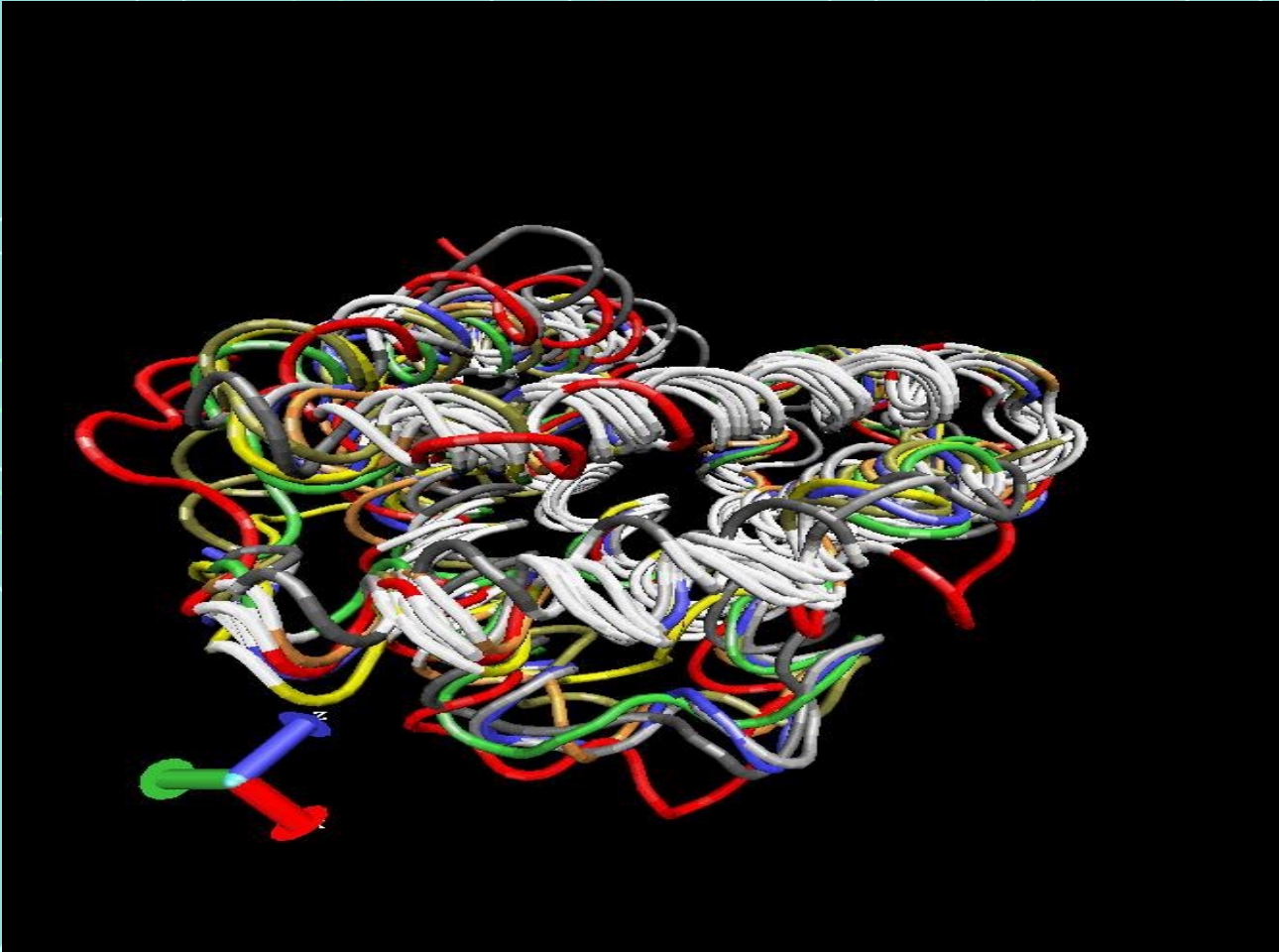


# Multiple Structural Alignment and Core Detection by Geometric Hashing



N. Leibowitz  
Z.Y. Fligelman  
R. Nussinov  
H.J. Wolfson

# Example of Multiple Structural Alignment





# Lecture Outline

- Problem definition and motivation.
- Previous work.
- Our MSTA Algorithm.
- Experimental Results.
- Summary.



# Definition of the problem

- Input:
  - N molecules ( $C_{\alpha}$  atom locations).
  - Min size of geometric core structure required.
- Output:
  - N-1 rigid transformations superimposing the molecules on a reference molecule.
  - A list of multiply aligned ( $C_{\alpha}$ ) atoms.



# Motivation

- Structural Analysis of protein ensembles.
- Fold databases for threading type techniques.
- Pharmacophore detection in drug ensembles.
- Active sites structural alignment.



# Pairwise Structural Alignment Methods

- Dynamic and double dynamic programming (e.g. Taylor and Orengo 1996).
- Pairwise inner distance matrix (DALI - Holm & Sander 1994).
- Geometric Hashing - sequence independent (Wolfson & Nussinov 1991).



# Previous Work on Multiple Structure Alignment

- Clustering of pairwise alignments:
  - Taylor and Orengo (mSSAP 1996);
  - Gerstein and Levitt (1996).
- Core detection given a sequence alignment:
  - Gelfand and Kister (1998);
  - Gerstein and Altman (1995).

# Comparison of Stages

## between string based and structural approach

- Detection of correspondence by multiple *string* alignment.
- Superimposition of the *structures*.

⊗ Extension and refinement of the initial core.

ॐ Detection of correspondence and superimposition of *structures*.

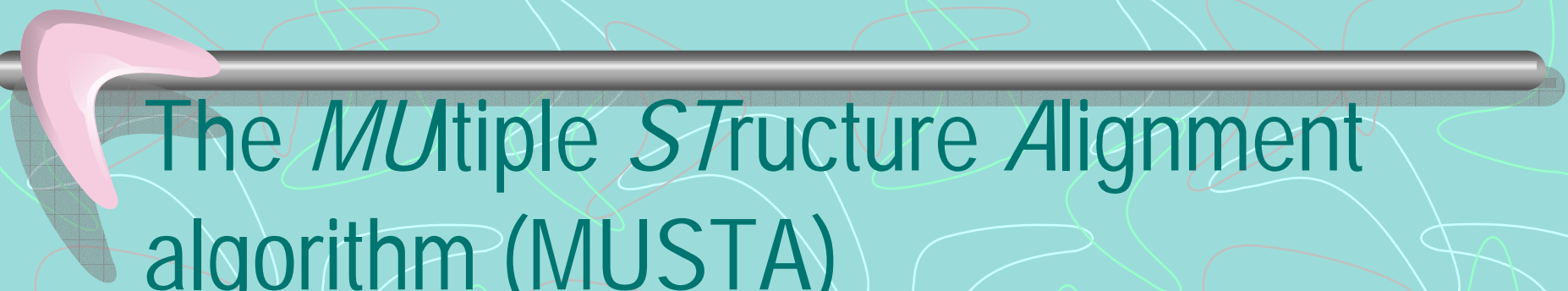
⊗ Extension and refinement of the initial core



# Problems to be solved

Correspondence - multiple alignment.

Superimposition - multi-transformation.



# The *M*ultiple *S*tructure Alignment algorithm (MUSTA)

- Input:  $N$  molecules,  $M_i$   $i=1\dots n$ .
- One molecule (e.g  $M_1$ ) - *reference molecule*.
- Output:  $(N-1)$ -vector of rigid transformations  
( $T_{12}$  ,  $T_{13}$  , ... ,  $T_{1n}$ ) - a *multi-transformation*.



# Algorithm Outline

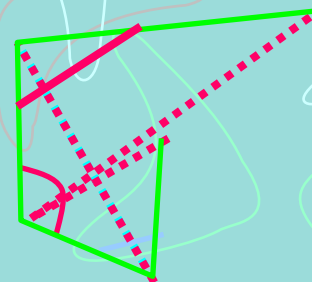
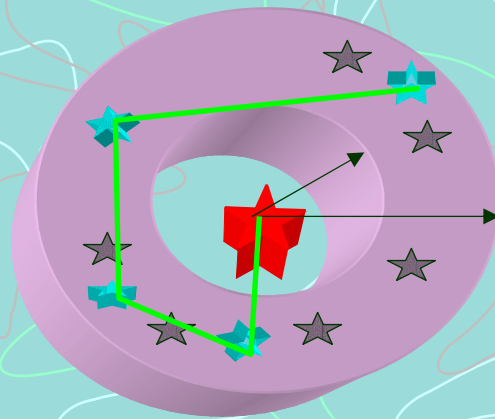
- Detection of congruent seed substructures and candidate multi-transformations.
- Clustering of the transformations and extension of the seed matches.
- Detection of the highest scoring hypotheses.



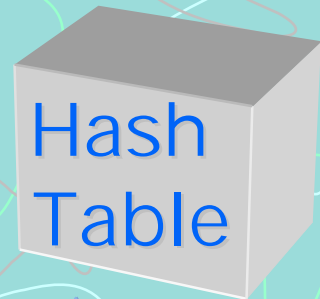
# Detection of Seed Matches

Build a look-up (hash) table of local  
geometric substructures, addressed  
by their (rigid motion) invariants.

# The Hashing Mechanism



(d1,d2,d3,d4)  
(a1,a2,a3)  
(t1,t2)





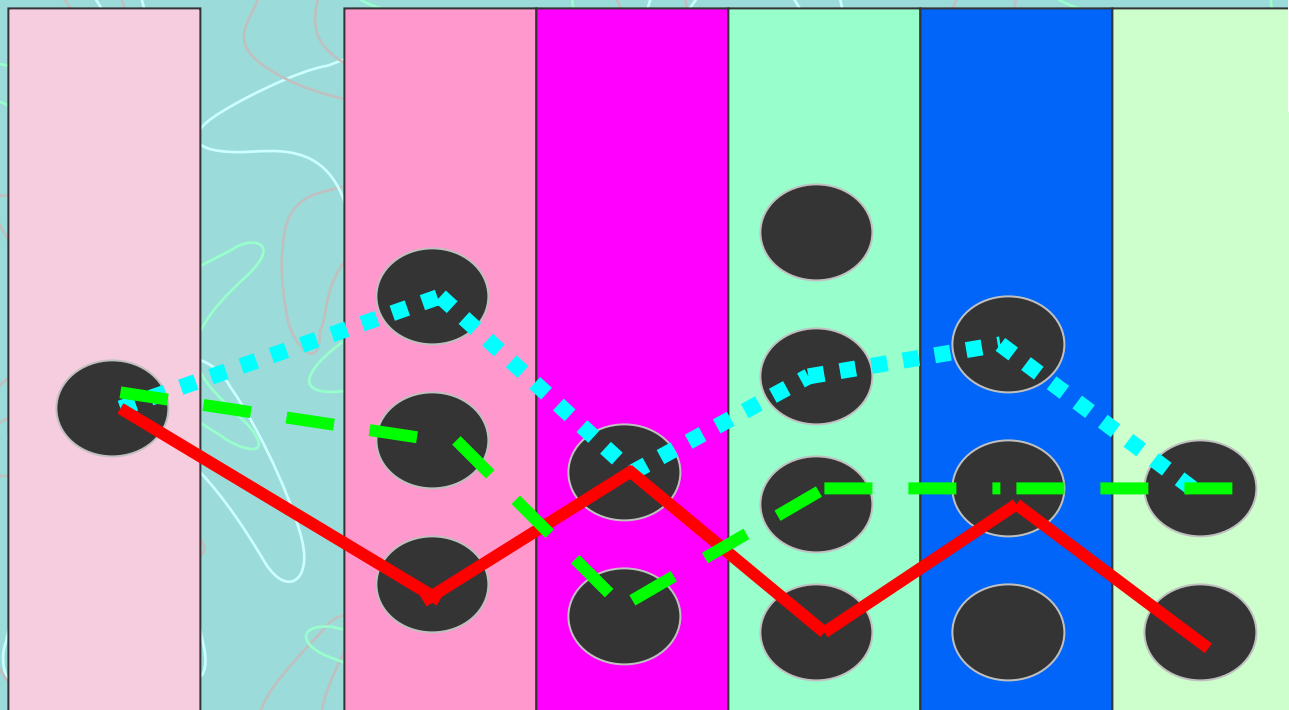
# Definition of Combinatorial Buckets

- Each  $k$ -tuple of the reference molecule defines a bucket.
- The bucket members are all the congruent  $k$ -tuples (from the other molecules)

# The Combinatorial Bucket

Number of potential multi- transformations in each combinatorial bucket:

$$\prod_{i=1}^{N-1} m_i$$



Reference  
k-tuple


Matching k-tuples from different molecules

Structural Bioinformatics - 2003;  
Prof. Haim J. Wolfson



# Candidate Multi-Transformation

- Every multi-transformation defined by a path in the bucket has *at least* **k**  $C_{\alpha}$  atoms as a *geometric core*.
- The solution space is restricted only to buckets that have at least one representative k-tuple from **each** molecule.



# Clustering and extending the seed matches

- Problems with the parametric clustering method:
  - Assignment of relative weights
  - numerical stability
  - compensation between rotations and translations.
- Solution: define the clustering in the means of the goal being achieved



# Component-wise Clustering of Multi-Transformations

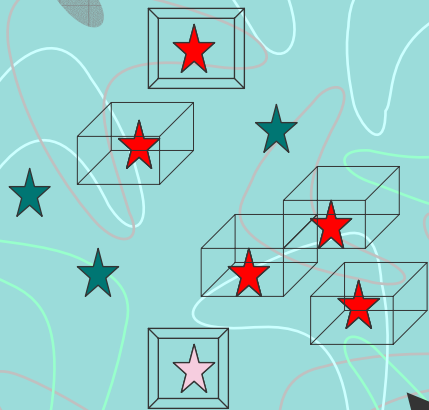
- Goal: as large as possible geometrically congruent core.
- Criteria: transformations from the same cluster map relevant 3-D points to *almost identical locations*.
- Distance between transformations: number of atoms mapped to different locations.

# Distance Between Transformations

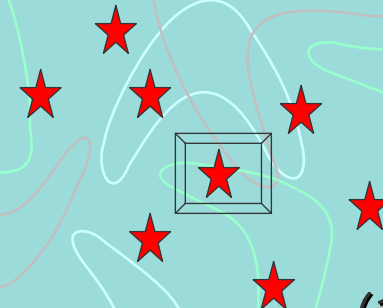
ref Mol

Mol i

Mol i



- (3,8)
- (4,9)
- (6,11)
- (1,5)
- (8,10)



- (1,5)
- (3,8)
- (4,9)
- (6,11)
- (17,12)



Agree : 4 pairs

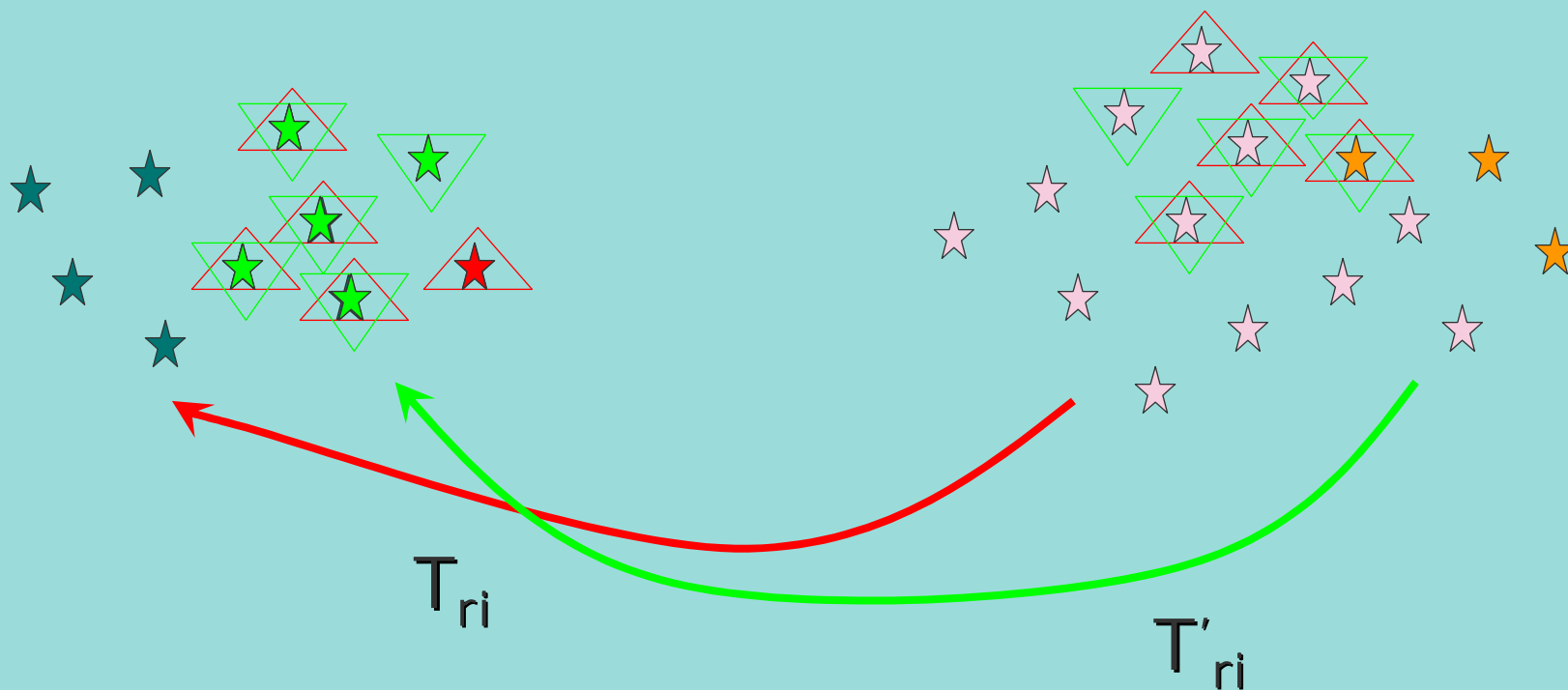
Disagree : 1 pair of each

Structural Bioinformatics - 2003;  
Distance = 2 pairs

# Distance Between Transformations

ref Mol

Mol i



Agree : 4 pairs

Disagree : 1 pair of each

Distance = 2 pairs



# Distance between transformations

- **Consistent pair** – a pair of atoms, where the first one is mapped to the vicinity of the second.
- All the consistent pairs of a transformation are the **transformation's mask**.
- The **distance** between transformations is the number of consistent pairs which are not shared by both lists = the size of the “***symmetrical difference***” between the transformations' masks.




# Clustering

- The transformations are clustered iteratively.
- At each step - *calculate a representative prototype* for each cluster.
- Prototype == The best superimposition of the union of the cluster consistent match list pairs in the least square (RMSD) sense.
- **NOTICE**: all the clusters emerge from match lists of atoms configurations appearing in all the molecules.



# Computing the highest score

- Restrict search to the combinatorial buckets.
- Remove buckets which have not survived the clustering.
- **Intersect** the match lists of the prototype transformations.
- Keep the ranked results that are **above a threshold**.



# Computing Highest Score- Pseudo Code

Replace transformations in combinatorial buckets by their prototypes.

Reduce bucket complexity by removing irrelevant transformations

For each combinatorial buckets

For all the combinations it defines

Compute match list intersection

If above a minimum

store as a solution

End-if

End-for

End-for



# Experimental Results

- Using the SCOP classification of the PDB (Murzin et al. 1995) we started a series of experiments going up from the deepest level of the SCOP tree.

The levels in the SCOP tree are:

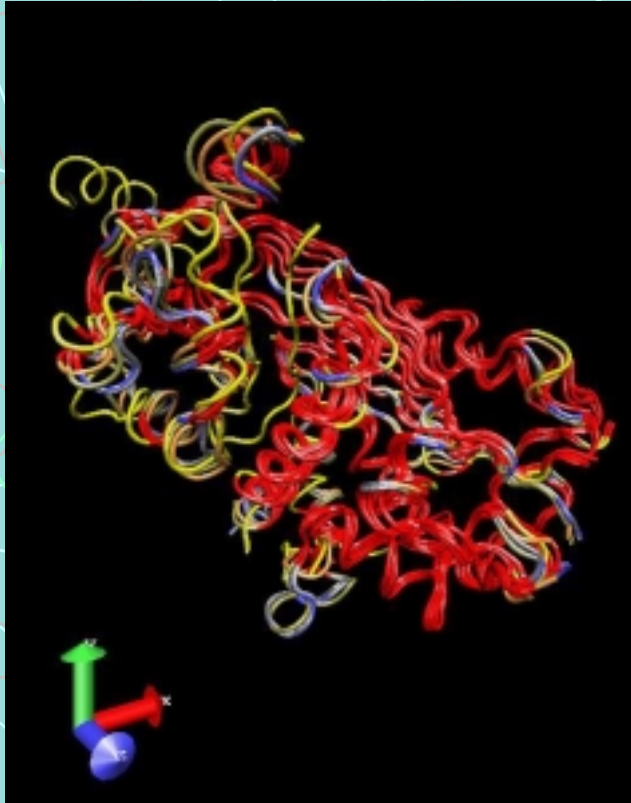
- Classes
  - folds - the alpha helix bundle experiments.
  - super-families - the TIM-BARREL experiments.
    - families - the cal-binding and globins examples.
      - » proteins - the serpins examples.
      - » species



# Experimental Results

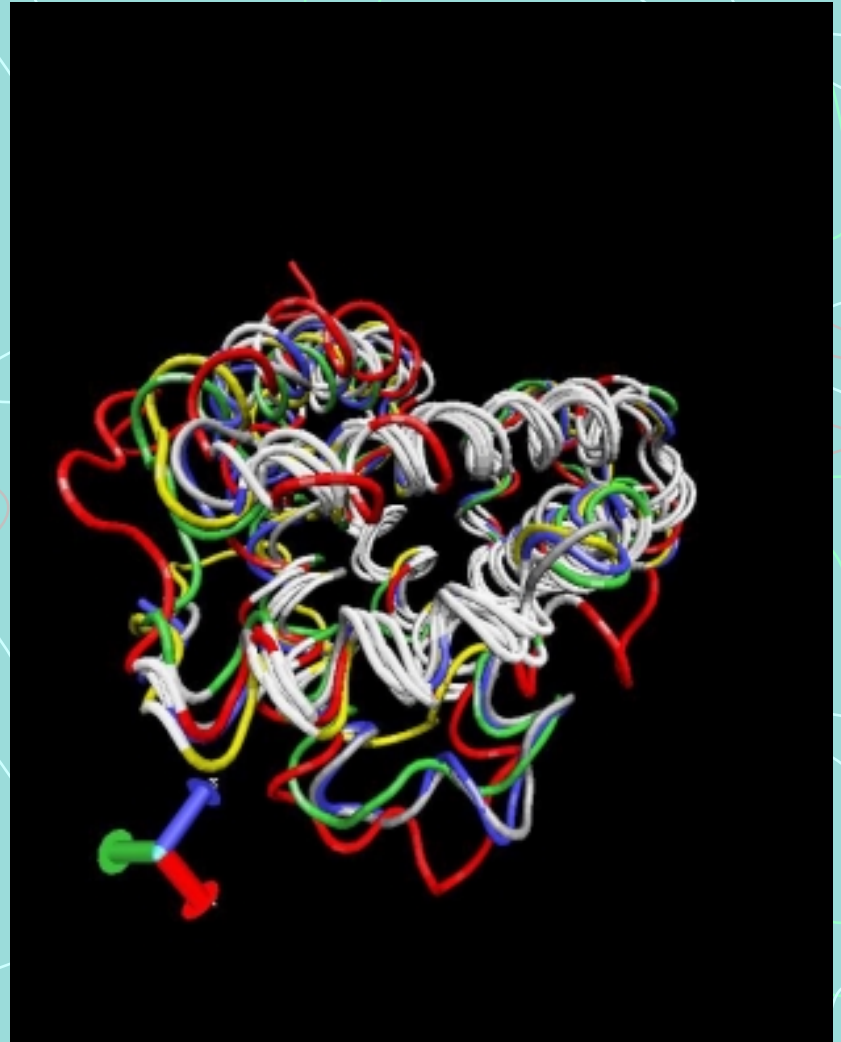
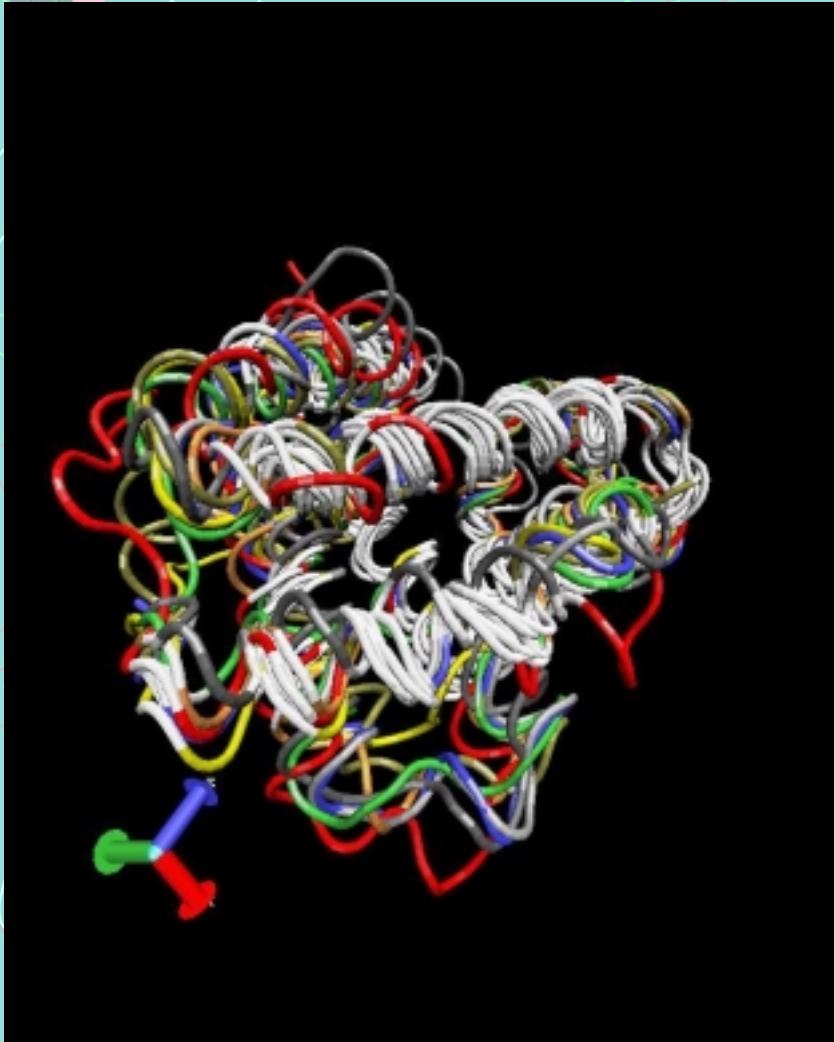
- We also tried the following:
- The triose phosphate isomerase family- taken from the HOMSTRAD (Mizuguchi et al. 1998), where a core of 216 C $\alpha$  atoms was found which is 87% of the smallest molecule
  - The seven globin molecules mentioned in Gerstein & Levitt (1996), where a core 71 C $\alpha$  atoms was found that 52% of the smallest molecule.

# Serpins - Example



Structural Bioinformatics - 2003;  
Imrich J. Wolfson  
Running times: 10-12 seconds

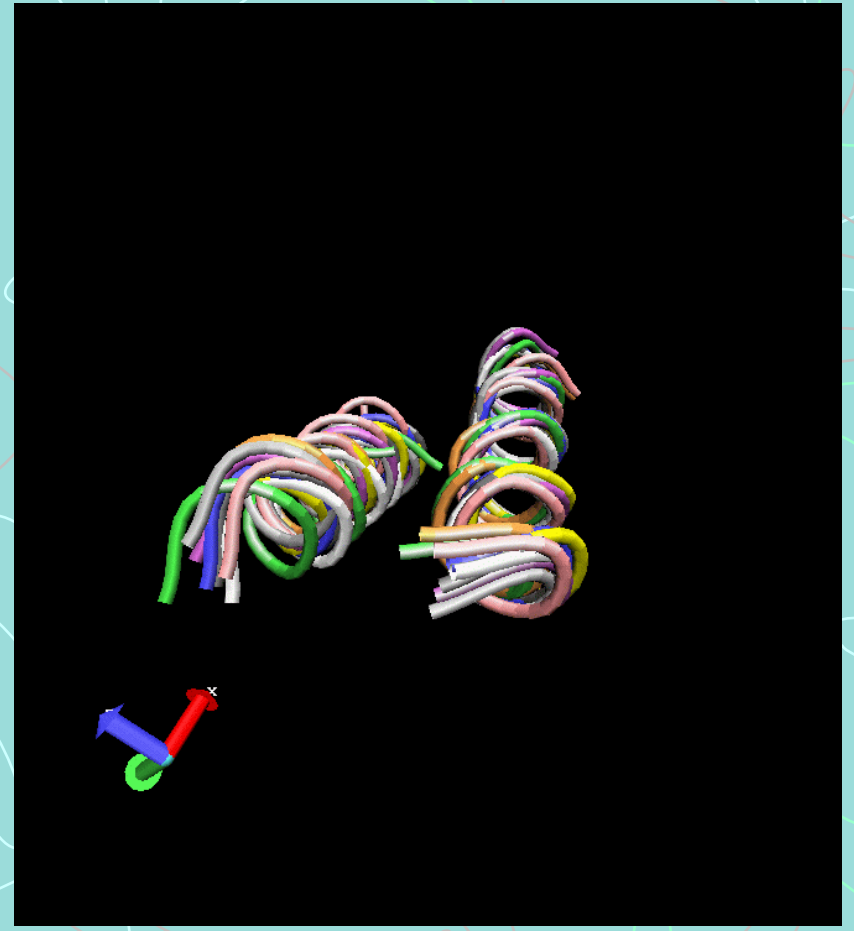
# Globins



Running times: **1 minute** on the average

Structural Bioinformatics - 2003;  
Prof. Hans-J. Wolfson

# The cal-binding results



Running time: 8 seconds

# The TIM BARREL results

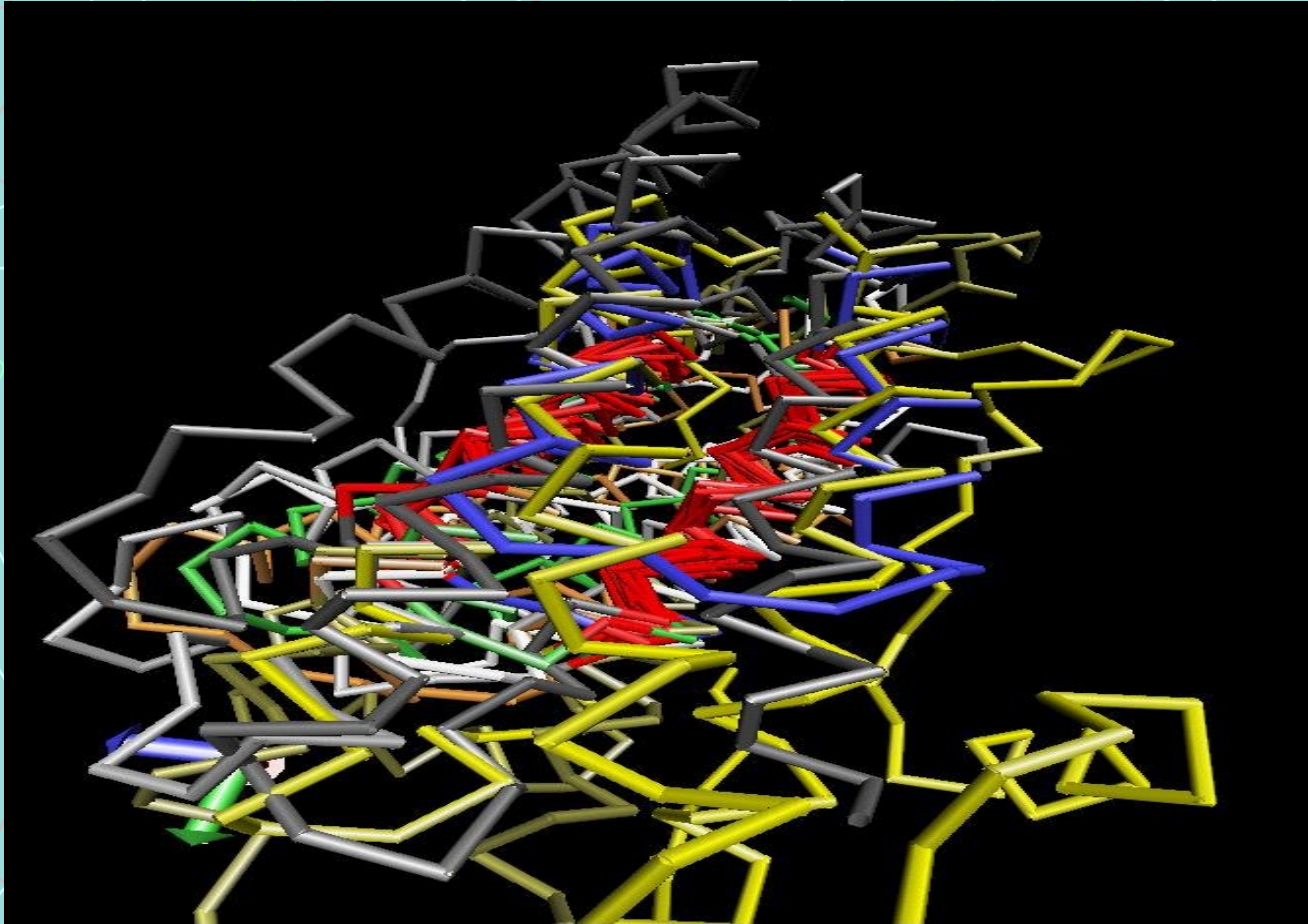


35:27 minutes



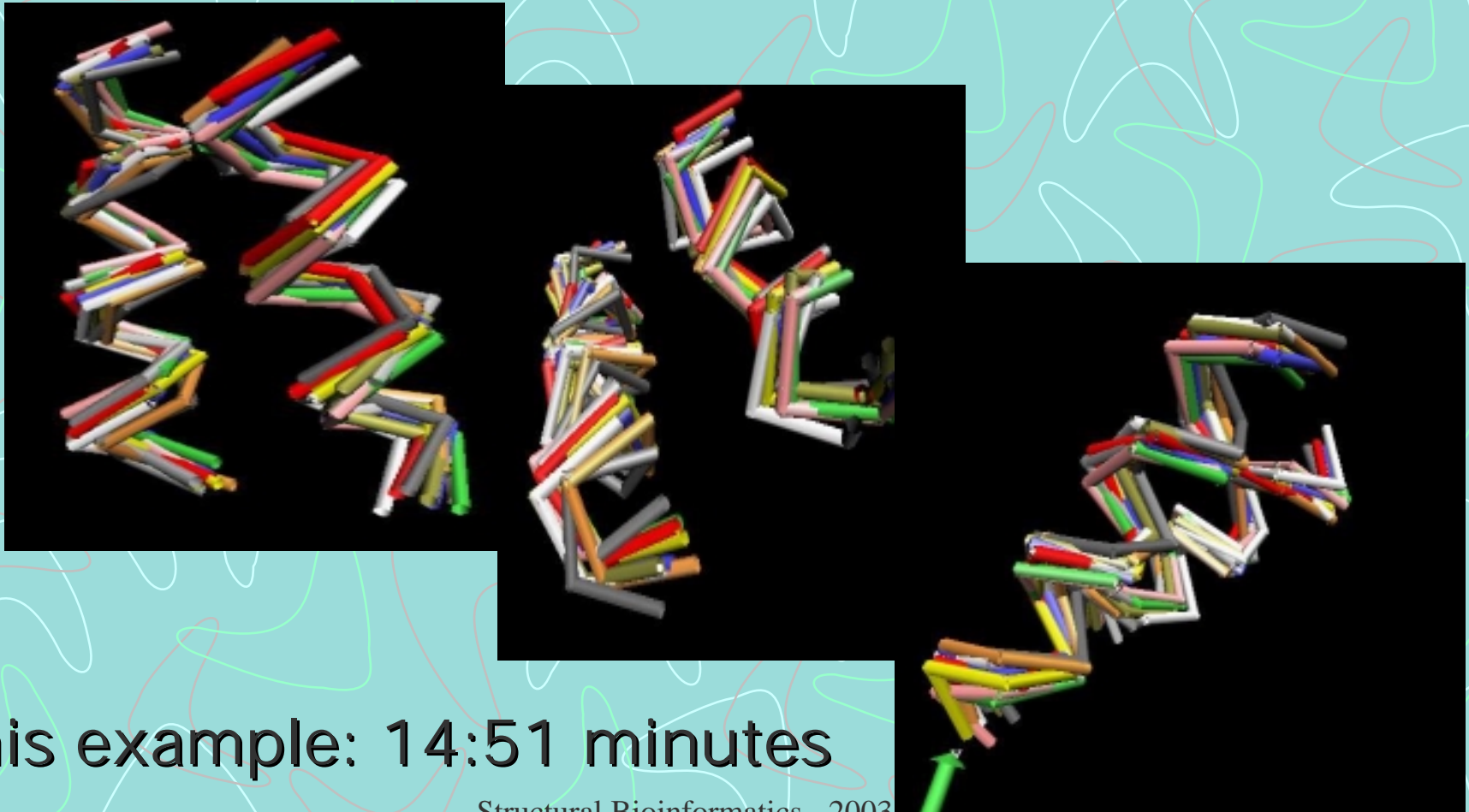
13:33 minutes

# 4 Alpha Helix Bundle



Structural Bioinformatics - 2003;  
Prof. Dr. J. Wolfson  
**Running times: 7-15 minutes**

# Helix Bundle - views of the Core



**This example: 14:51 minutes**



# Publications :

1. **N. Leibowitz, Z.Y. Fligelman, R. Nussinov, H.J. Wolfson**, *Multiple Structural Alignment and Core Detection by Geometric Hashing*, Proc. of the 7'th International Conference on Intelligent Systems in Molecular Biology, Heidelberg, Germany, August 1999, pp. 169--177, (T. Lengauer et al., ed.'s), AAAI Press, Menlo Park, California.
2. **N. Leibowitz, R. Nussinov, H.J. Wolfson**, *MUSTA - a General, Efficient, Automated Method for Multiple Structure Alignment and Detection of Common Motifs: Application to Proteins*, J. of Computational Biology, 8(2), 93--121, (2001).
3. **N. Leibowitz, Z.Y. Fligelman, R. Nussinov, H.J. Wolfson**, *An Automated Multiple Structure Alignment and Detection of a Common Substructural Motif*, PROTEINS: Structure, Function and Genetics, 43, 235--245, (2001).



# Further Developments

- **MultiProt** – runs much faster and allows partial multiple structural alignment, yet sequence order dependent :  
M. Shatsky, R. Nussinov, H.J. Wolfson, *MultiProt - a Multiple Protein Structural Alignment Algorithm*, 2'nd Workshop on Algorithms in Bioinformatics (WABI'02 as part of ALGO'02), Rome, Italy, Sept. 2002, Lecture Notes in Computer Science 2452, pp. 235-250, Springer Verlag.
- **MASS** – Dror-Shem Tov, Nussinov, Wolfson – SSE based multiple structure alignment (to appear in ISMB 2003, Australia)