

Clustering of peptide fragment structures reveals nature's building block approach

Ashish V. Tendulkar
Research Scholar
Kanwal Rekhi School of I.T.
I.I.T. Bombay

Guide: **Prof. P. Wangikar**

Co-guide: **Prof. Sunita Sarawagi**

Outline

- ◆ Terms
- ◆ Objectives
- ◆ Approach
- ◆ Results
- ◆ Conclusion

Terms

- ◆ Protein is made up of amino acids. There are in all 20 different types amino acids.
- ◆ Protein is a linear sequence of amino acid.
- ◆ Protein takes up 3-D structure. The structure is result of its amino acid sequence.

Protein Structure

- ◆ Primary Structure:
ACGADSTYKSTYSC
PLA
- ◆ Secondary structure
- ◆ 3-D structure



Objectives

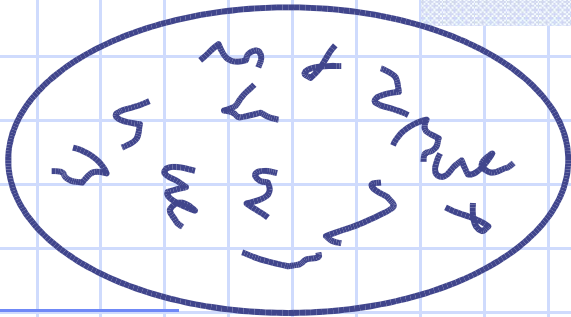
- ◆ Prediction of protein structure from merely its sequence.
- ◆ Protein sequence is believed to take up vast number of conformations
- ◆ Learn relation between sequence and structure by example of known protein structures.
- ◆ Build library of sequence-structure mapping

Salient Features

- ◆ **Geometric invariant:** A quantity, which is unchanged under a group of geometric transformations, in this case, the group of translations and rotations in 3-dimensional space.
- ◆ Examples of continuous invariants: signed volumes, areas, lengths.
- ◆ For our group of transformations, it has been shown that invariants suffice to decide superimposability of two structures. Thus, if two patterns ***K1*** and ***K2*** are not superimposable then there is an invariant f such that $f(K1) \neq f(K2)$.

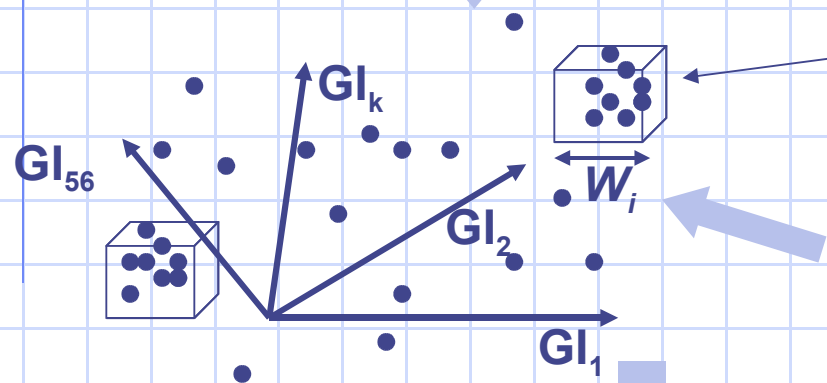
Salient Features

- ◆ We discretize a structure by its evaluations on a fixed suite of N invariants and mapped into the N -dimensional space as a vector.
- ◆ We examine 1.2 million peptides from 4,500 non-redundant protein structures.
- ◆ This collection may now be subjected to the tools of data-mining.
- ◆ **Clustering of Patterns:** A cluster is a small region in this N -space, which has a large number of pattern-vectors.
- ◆ Closeness of points and density is decided via a training regime



All overlapping octapeptide fragments from PDB_95

Geometric invariant based representation of each peptide as a point in 56-dimensional space and clustering



Dense cluster of peptides in a 56-dimensional box

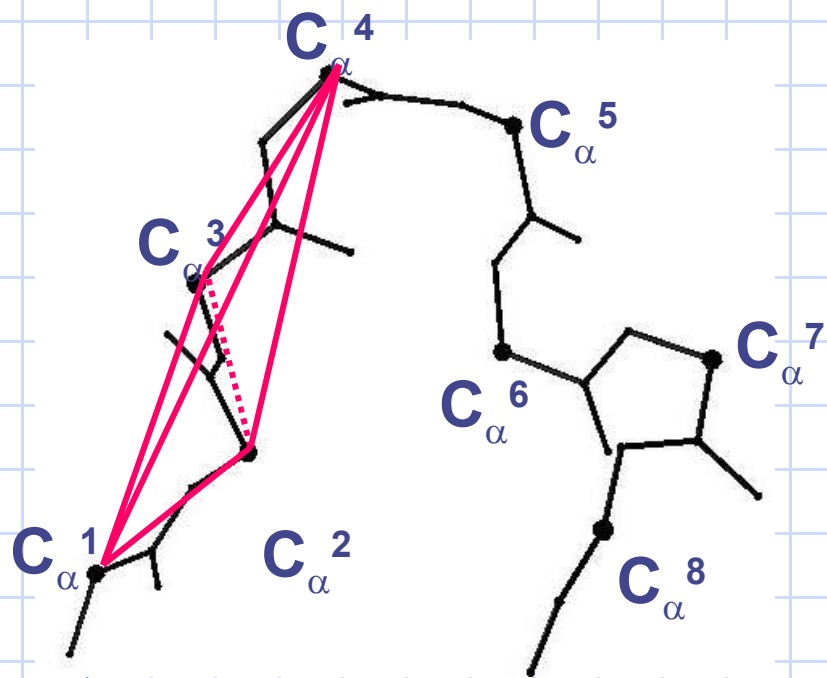
Training regime to decide the tolerance window W_i in each dimension based on known superimposable peptides.

Categorization of clusters

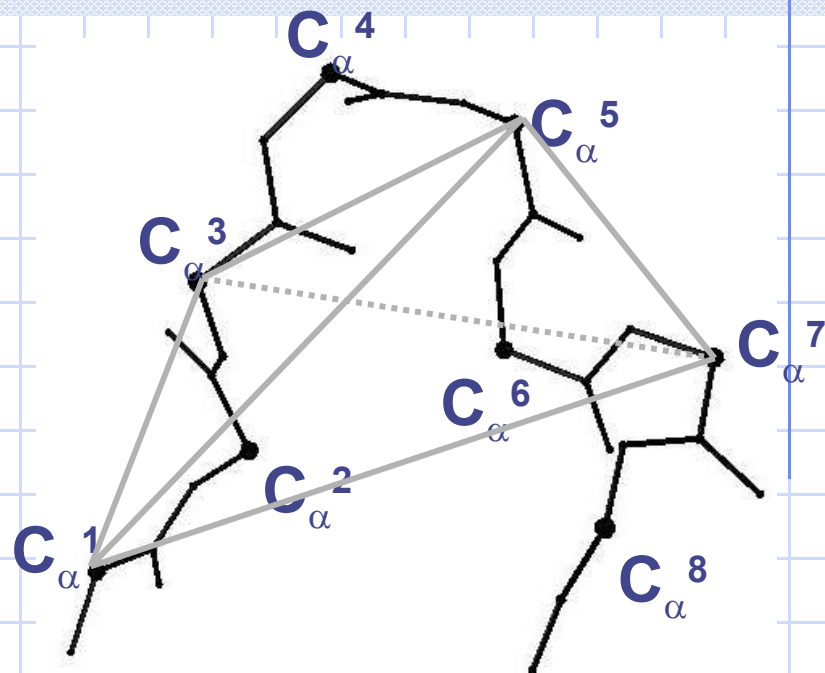
“*Functional*” clusters with majority of peptides drawn from a single SCOP superfamily.

“*Structural*” clusters

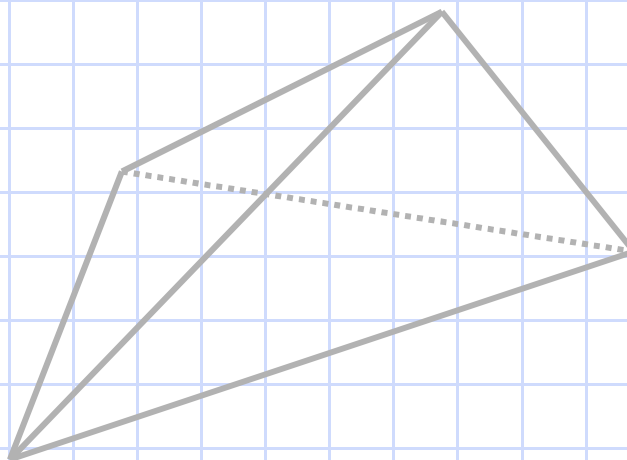
Hierarchical clustering based on closeness of centroids of clusters



a) Tetrahedron_gap_0: constructed from consecutive C_{α} atoms.



b) Tetrahedron_gap_1: constructed from alternate C_{α} atoms.



c) Geometric invariants associated with a tetrahedron

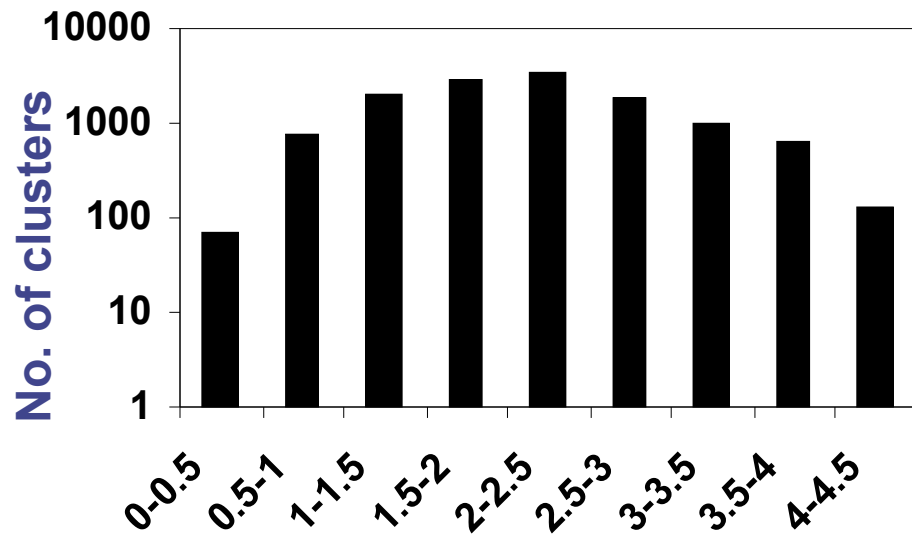
Examples of G.I.

- Surface area
- Volume
- Perimeter
- Sum of squares of edges
- Sum of centroid to node distances

Summary of Peptide Library

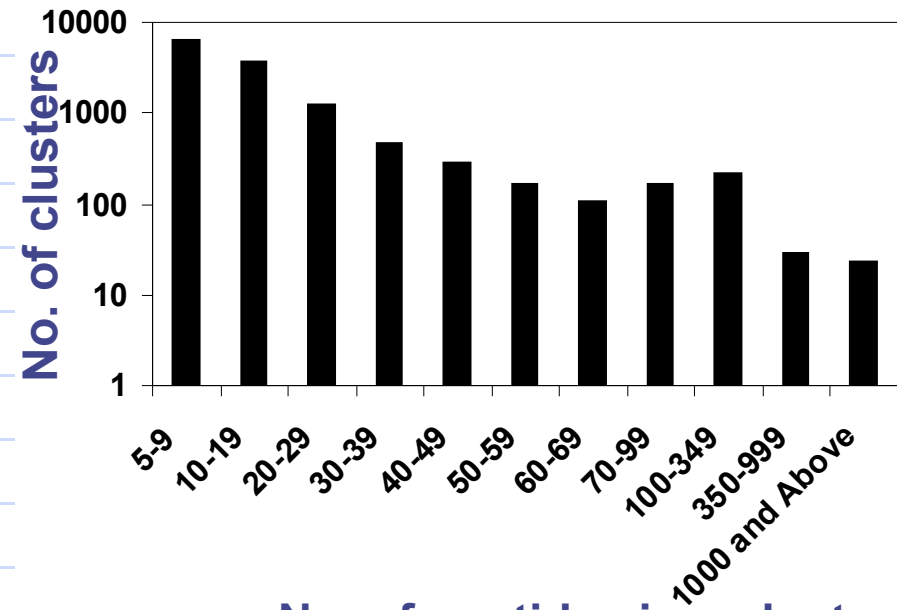
- ◆ 12000 clusters, size range from 5-160,000.
- ◆ 2000 functional clusters.
- ◆ Demonstrates nature's bias toward a selected conformations.
- ◆ Potential applications in protein structure prediction.

Distribution of clusters By Information Content



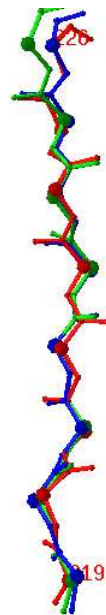
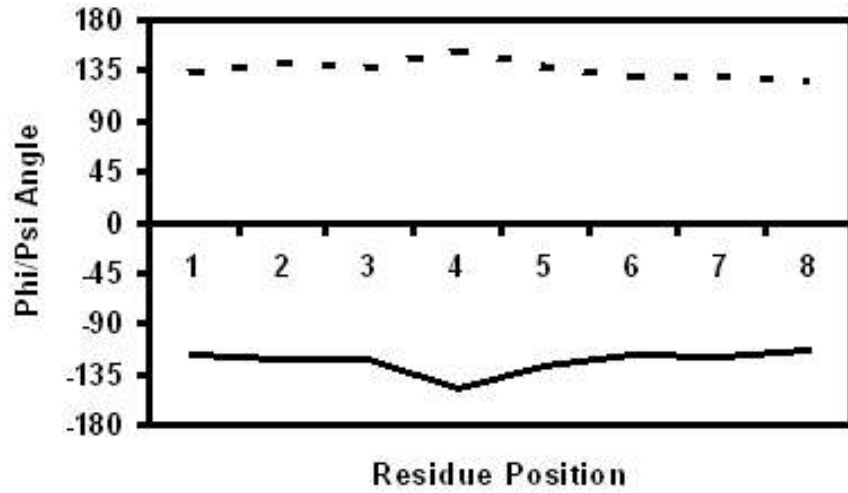
Avg. information content of the cluster

Distribution of clusters By Cluster size.

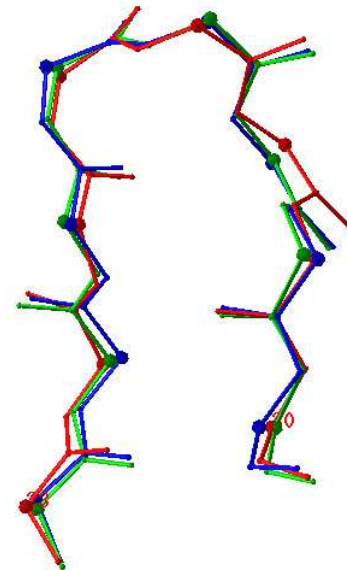
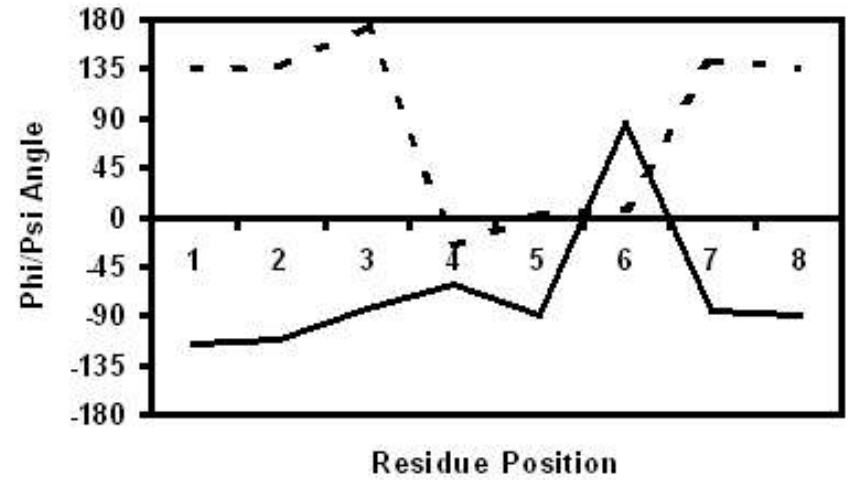


No. of peptides in a cluster

Structural Clusters

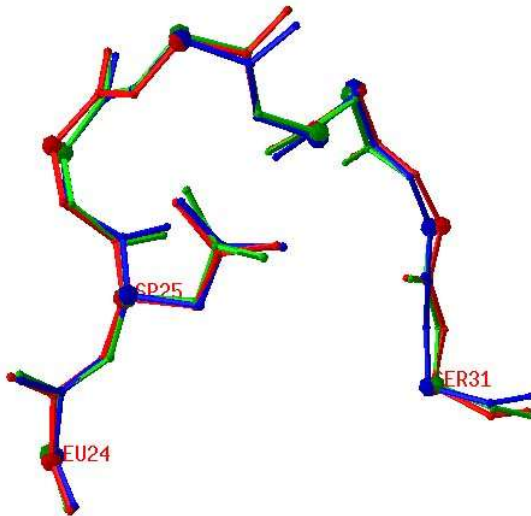
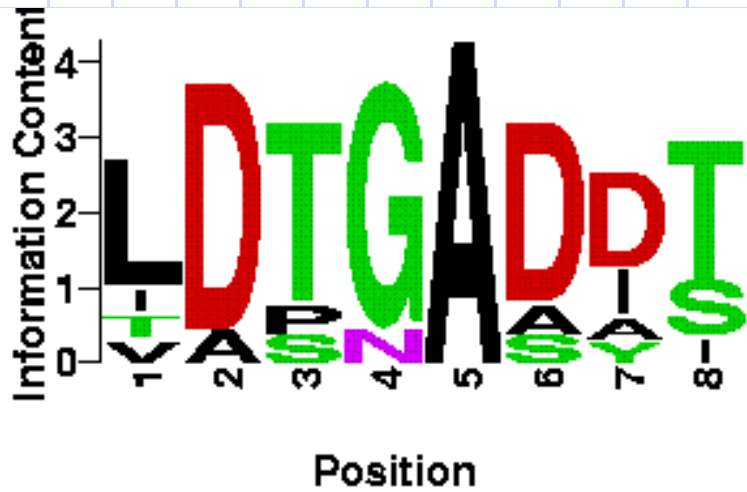


Twisted β -strand (S.2.10.1.23.389)

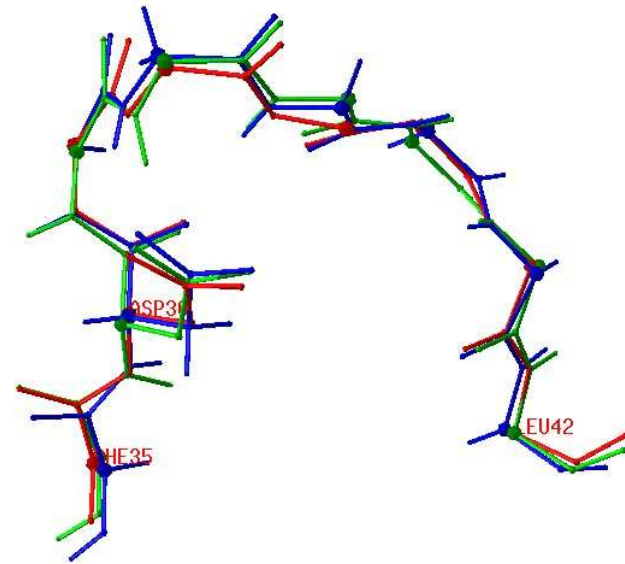


Known β -hairpin (S.1.6.1.6.19)

Functional Clusters



Acid Proteases: Active site loop conformation I (F.b.50.1.3.11.7870)



Acid Proteases: Active site loop conformation II (F.b.50.1.4.9.3460)

Conclusions

- ◆ Century old “Geometric Invariant theory” applied to protein structure for the first time.
- ◆ Peptide fragment library(DPFS) can be used in protein structure prediction. It is available on web at www.it.iitb.ac.in/dpfs/

Acknowledgements

- ◆ Prof. Milind Sohoni for his inputs on Geometric Invariants
- ◆ Anand Joshi for his contribution in the project