**BINF 731**

# Protein Structure Analysis

Iosif Vaisman

2009

## Root mean square deviation

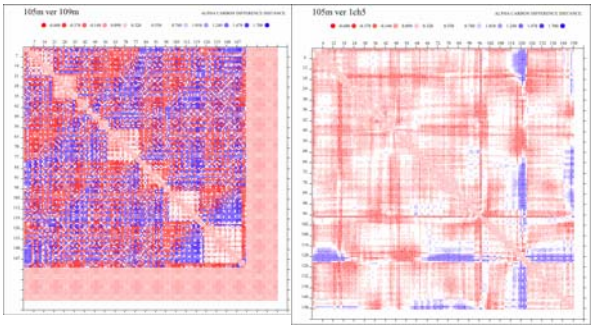
Coordinate based RMSD

$$RMSd(A,B) = \min_T \sqrt{\sum_{i=1}^{n} (A_i - TB_i)^2}$$

Distance based RMSD

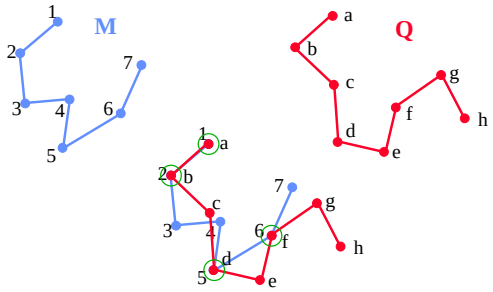
$$RMSd_D(A,B) = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(d_{ij}^A - d_{ij}^B\right)^2}$$

## Difference Distance Matrix Plot (DDMP)



## Geometric Hashing

Find common subsets, invariant under rotation and translation in two point sets M (model) and Q (query).




Finding the maximum coincidence set is an NP-hard problem

## Geometric Hashing
### Reference frames

- Two points (basis pair) define a reference frame
- The coordinates of all points are computed in the reference frame (reference frame system)
- There will be pairs of points (from M and Q) with the same coordinates
- The number of such pairs depends on selection of reference frame and reference frame system resolution

## Geometric Hashing Algorithm

**Preprocessing**

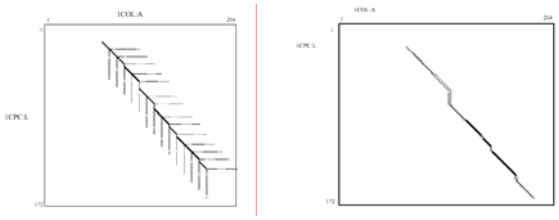
Hash table H is created. It has a bin for each cell in the frame systems. The coordinates of all points in each model frame system are calculated. If there is a point in the cell (p,q) in the frame system with basis $(a_i,a_k)$, then $(a_i,a_k)$ is placed in the bin H(p,q)

**Recognition**

A pair of points in the query is chosen as basis, and the coordinates of the other points are calculated. These coordinates are used as indices for H, and for each cell being indexed, a vote is given for the (model) basis pairs in the cell. The number of votes for a model basis pair is the number of coinciding points to the query (using the specified query basis pair)

## Combinatorial Extension (CE) Algorithm

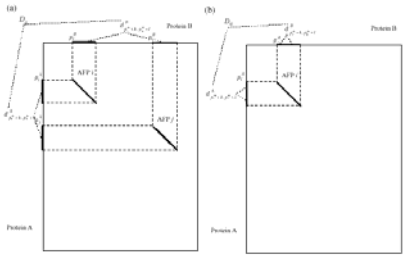
Structure alignment of phycocyanin (1CPC:L) to colicin A (1COL:A).



The solid line represents the optimal path built from AFPs. The dotted line represents the search area at every step of path extension.

The thick solid line represents alignment overlap both before and after optimization.

## Combinatorial Extension (CE) Algorithm



Calculation of distance

*Dij* for two AFPs *i* and *j* from the path

*Dii* for single AFP *i* from the path.

## Search for common substructures by clique detection