

Presentation at ISCB Student Symposium, Madrid, Sept 28, 2005

Visualization and Clustering of Protein Local Conformational Space using Geometric Invariant Theory

By

Ashish V. Tendulkar
Prof. Pramod P. Wangikar



Indian Institute of Technology, Bombay
Mumbai, India-400 076

Motivation

- Protein local conformations are broadly classified into three categories viz. α -helix, β -strand, and loops.
- Ramchandran's plot^a and our earlier work^b shows that the protein local conformations are biased in favor of a finite number of conformations.
- Visualization and characterization of restrictiveness of conformational space remained a challenge owing to lack of unilateral structure descriptors and computationally expensive step ($O(n^2)$) of matching local conformations.

^aRamchandran, G. N. et. al.(1963), J. Mol. Biol., 7, 95-99

^bTendulkar, A. V. et. al.(2004), J. Mol. Biol., 338, 611-29

Key Techniques in Approach

- Octapeptide^a as a **unit of local conformation**^b. C_α atoms are used to approximate backbone geometry.
- Geometric Invariants(GI) are used as a **unilateral structure descriptors**. GIs remains unchanged under transformations like rotation and translation. E.g. perimeter a triangle. GIs are calculated from x, y, z coordinates of C_α atoms.
- Principal component analysis(PCA) is used for **dimensionality reduction**.
- K-means clustering is applied to **group similar conformations**.
- **Visualization** in terms of **conditional bivariate distribution plots**.

^a solutions

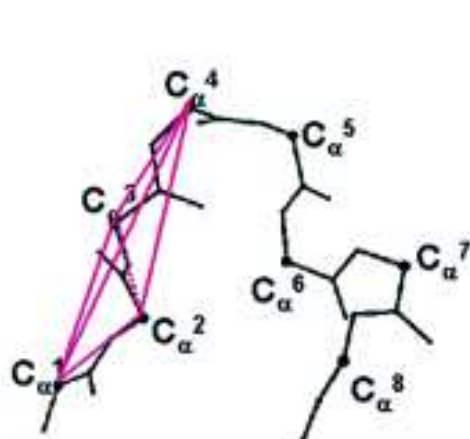
^b gaps in literature

Extraction of Local Conformations

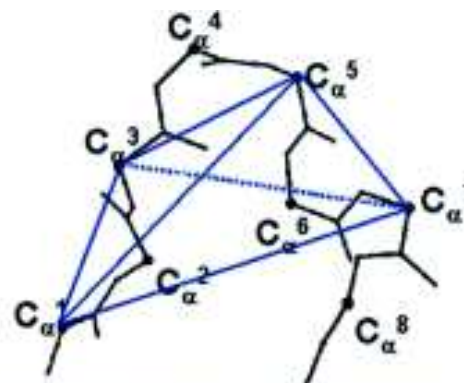
- Input proteins: **Astral^a 95 version 1.67 domains**
- The octapeptides are extracted in overlapping manner, with the neighboring octapeptides sharing an overlap of 7 residues.
- We used **Astral RAF** format sequence to keep track of gaps in sequence, thereby avoiding illegitimate octapeptides.
- Thus, **we extract total of 1,770,147 \approx 1.7M octapeptides.**

^aBreener et. al.(2000), Nucleic Acids Res. 28, 235-242

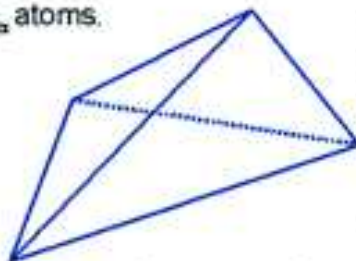
Construction of Structure Descriptors(GIs)



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



b) Tetrahedron_gap_1: Tetrahedron 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

Significance of Signed Tetrahedron Volume

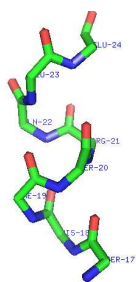
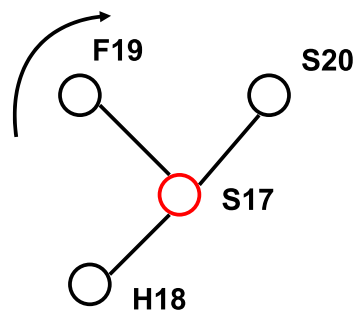


Fig. 2A

Tetrahedron _{$i, i+1, i+2, i+3$}



Tetrahedron _{$i, i+2, i+4, i+6$}

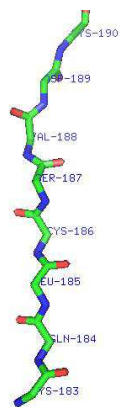
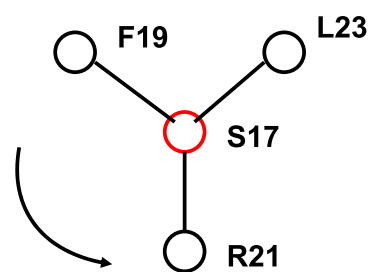
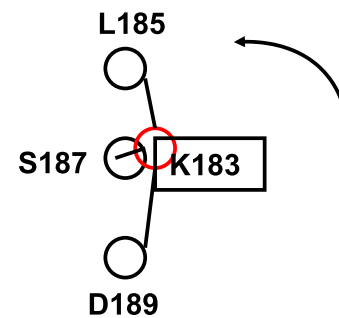
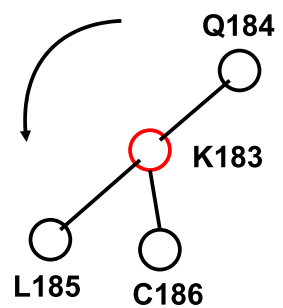
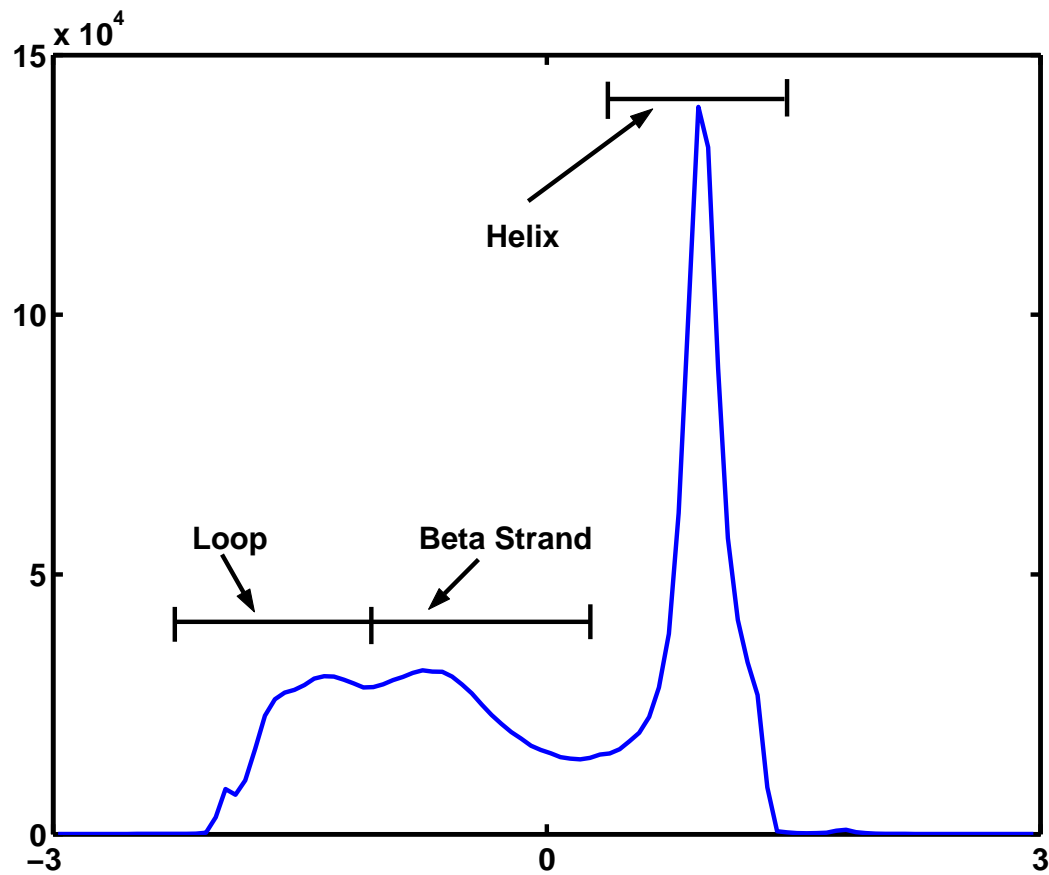


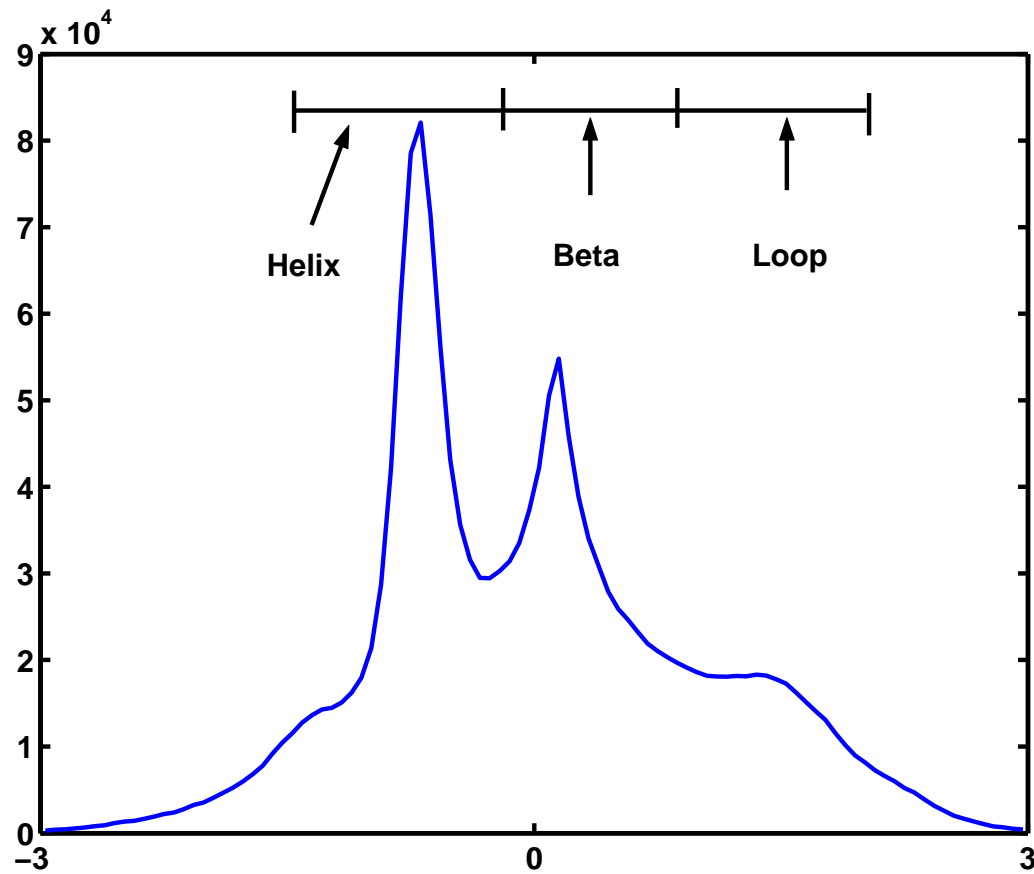
Fig. 2B



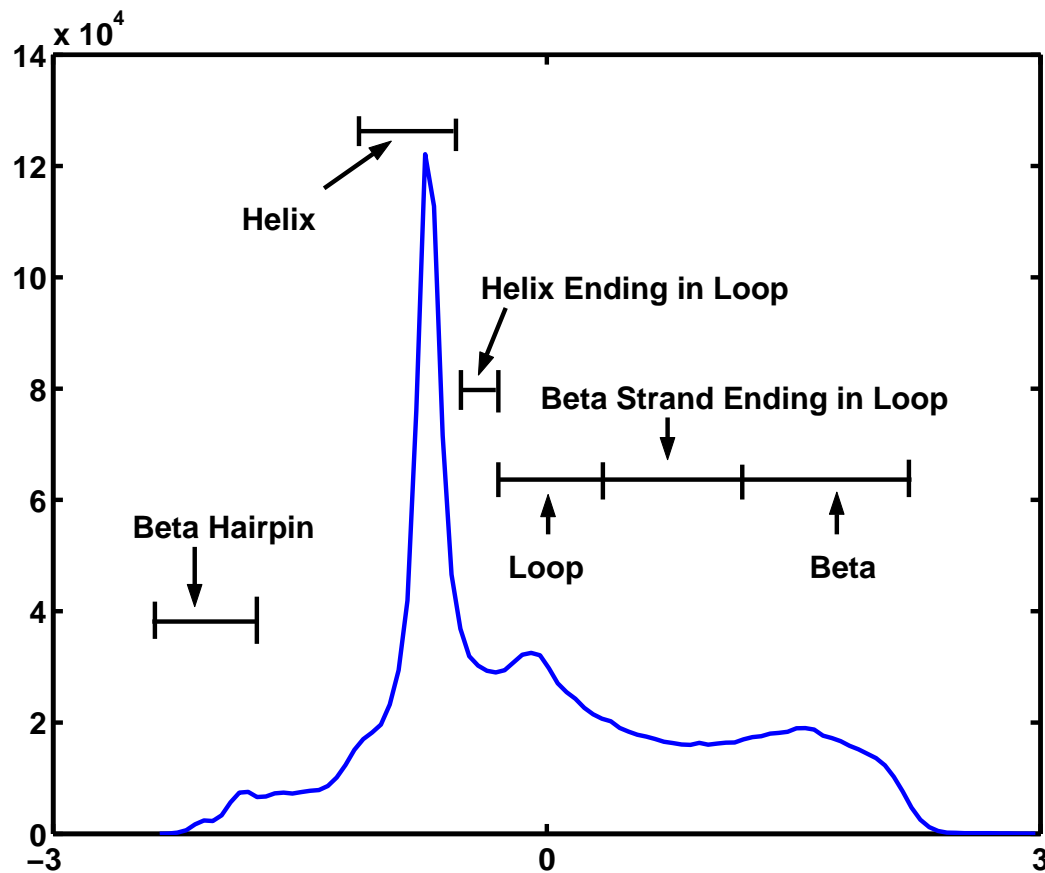
Distribution on Volume of Tetrahedron _{$i, i+1, 1+2, i+3$}



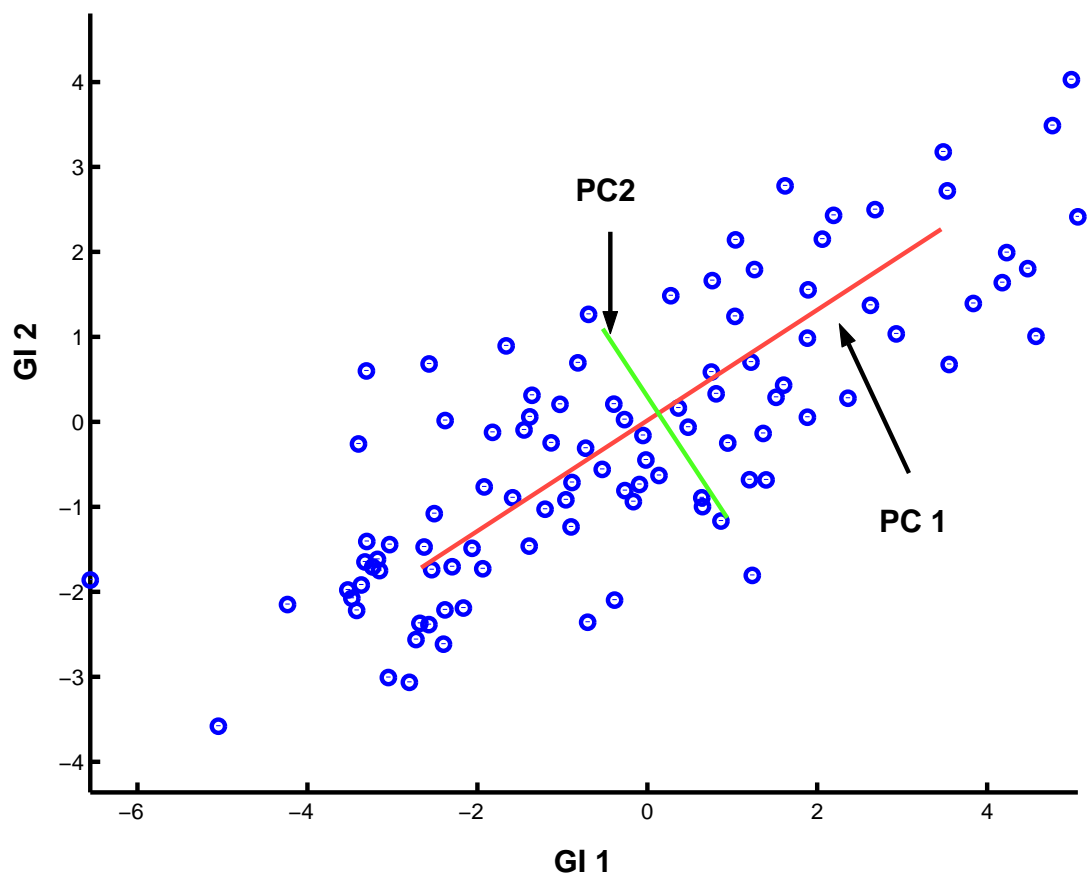
Distribution on Volume of Tetrahedron _{$i, i+2, 1+4, i+6$}



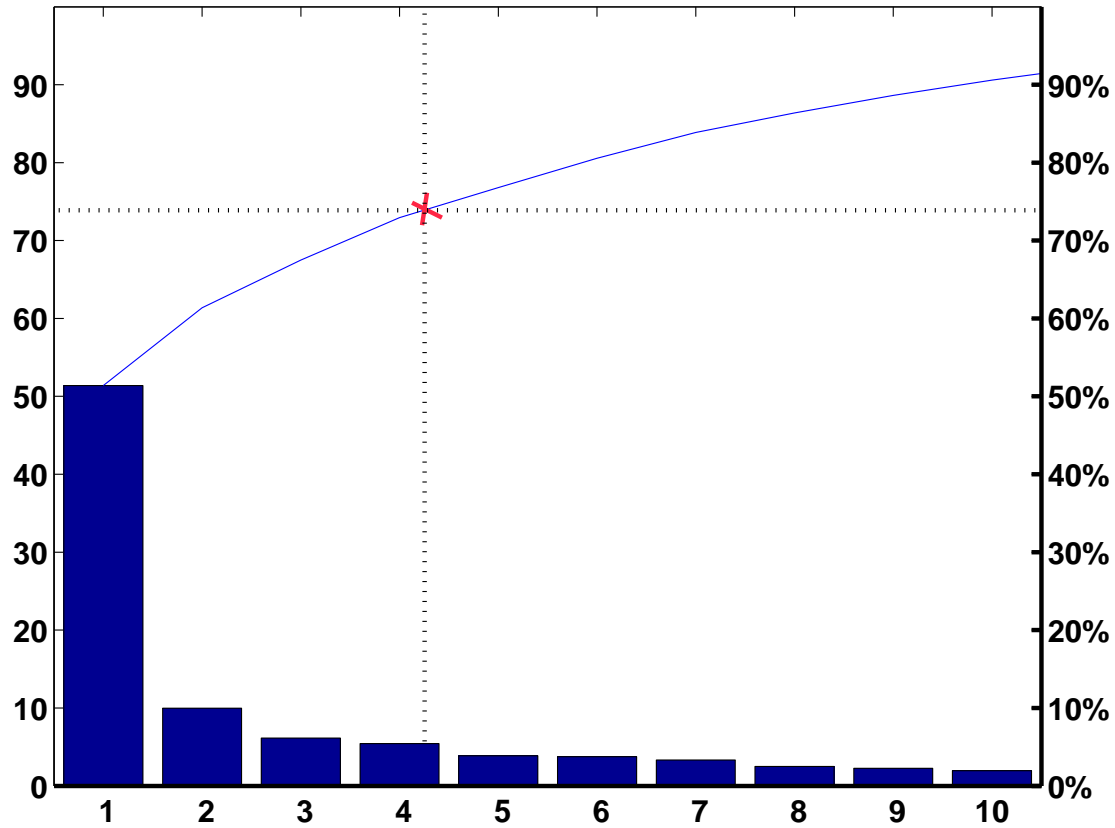
Distribution on End to End Distance



Principal Component Analysis



Principal Components

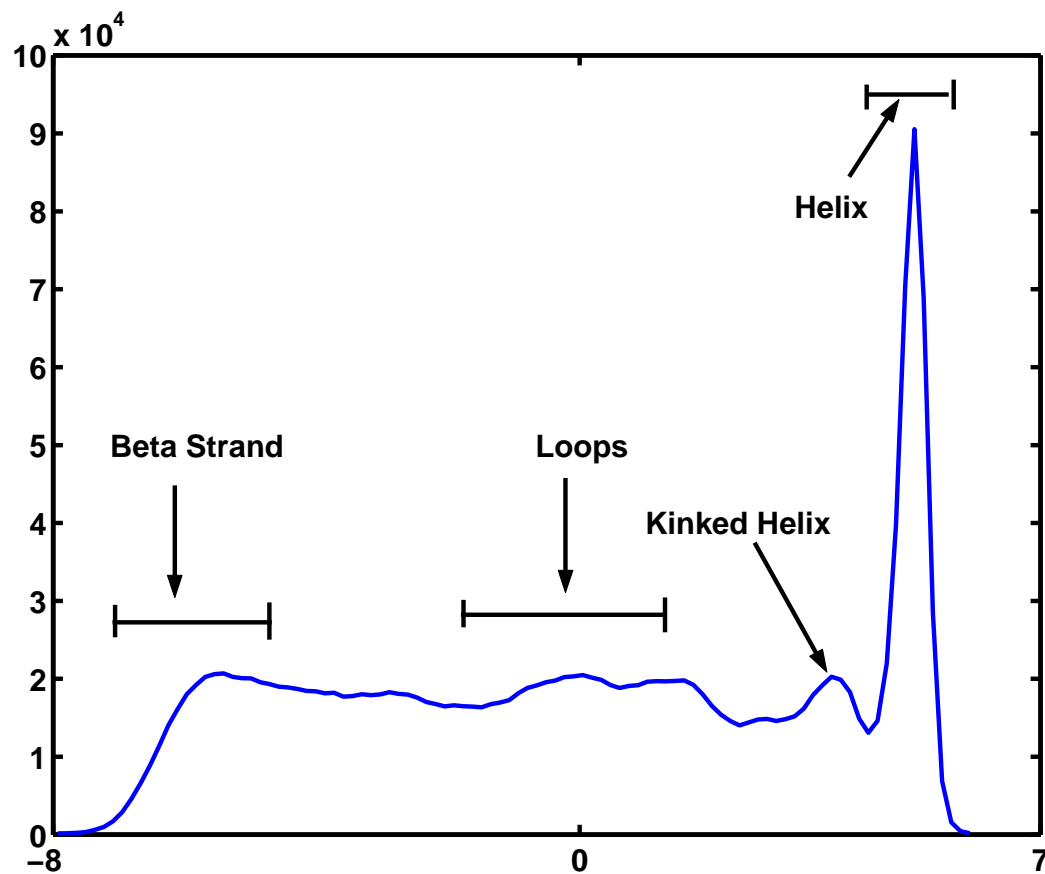


Analysis of First PC

Topmost GIs	Weight(w_i)	α -helix(x_i)	Partial Sum
$P_{1,3,5,7}$	-0.243	large -ve	large +ve
$P_{2,4,6,8}$	-0.241	large -ve	large +ve
$P_{1,5,8}$	-0.239	small -ve	medium +ve
$V_{2,3,4,5}$	0.154	large +ve	large +ve
$V_{4,5,6,7}$	0.159	large +ve	large +ve
$V_{3,4,5,6}$	0.160	large +ve	large +ve
Location of α -helix on $PC_1 = \sum w_i x_i$			= large +ve

Thus, α -helices occupy extreme +ve region, β -strands occupy extreme -ve, while *loops* occupy central region. **PC_1 distinctly separates extended structures from the compact ones.**

PC₁ Univariate Distribution



PC Univariate Distributions At a glance

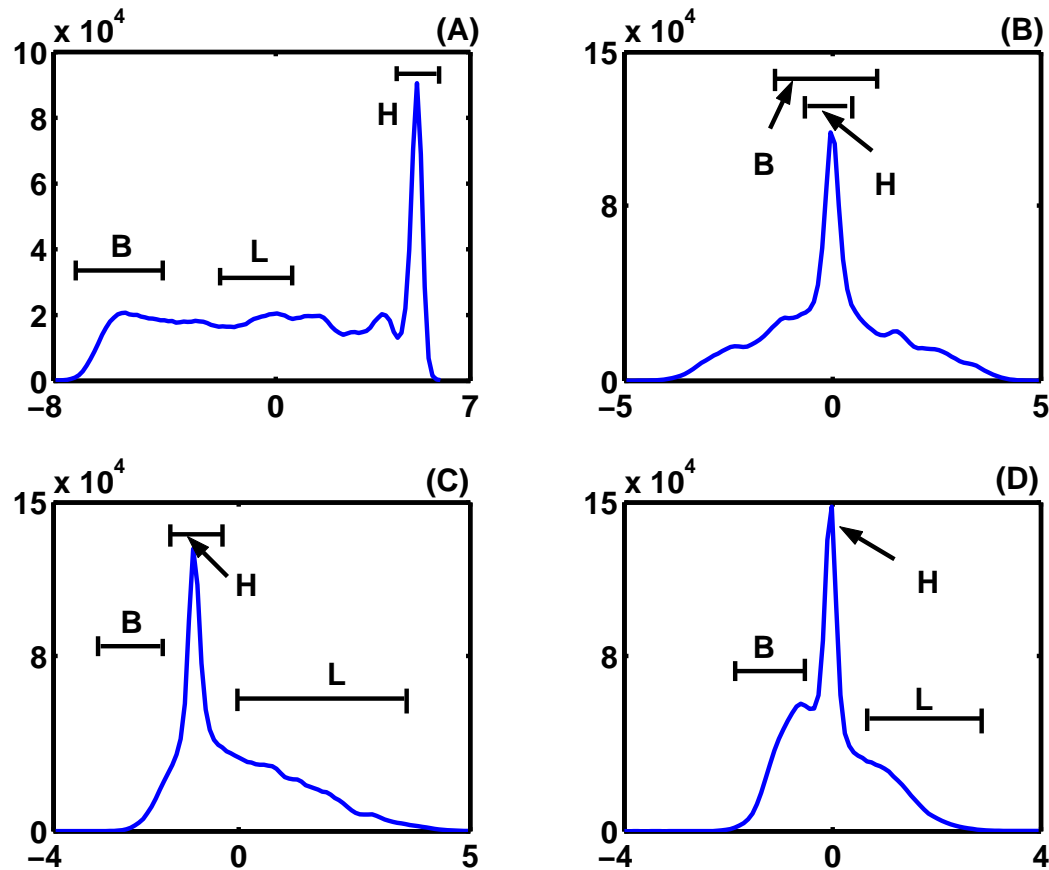
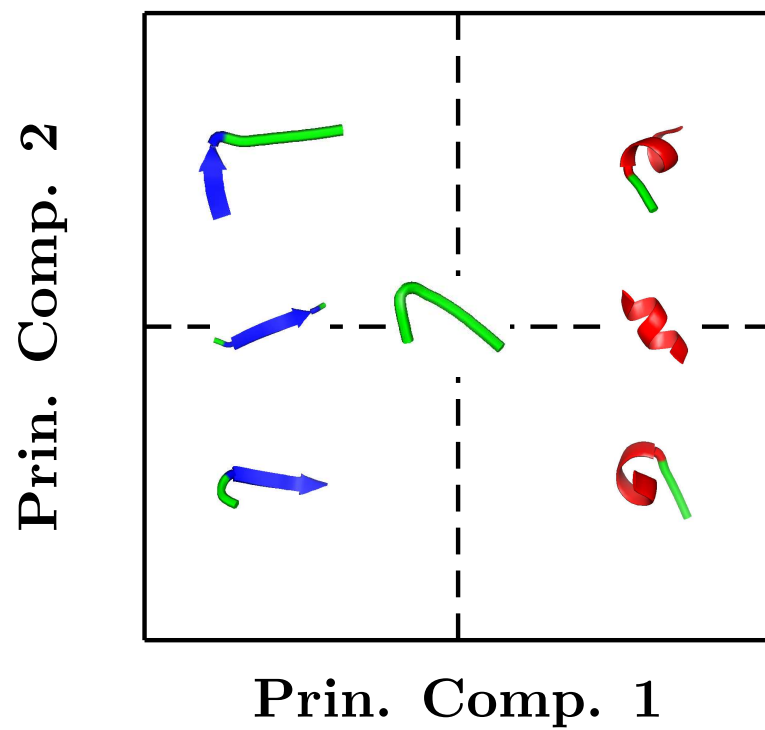


Fig. 3

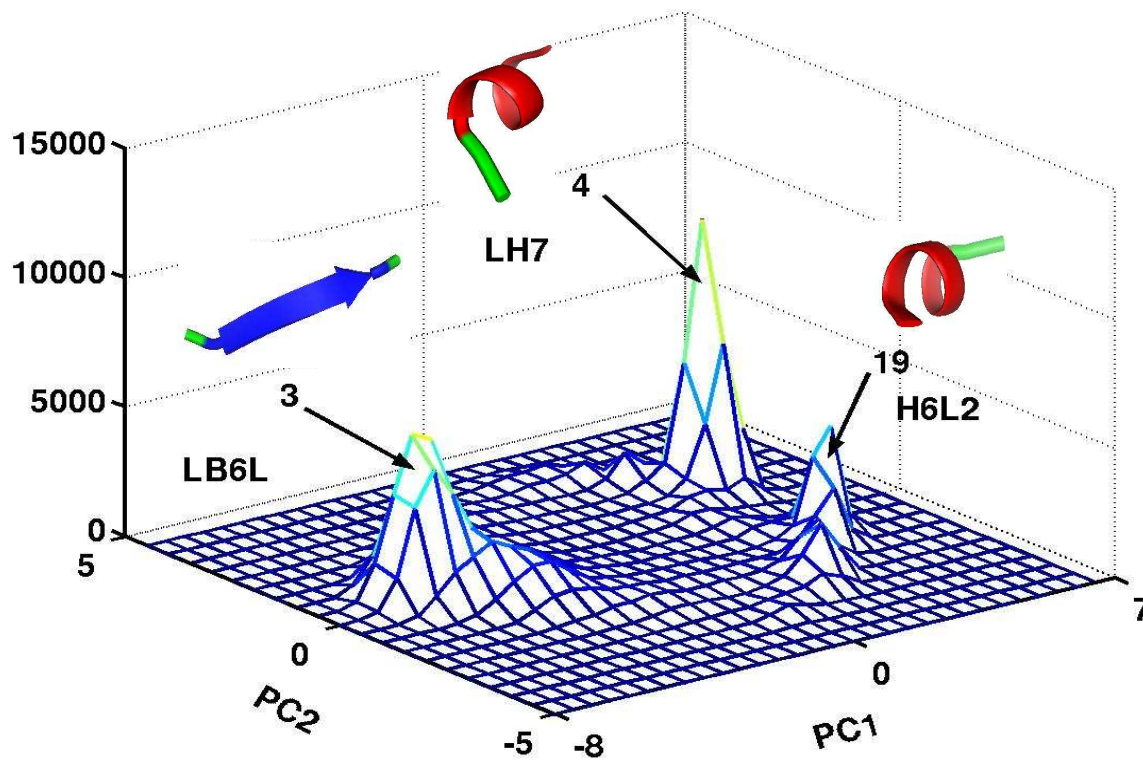
Clustering of Local Conformations

- K-means clustering with $K = 150$
- Input data matrix, $T_{N \times s}$ of N octapeptides with first s significant.
- Consensus secondary structures are found from the clusters.
- Clusters are visualized on a conditional bivariate probability distribution plot conditioned on PC_3 and PC_4 .

Local Conformations Occupancy Chart



Sample Bivariate Distribution



Conclusion

- Our method has proposed a general framework for visualization and analysis of local conformation space by using a peptide of arbitrary length as a unit local conformation.
- The conditional bivariate distribution plots provides visual blue-print of allowed and disallowed protein conformations.
- The visualization of restrictions in conformational space provides leads to guide conformational sampling in protein structure modeling.
- This is useful in checking the integrity of a predicted as well as experimentally deduced structure.

Publication

Geometric Invariant based Framework for Analysis of Protein Conformational Space, Ashish V Tendulkar, Babatunde Ogunnaike, Milind Sohoni, Pramod Wangikar, *Bioinformatics*.
Volume 21, Issue 18, Sept. 2005, Pages 3622-3628

Emails of Presenters

ashish@it.iitb.ac.in

pramodw@iitb.ac.in

The slides can be obtained from www.it.iitb.ac.in/~ashish/