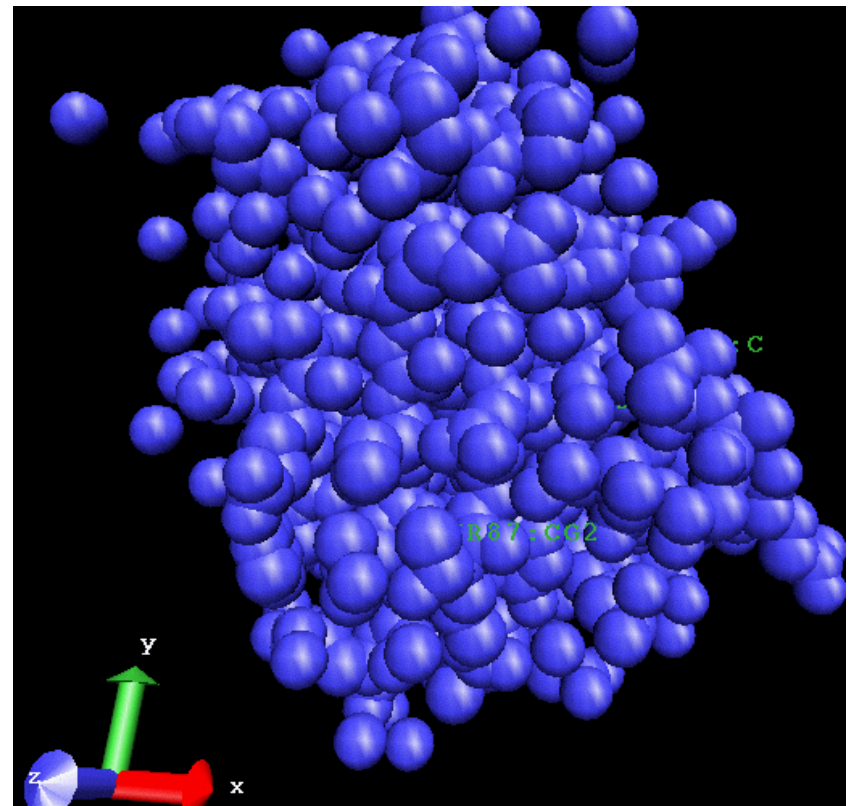
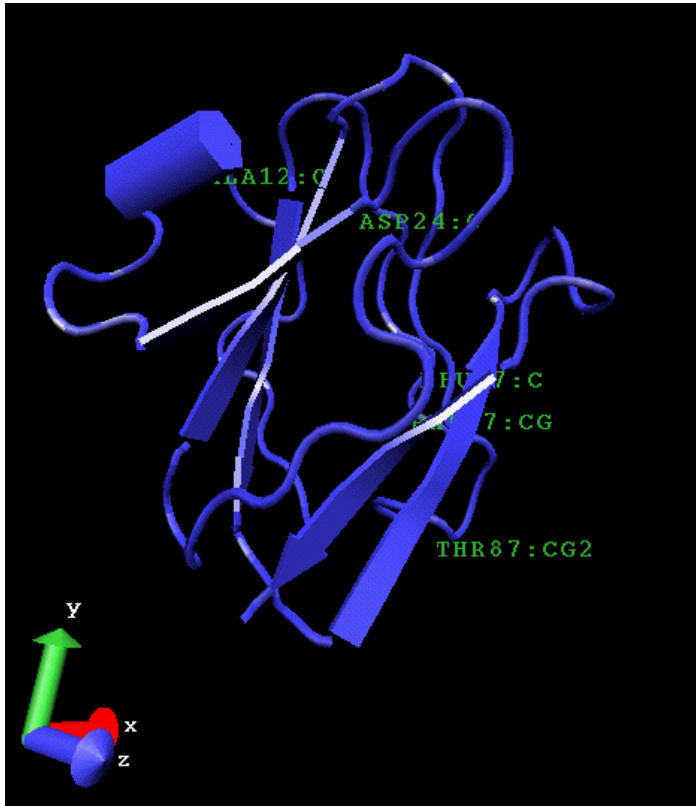


Lecture 3 - Structural Alignment of Proteins

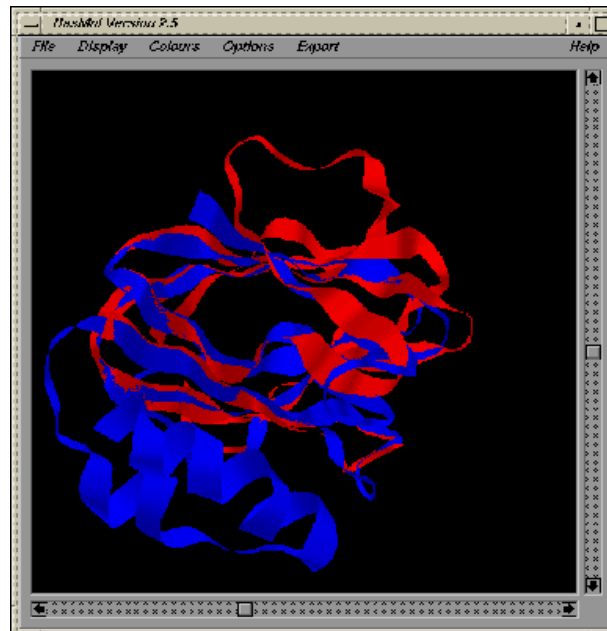
Rigid 3-D Object Matching and
Superposition

Protein Tertiary Structure



Protein Structural Alignment

The Rigid Case



Recommended Reading (1)

- W.R. Taylor and C.A. Orengo, Protein Structure Alignment, *J. Molecular Biology*, vol. 208, pp. 1-22, (1989).
- R. Nussinov, R. and H. J. Wolfson, Efficient detection of three-dimensional motifs in biological macromolecules by computer vision techniques, *Proc. Nat. Acad. Sc.*, vol. 88, pp. 10495-10499, (1991).
- G. Vriend, and C. Sander, Detection of Common Three-Dimensional Substructures in Proteins, *Proteins*, 11, pp. 52-58, (1991).

Recommended Reading (2)

- L. Holm, and C. Sander, Searching protein structure databases has come of age, *Proteins*, vol. 19. pp. 165-173, (1994).
- D. Fischer, R. Tsai, R. Nussinov, and H.J. Wolfson, A 3-D Sequence-Independent Representation of the Protein Databank, *Prot. Engineering*, vol. 8(10), pp. 981-997, (1995).
- I. Eidhammer, I. Jonassen, and W.R. Taylor, Structure Comparison and Structure Pattern, *J. Comp. Biology*, vol. 7(5), pp. 685-716, (2000).

Why bother with structures when we have sequences ?

- In evolutionary related proteins structure is much better preserved than sequence.
- Structural motifs may predict similar biological function .
- Getting insight into protein folding. Recovering the limited (?) number of protein folds.

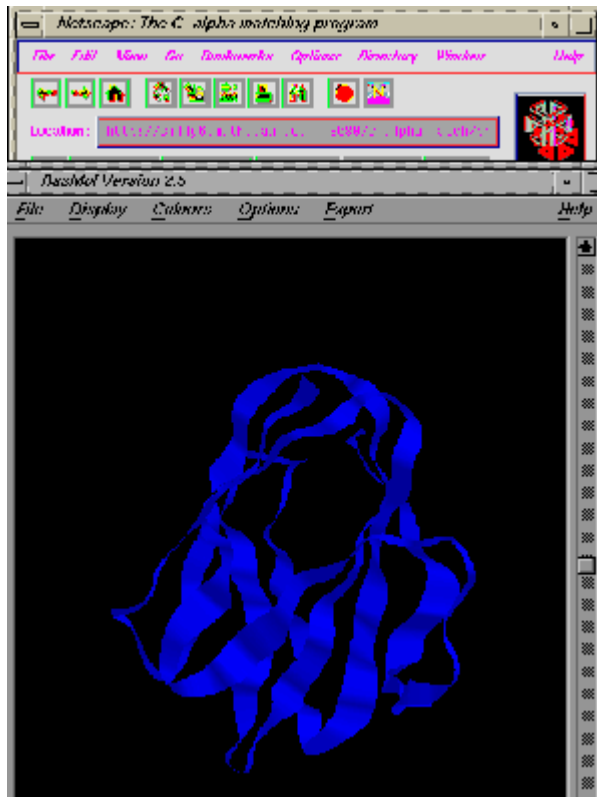
Who needs automated algorithms ?

- Emergence of *large* structural databases which do not allow manual (visual) analysis and require efficient 3-D search and classification methods.
- *Structural Genomics* effort.

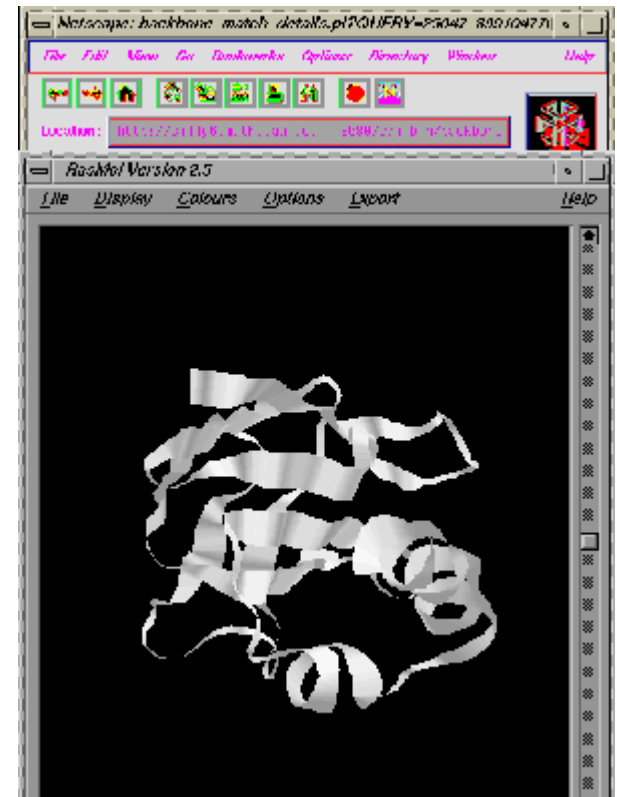
Additional Applications of Structural Alignment Methods

- Similar substructures in drugs acting on a given receptor - pharmacophore.
- Structurally similar receptor cavities could bind similar drugs.
- Docking.
- Biomolecular recognition.

Protein Structural Alignment Input



ApoAmicyanin - 1aaj



Pseudoazurin - 1pmy

Protein Shape Representation by Discrete 3D “critical features”

- Backbone C_{α} atomic centers.
- $C_{\alpha} \rightarrow C_{\beta}$ vectors.
- Secondary structure elements.
- Molecular surface representations.

The Major tasks in Structural Comparison :

- The correspondence (matching) task - difficult.
- The best superposition of matching features - minimal RMSD superposition has a closed solution.

Superposition - best least squares (RMSD) rigid alignment

Given two sets of 3-D points :

$P=\{p_i\}$, $Q=\{q_i\}$, $i=1,\dots,n$;

find a 3-D rotation R_0 and translation a_0 ,
such that

$$\min_{R,a} \sum_i |Rp_i + a - q_i|^2 = \sum_i |R_0p_i + a_0 - q_i|^2 .$$

*A closed form solution exists for this task.
It can be computed in $O(n)$ time.*

Several algorithms have been developed for the detection best RMSD 3-D rigid alignment both in Molecular Biology (Kabsch), Computer Vision (Schwartz and Sharir, Horn, Arun et al., Umeyama , Faugeras et al.).

The problem is related to the well known Procrustes problems in statistics and involves eigenvalue analysis of a correlation matrix of the points.

Solving the Correspondence (Matching) Problem

- Main difficulty arises because of the required *local* match in an a-priori unknown site.
- Exploit the fact that the objects handled are *rigid*.
- The correspondence of a pair of ordered triplets of points, which define (*fat enough*) **congruent triangles**, uniquely defines a 3-D rigid transformation.

Sequence order dependence

- Matching set should follow sequence order (Taylor and Orengo).
- Fragments of the chain (10-20 a.a) should follow sequence order (Vriend and Sander).
- Sequence order independent (Nussinov and Wolfson).
- Sec. str. elements are sequence order independent (Mitchell et al., Alesker, Nussinov and Wolfson).

Dynamic Programming SSAP

Orengo and Taylor (1989)

- For each residue define a local, rotation and translation invariant **structural** environment .
- For each pair of residues compute their similarity/distance based on their structural environments.
- Use the above computed distances as entries of a dynamic programming matrix.
- Find optimal path in the matrix.

Local - rigid motion invariant environment

- Represent each residue by the set of vectors btwn its C_β and the C_β atoms of all other residues in a fixed reference frame based on the C_α tetrahedral geometry of this residue.

Proximity between residues

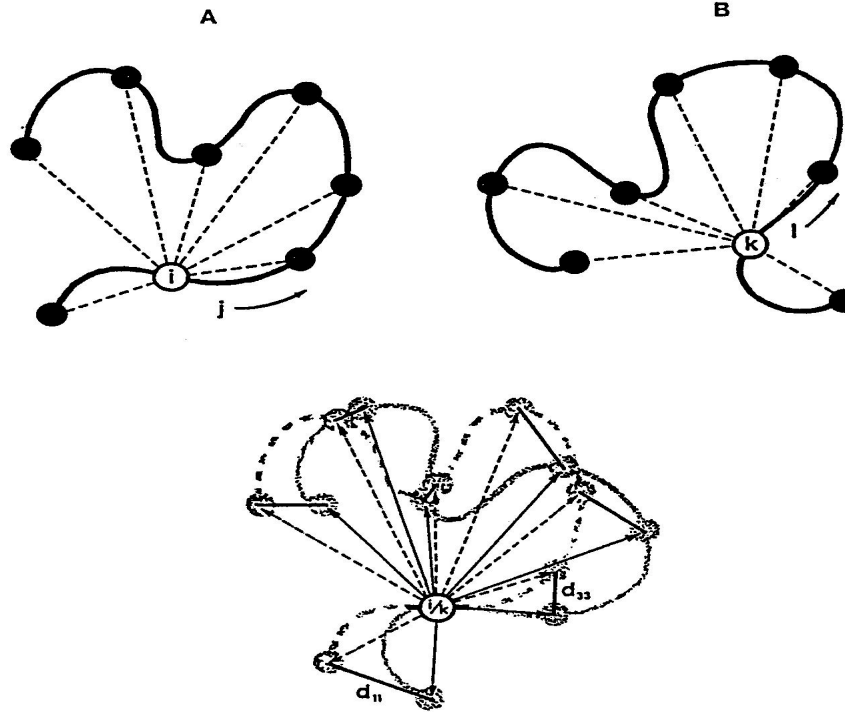


FIGURE 8. Structure comparison by the method of Taylor and Orengo. The two chains A and B are simple two dimensional representations of two similar protein structures. Two positions in these structures, i in A and k in B are compared. In C the structures are aligned on residues i and k and the distances between positions (all j in A and all l in B) are compiled in a matrix. To avoid confusion, only the distances between sequentially equivalent positions are drawn in C (these constitute the diagonal of the matrix). This matrix is then processed by a sequence alignment algorithm and the best correspondence of positions found. The process is repeated for all pairs of positions (all possible locations of i in A and k in B) and the results accumulated into an overall consensus alignment. Reproduced by kind permission of Protein Engineering.

Local environment similarity matrix

- Define similarity btwn two vectors $I \rightarrow V$ from protein **A** and $J \rightarrow W$ from protein **B** by $S_{ij} = a / (\Delta + b)$, where Δ is the length of their difference vector and a, b are constants (500, 10 respectively).

Double Dynamic Programming

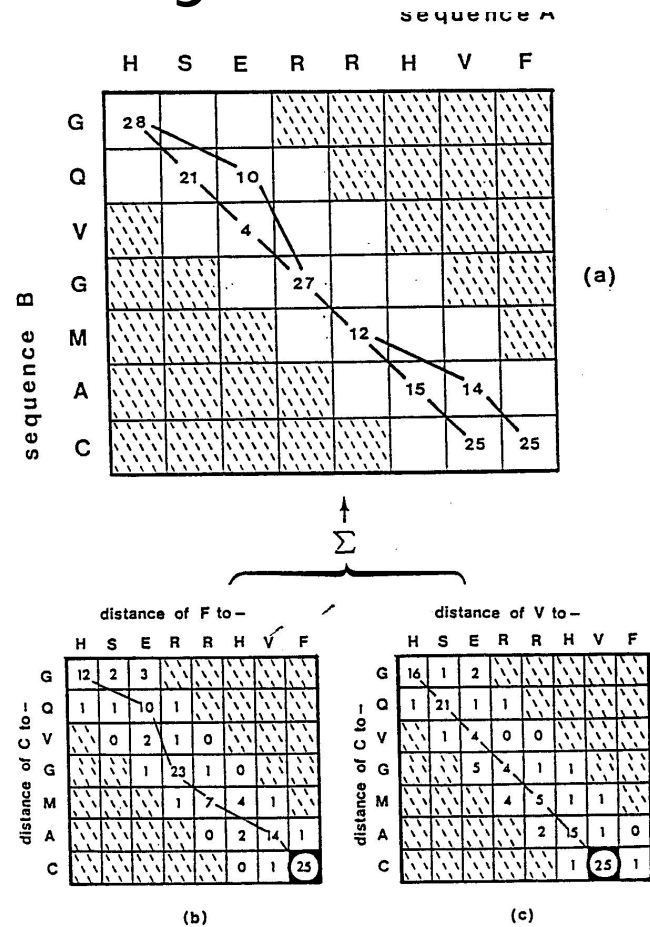


FIGURE 9. Application of the dynamic programming method to structure alignment in the method of Taylor and Orengo¹.

The DDP used by SSAP

- Detection of best equivalence between a pair of residues, e.g. (b) represents the comparison of all the distances viewed from residue C (in protein B) with all the distances centered at residue F (in A); [c] represents a similar matrix for residue C (in B) with residue V (in A).
- Score of the best path is the entry of the matrix in (a).

Sequence Order Independent Matching - Geometric Task :

**Given two configurations of points in the three dimensional space,
find those rotations and translations of one of the point sets which produce “large” superimpositions of corresponding 3-D points.**

Remarks :

The superimposition pattern is not known *a-priori* - pattern detection.

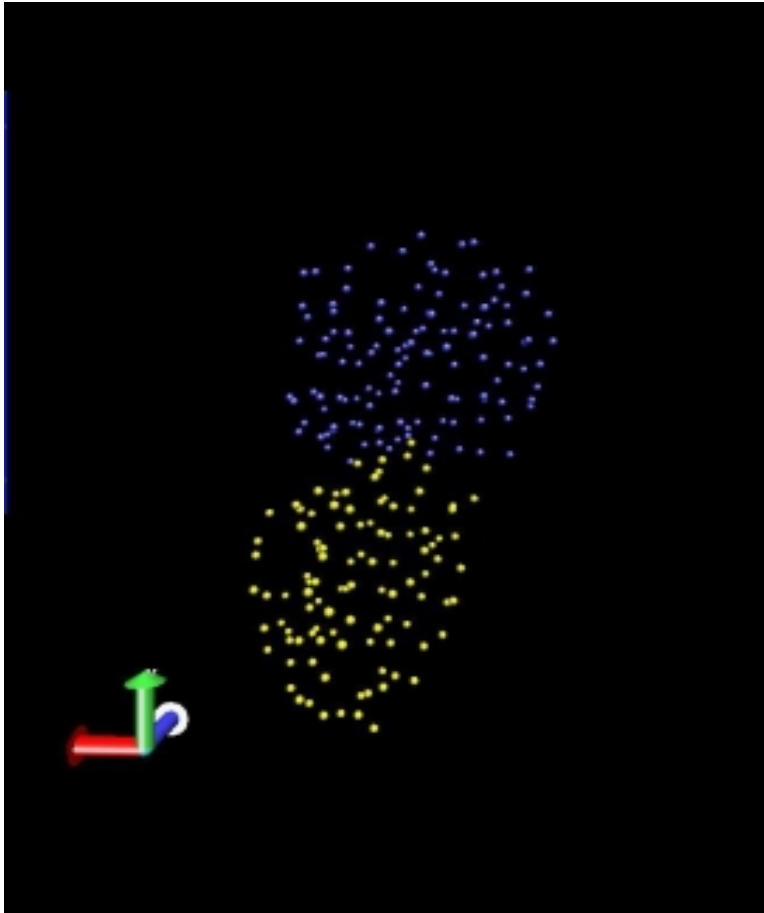
We are looking not necessarily for the largest superimposition, since other matchings may have biological meaning.

Analogous to local similarity in sequence alignment.

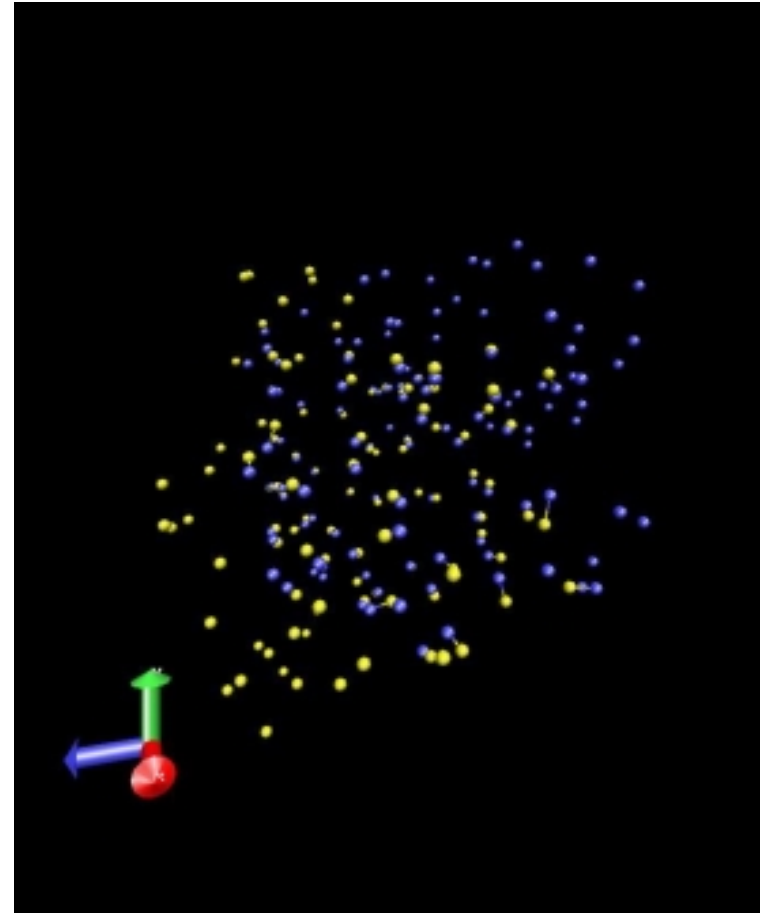
Sequence order dependence vs independence - geometric complexity

- Sequence order dependent alignment = 3-D *curve* matching - an inherently 1-D task.
- Sequence order independent alignment - a “real” 3-D task.

Sequence Independent Approach

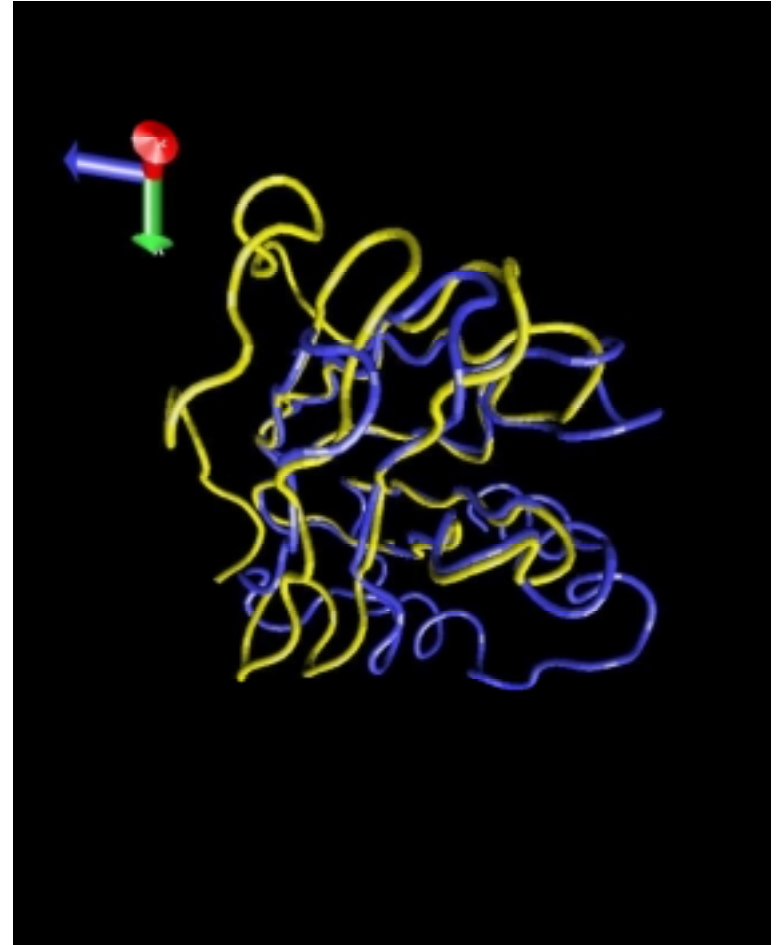
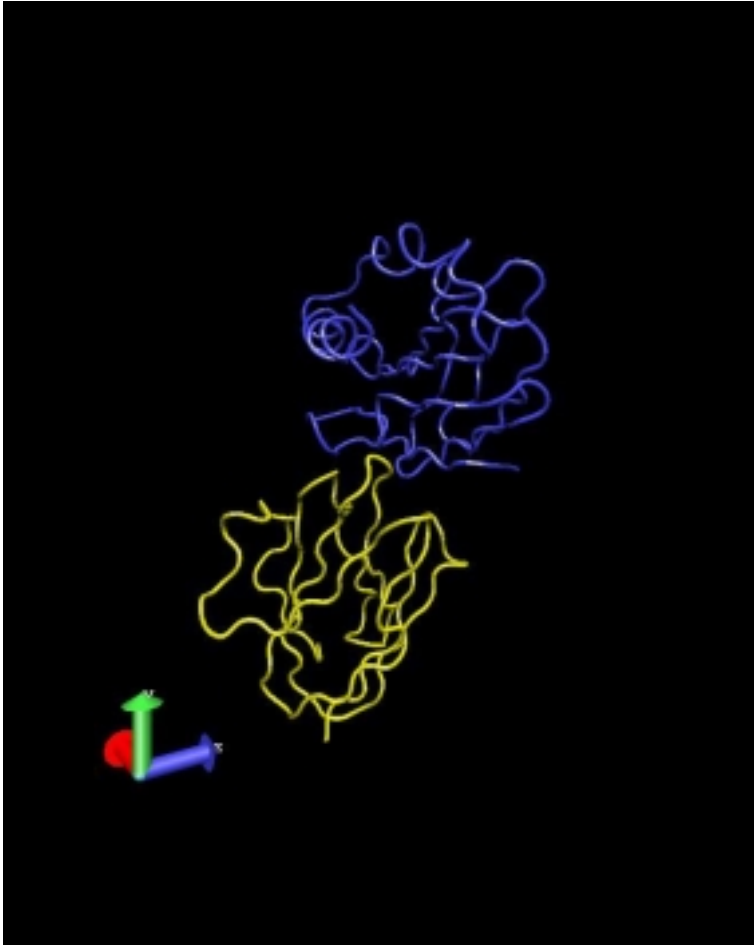


C_{α} constellations - before



Superimposed constellations

Backbone Trace



Advantages of the sequence independent alignment

- Enables detection of non-sequential motifs in proteins, e.g. molecular surface motifs, especially, similar binding sites.
- Allows search of structural databases with only **partial and disconnected** structural information.
- Same algorithm applies to other molecular structures, e.g. drugs.

Solution of the superimposition

File Edit View Go Bookmarks Options Directory Window Help

Location: http://silly6.math.tau.ac.il:8080/cgi-bin/backbone_match.cgi?PR

Mail What's New? What's Cool? Destinations Net Search Welcome

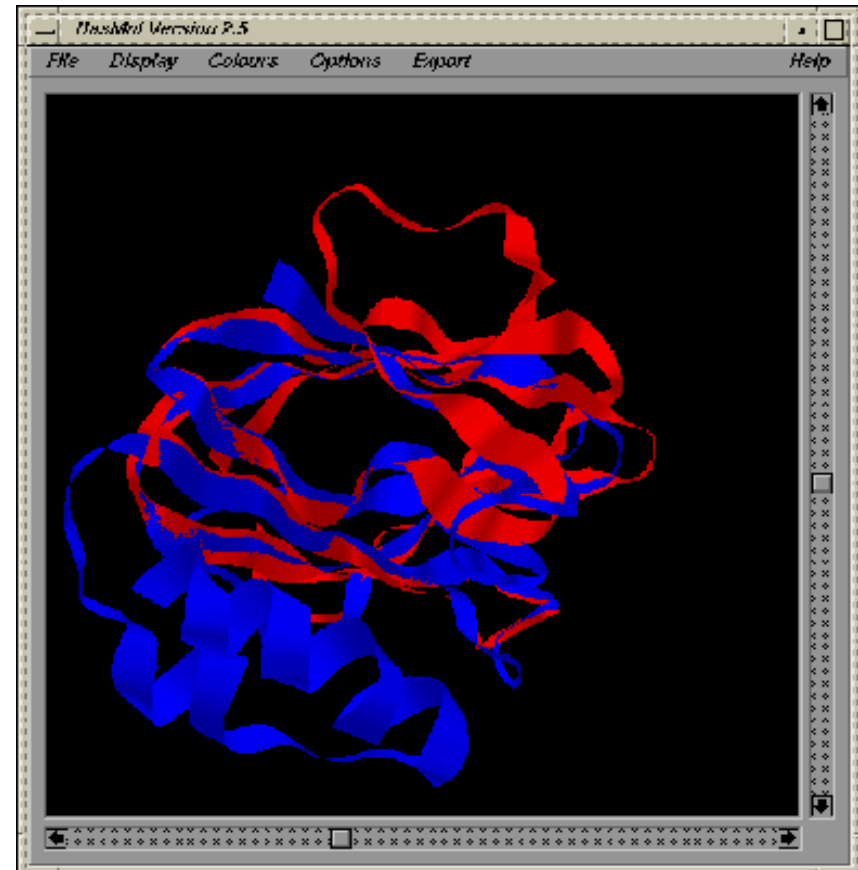
Results for matching 1PMY with 1AAJ

Results

#	Score	Match Size	RMS	Rotation			Translation		
Result 1	78.00	78	1.44	1.178	-0.059	-2.615	30.230	14.864	17.912
Result 2	61.00	61	2.05	-0.952	0.393	0.832	-1.717	7.031	-8.936
Result 3	60.00	60	1.82	1.999	-0.582	0.353	-4.668	19.664	30.651
Result 4	48.00	48	2.14	1.270	0.128	-2.818	31.964	9.542	13.826
Result 5	47.00	47	1.86	1.544	-1.136	-0.213	6.629	25.553	27.876
Result 6	45.00	45	1.82	-1.323	-0.208	0.324	-14.846	6.148	-1.035
Result 7	43.00	43	1.64	-1.133	0.232	0.291	-7.763	7.696	-11.996
Result 8	42.00	42	1.57	1.535	0.090	-3.050	26.773	7.332	12.549
Result 9	41.00	41	1.89	-2.113	0.924	-2.010	6.091	33.671	-16.447
Result 10	41.00	41	1.82	-2.066	0.590	-1.669	6.098	37.415	-6.752

- Rotations are given in radians for X-Y-Z axes. Rotating space around the X-axis, then around the Y-axis and finally around the Z-axis would give the required rotation.
- X-Y-Z translation coordinates are given in Angstrom units.

http://silly6.math.ta...78_899105078&RESULT=5



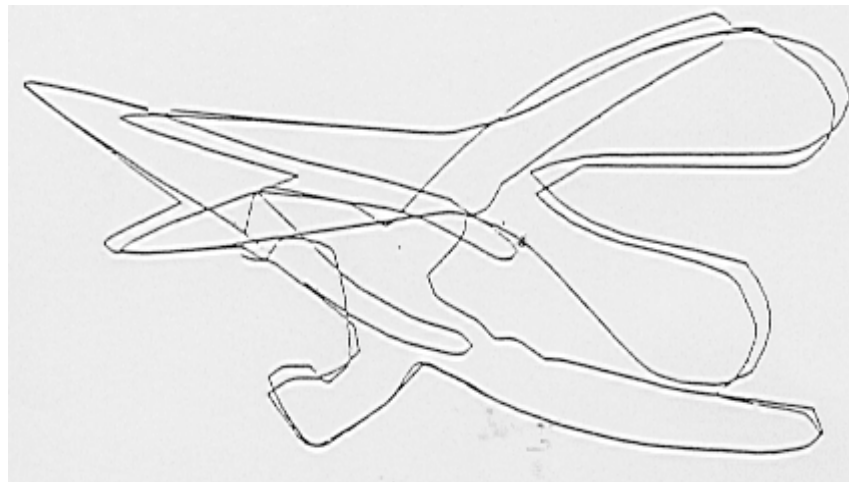
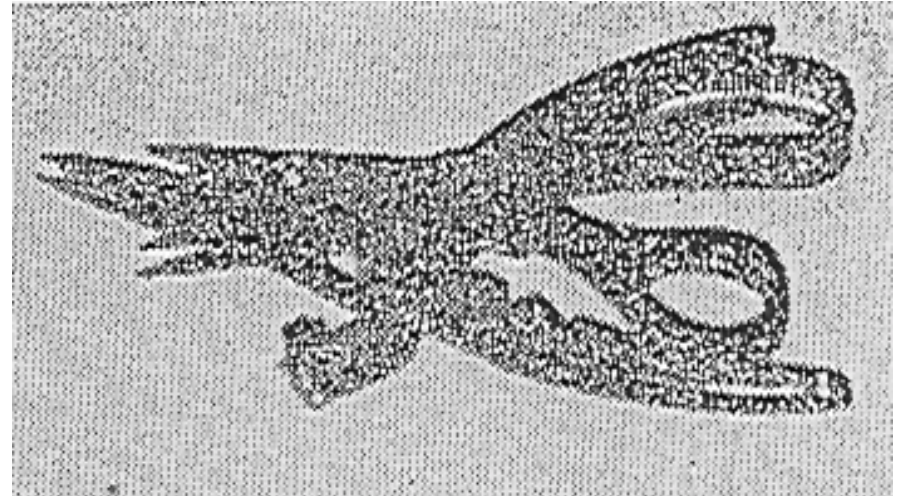
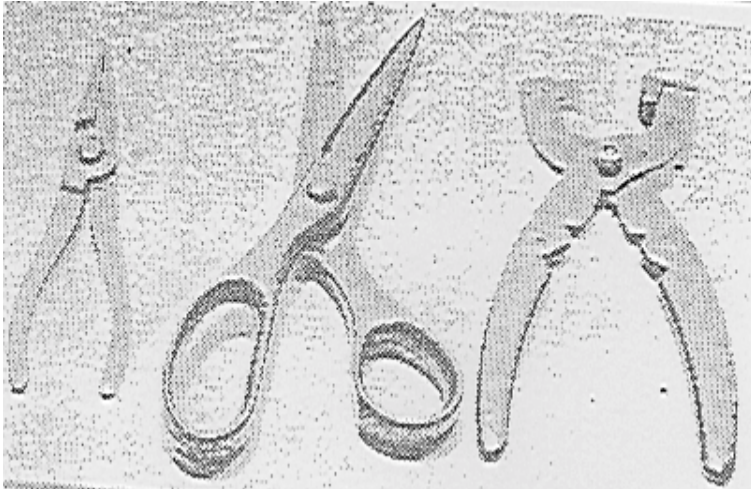
Potential disadvantages of neglecting sequential order info

- Motifs preserving sequence order might be biologically more meaningful than similar size non-sequential motifs.
- The computational task becomes much more complex, when sequence order is not exploited.

Answer :

If the use of sequence order is advantageous, one can always exploit it. This info does not disappear, and can be incorporated.

Analogy with Object Recognition in Computer Vision



Straightforward Algorithm

- For each pair of triplets, one from each molecule which define ‘almost’ congruent triangles **compute the rigid motion that superimposes them.**
- Count the number of point pairs, which are ‘almost’ superimposed and sort the hypotheses by this number.

Naive algorithm (continued)

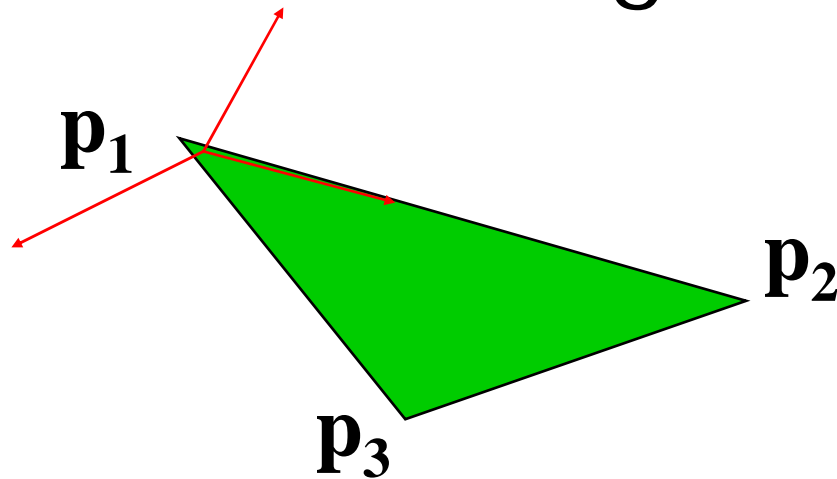
- **For the highest ranking hypotheses improve the transformation by replacing it by the best RMSD transformation for all the matching pairs.**
- ***Complexity : assuming order of n points in both molecules - $O(n^7)$.***

($O(n^3)$ if one exploits protein backbone geometry.)

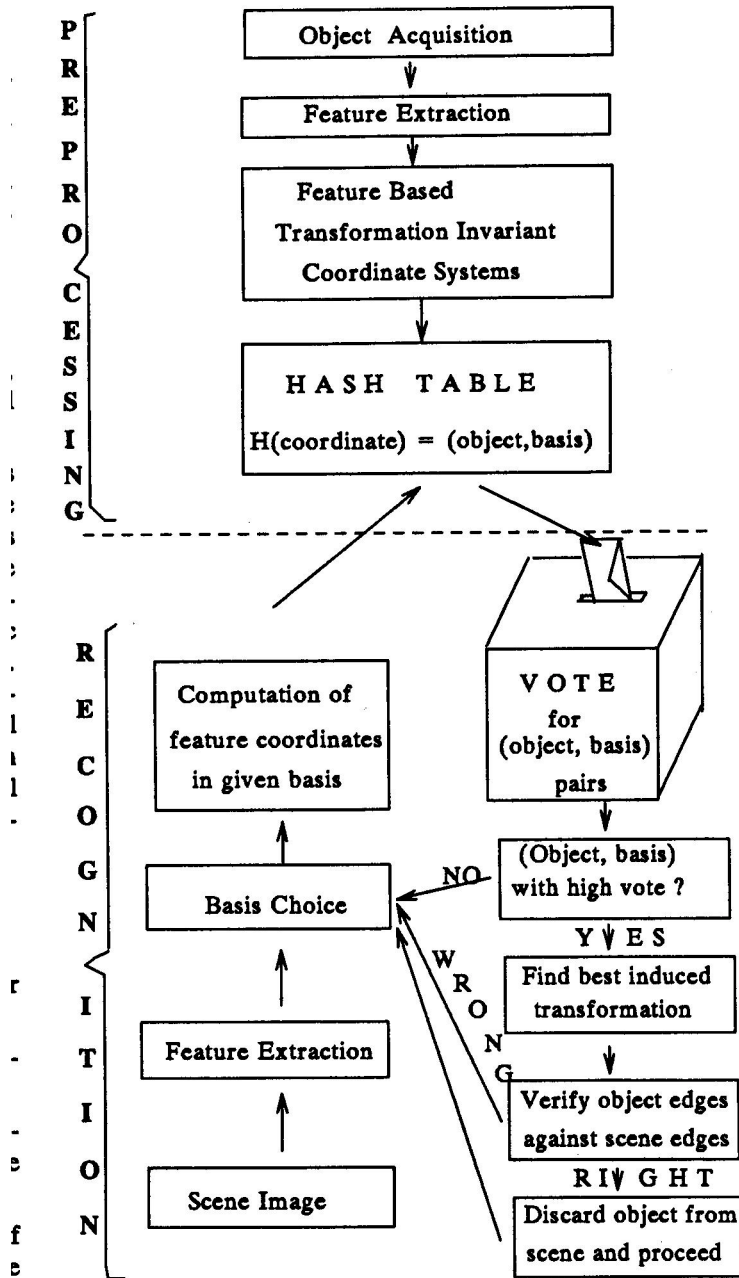
Geometric Hashing

- Developed for object recognition in Computer Vision (Lamdan, Schwartz, Wolfson, 1988 - rigid, Wolfson, 1991 -flexible).
- Adapted to Molecular Biology (Nussinov, Wolfson, 1989).
- Motivated by *associative memory* ideas and efficient *hashing* techniques.

A 3-D reference frame can be uniquely defined by the ordered vertices of a non-degenerate triangle



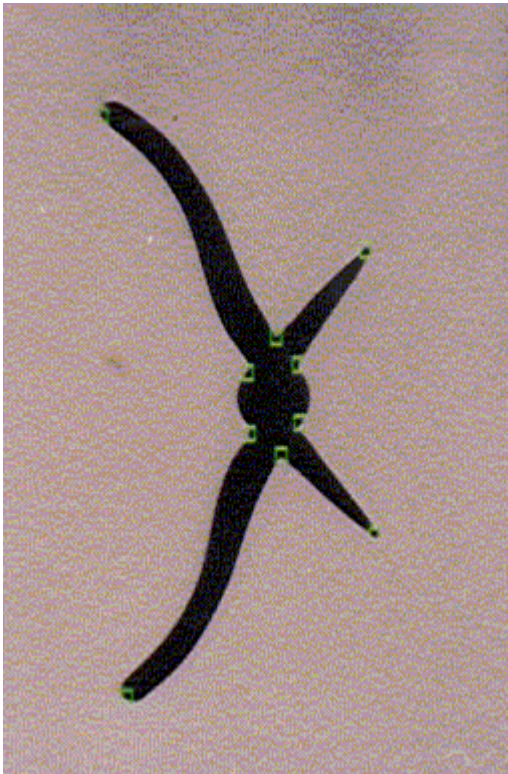
The lengths of the triangle sides are rigid motion *invariant*.



Lamdan & Wolfson, Geometric Hashing, ICCV'88

Figure 1 : The general scheme of the object recognition algorithm.

Model Database



Scene



Recognition



Lamdan, Schwartz, Wolfson, “Geometric Hashing”, 1988.

Protein Structure Alignment

- Define local neighborhoods of residues (in practice an annulus defined by min and max radii).
- Using Geometric Hashing detect seed matches defined by a transformation and a match-list.
- Cluster *seed matches* and merge match-lists.
- Extend the seed matches and detect best RMSD transformations.
- Iterate last step.

Geometric Hashing - Preprocessing

- Pick a *reference frame* .
- Compute the coordinates of all the other points (in a pre-specified neighborhood) in this reference frame.
- Use each coordinate as an address to the hash (look-up) table and record in that entry the (protein, ref. frame, shape sign., point).
- Repeat above steps for each *reference frame*.

Geometric Hashing - Recognition 1

For the target protein do :

- Pick a *reference frame* satisfying pre-specified constraints.
- Compute the coordinates of all other points in the current *reference frame* .
- Use each coordinate to access the hash-table to retrieve all the records (prot., r.f., shape sign., pt.).

Geometric Hashing - Recognition 2

- For records with matching shape sign. “vote” for the (protein, r.f.).
- Compute the transformations of the “high scoring” hypotheses.
- Repeat the above steps for each r.f.

Complexity of Geometric Hashing

N- number of structures (proteins).

O(n)- no. of “features” in a structure.

R - no. of reference frames (bases).

Typically, $R = n, n^2, \text{ or } n^3$.

If the reference frame is based on more than one point additional invariants (shape signatures) arise, e.g. for 2 pts. - distance; for a triplet - triangle sides length.

Complexity (continued)

Preprocessing: $O(N * R * n)$.

**Match Detection/Recognition :
 $O(R * n * s)$.**

s - size of a hash-table entry. Can be kept low by not processing “fat” entries. These entries are known in advance after *Preprocessing*.

Advantages :

- Sequence order independent.
- Can match partial disconnected substructures.
- Pattern detection and recognition.
- Highly efficient.
- Can be applied to protein-protein interfaces, surface motif detection, docking.

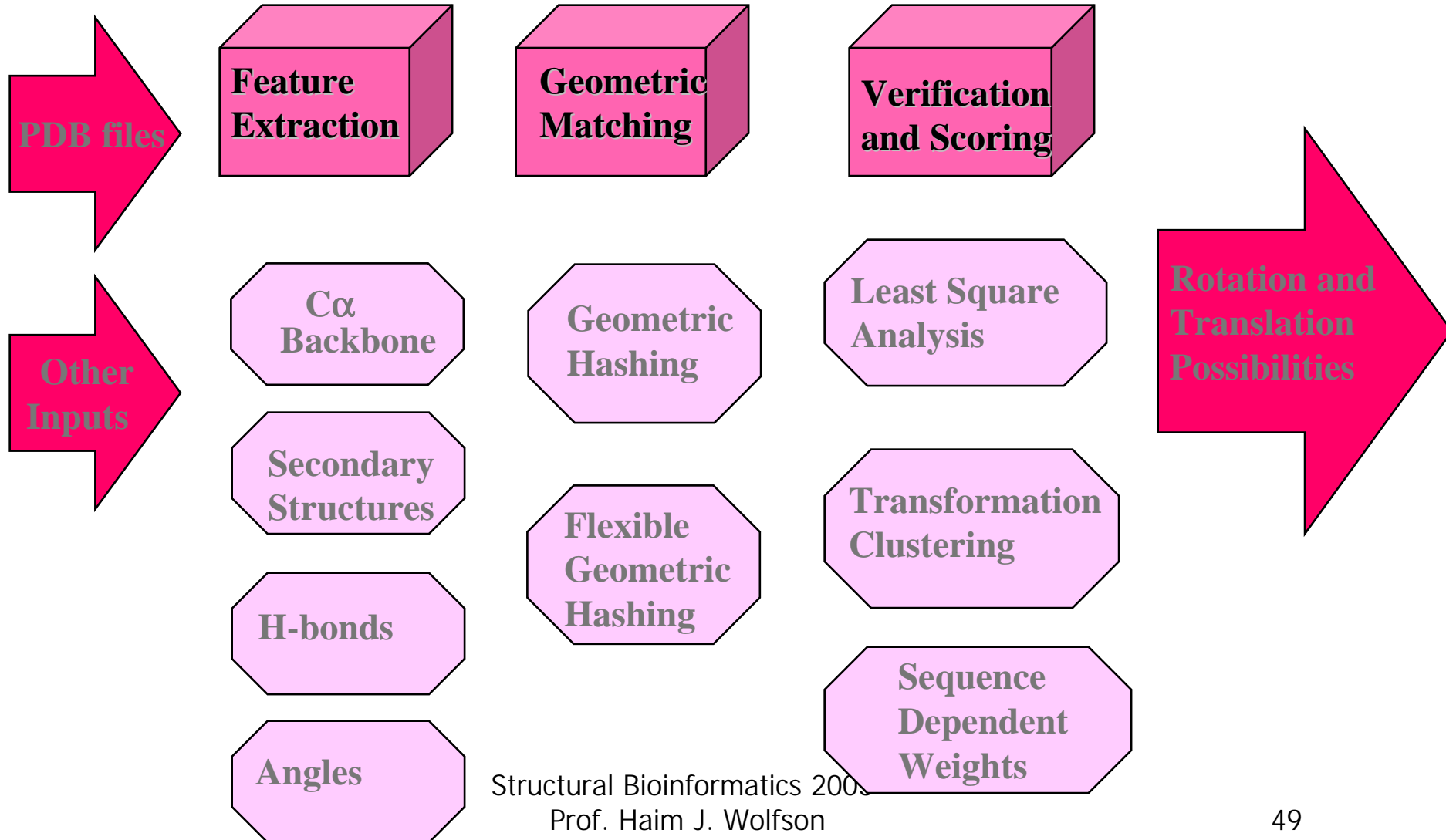
Structural Comparison Algorithms implemented with GH

- C_{α} backbone matching.
- Secondary structure configuration matching.
- Structural comparison of protein-protein interfaces.
- A representative set of the PDB monomers and interfaces.

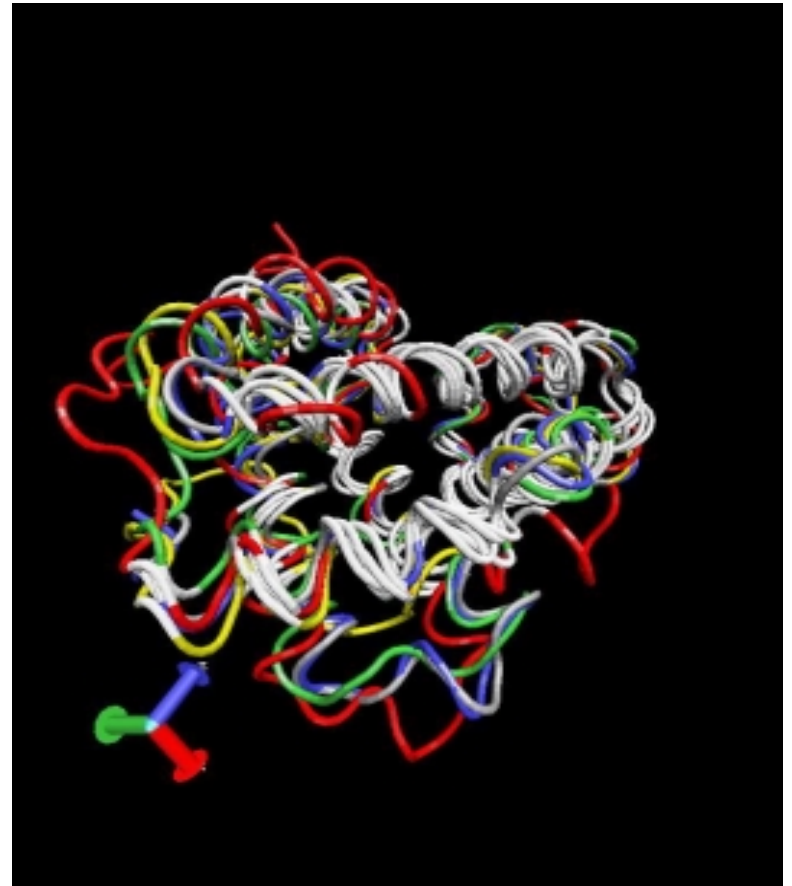
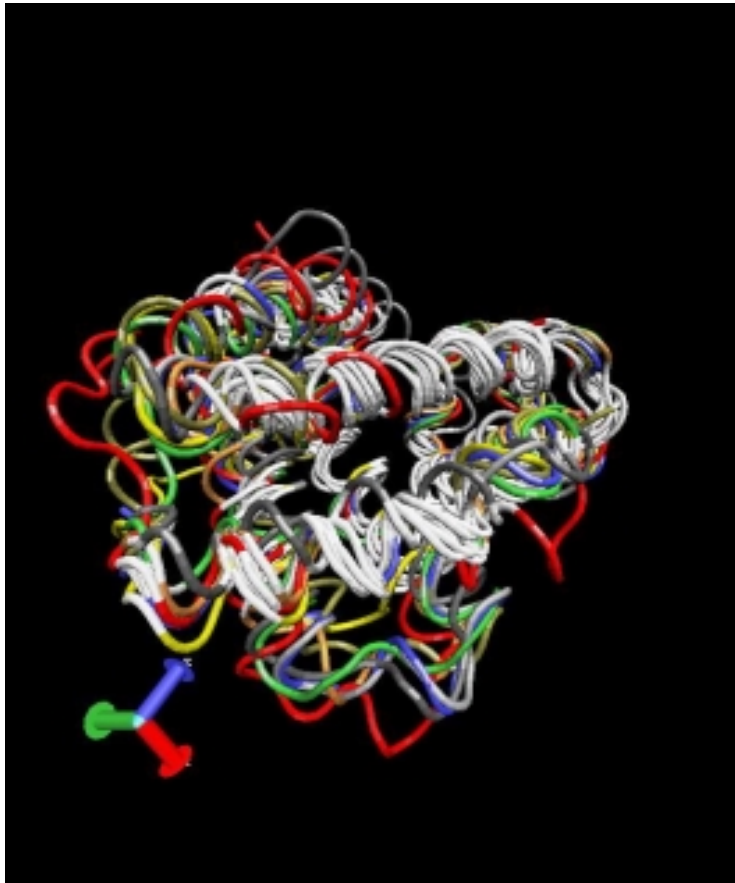
Structural Comparison Algorithms (continued)

- Amino acid substitution matrices based on structural comparison statistics.
- Molecular surface motifs.
- Multiple **Structure** Alignment.
- Flexible (Hinge - based) structural alignment.

Protein Structural Alignment



Multiple Structural Alignment Globins



Multiple Structural Alignment

Tim Barrels

