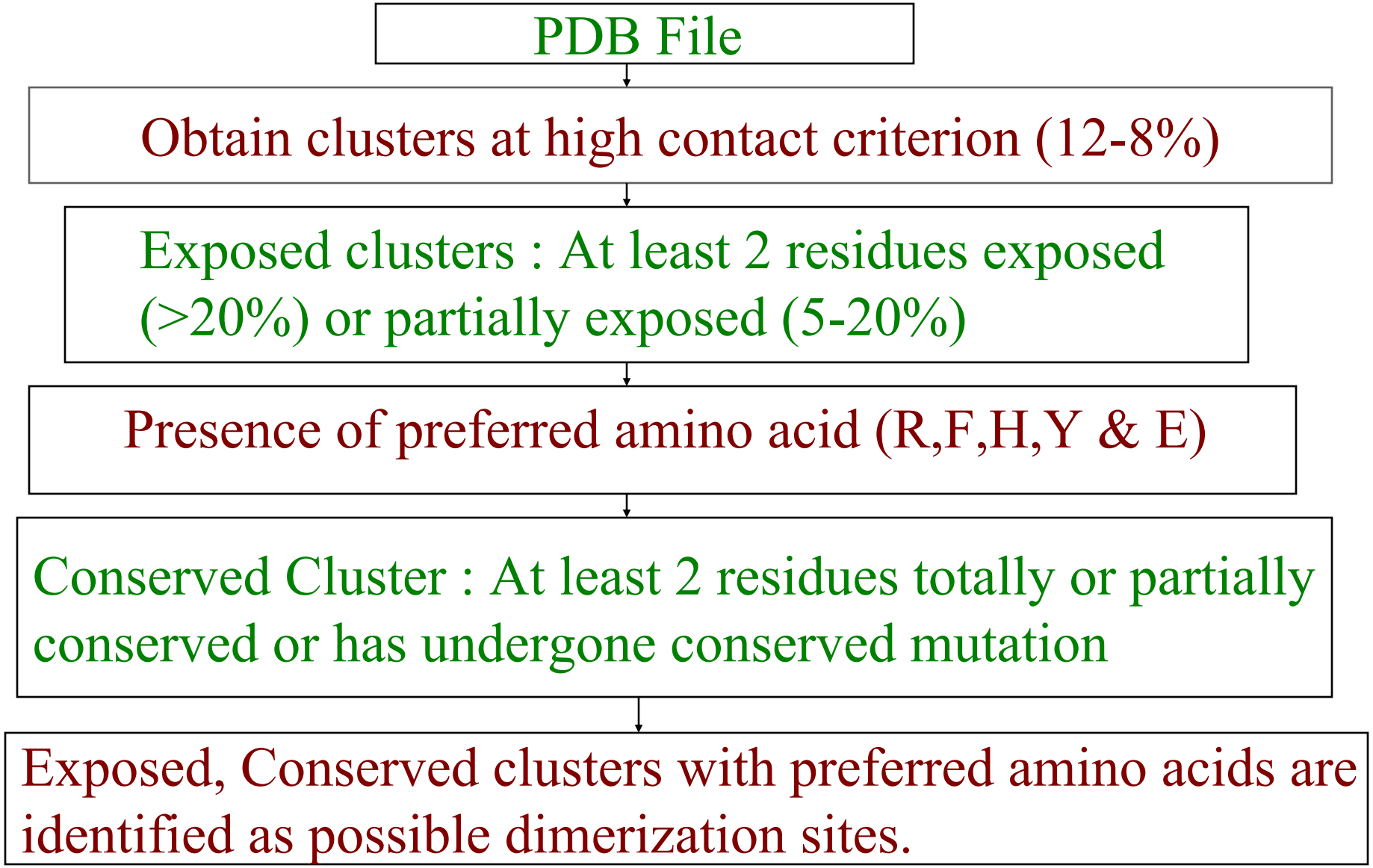
Accessible surface area (Connolly's method):

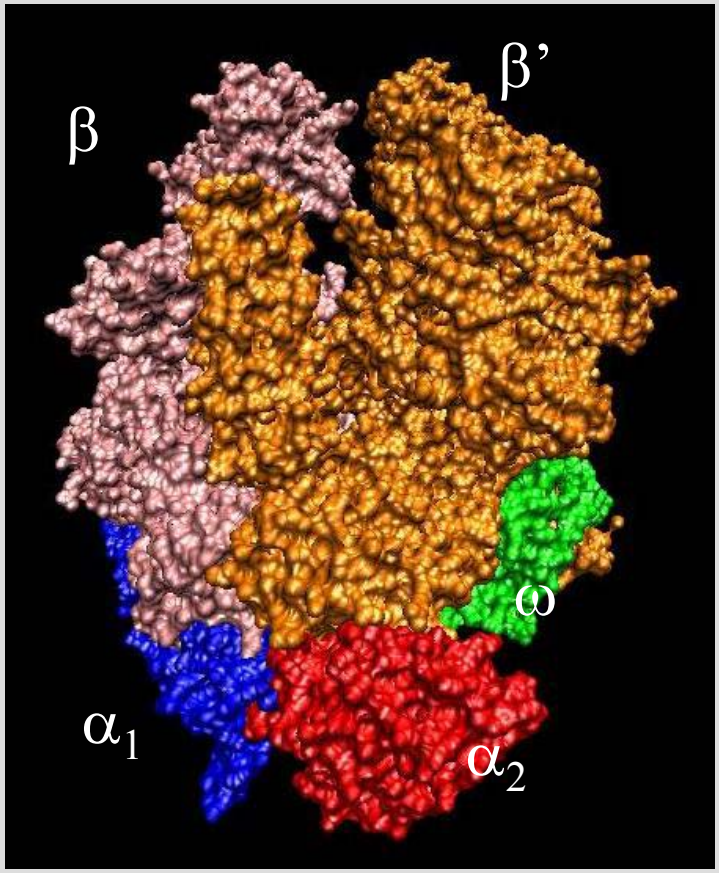
- Change in accessible area when monomer dimerizes
- Any interface cluster residue that has lost ASA

Conservation (Multiple sequence alignment using ClustalW):

- Conservation of interface residue in homologous sequences
- Totally conserved or partially conserved or conserved mutations.

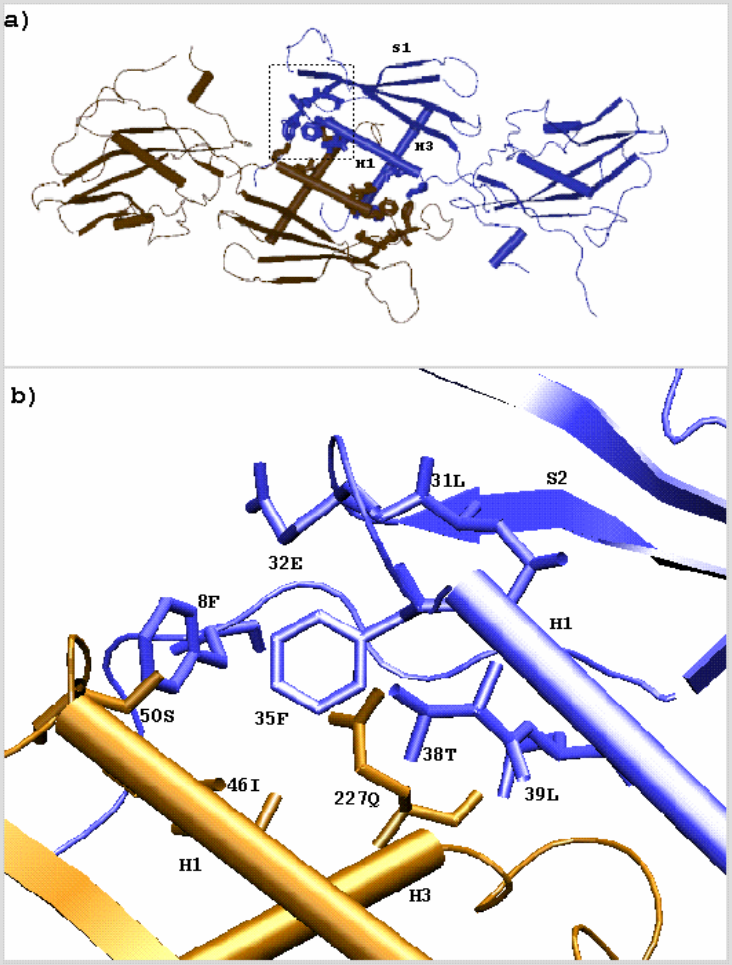
III. Identification of dimerization Sites on the Monomers :





RNA Polymerase

(structure from Thermus Thermophilus)



α_1-α_2 Interface

Mutation of 35F lead to the loss of dimerization

Kannan, et.al,, Protein Sci,**10**, 46 (2001)

Summary:

- The Graph Spectral method takes into account the global topology of the protein.

•Analysis of clusters of spatially connected residues yield better results than considering pair-wise residue interactions and direct identification of the cluster is identified given the percentage interaction criteria.

•Clusters at the Protein-Protein interface and the cluster centres (Hot spots) are identified by Eigen values and their vector components of the Laplacian Matrix of the protein graph

• Clusters at the dimer interface can be an (a) Expanding one or (b) New one

•Charged and aromatic residues are often found at the interface

•Predicted hotspots and the interface cluster on the monomer based on (a) Cluster centre, (b)Accessible surface area and (c) Conservation among homologous sequences show agreement with experimental results

•Further studies: Extended data set (~400 dimers), Protein-DNA/RNA interaction

Acknowledgement

Kannan

Brinda

SathyaPriya

Rakesh K Pandey

S.M. Patra

Support: Department of Biotechnology, India for the project

“ Genomics Initiative at IISc- Computational Genomics”