

# Protein Structure Analysis

BME 110: Computational Biology Tools

# Topics

- Why do we care about Protein structure?
- What do we really want from an alignment?
- Structure Alignments - Tools and Databases
- How those tools work
- A case study in distant homology - PAZ domains
  - Pf\_Ago (Pyrococcus furiosus) and H\_Ago (Homo sapiens)
  - What the Author thinks
  - What DALI and VAST suggest
  - What do you think?

# Why Examine Protein Structures?

- Structure more conserved than sequence
  - Similar folds often share similar function
  - Remote similarities may only be detectable at structure level
- Interpreting experimental data
  - Locating sites of interesting mutations
  - Locating splice sites
- Designing experiments
  - *In silico* mutagenesis

# Structure Analysis

- Identify interesting sites on protein
- Measure distances, angles, etc.
- Examine surface properties (shape, charge)
- Compare two structures
  - Homologs
  - Mutants
  - With and Without Ligands

# Comparing Protein Structures

- Defined alignment
  - Mutant-wildtype, model-native, two different conformations.
  - Unique solution exists -- we know the true alignment
- Derived alignment
  - Query is an unknown protein
  - Known Parent (assumed homolog)
  - calculate a computationally 'Optimal' alignment
  - infer annotation from parent to query

# What do we want from an Alignment?

- 'Optimal alignment'
  - Important parts of protein should associate (align) with each other
    - Catalytic residues and their position in 3-space
    - Important structures (hinges, binding sites)
    - Protein interface residues and their position in 3-space
    - Evolutionary History
  - Natural selection only selects for successful Function
  - Sequences (and alignments) are assumed to be sequential
- Sequence alignments can be improved when we have structural information
  - No unique solution though (more residues or closer match?)
  - Structural alignment implies a sequence alignment

# Tools and Databases

- Structure Databases and search tools
  - NCBI Structure (VAST and MMDB)
    - <http://www.ncbi.nlm.nih.gov/Structure/>
    - Molecular Modeling DataBase
      - Experimentally derived structures from PDB (not theoretical)
  - FSSP (DALI)
    - <http://www.ebi.ac.uk/dali/>
    - <http://ekhidna.biocenter.helsinki.fi/dali/start>
    - Families of Structurally Similar Proteins
      - Maintains database of Protein Neighbors organized by PDB code
  - CE
    - <http://cl.sdsc.edu/>
    - Combinatorial Extension
      - Maintains database of Protein Neighbors by PDB code

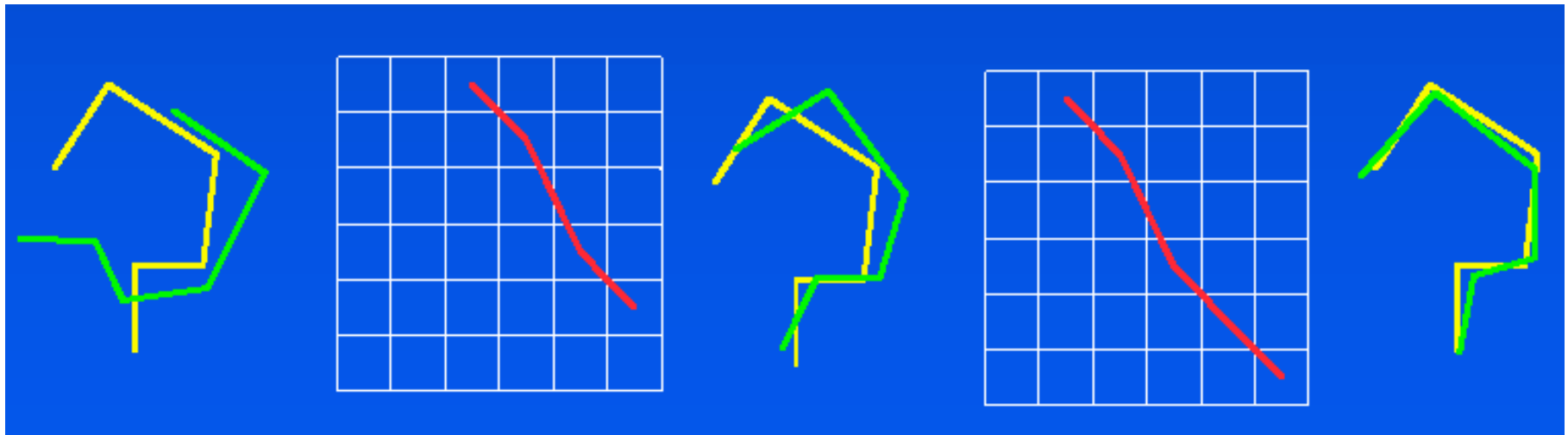
# Tools and Databases(2)

- Structure classification by domain
  - Classifications based on Secondary structure
  - SCOP Structural Classification of Proteins
    - <http://scop.berkeley.edu/>, Alessi Mursin et al.
    - Last release November 2007
  - CATH Class Architecture Topology Homology
    - <http://www.cathdb.info/>, Automated and manual classification
    - Last release Jan 2007, v. 3.1.0
- CEMC – Multiple Structure Alignment
  - <http://bioinformatics.albany.edu/~cemc/>

# How Structure alignments work

- Methods
  - Structal
  - DALI
  - VAST
- Structure similarity measures
  - RMSD
  - Pvalues

# Iterative Dynamic Programming

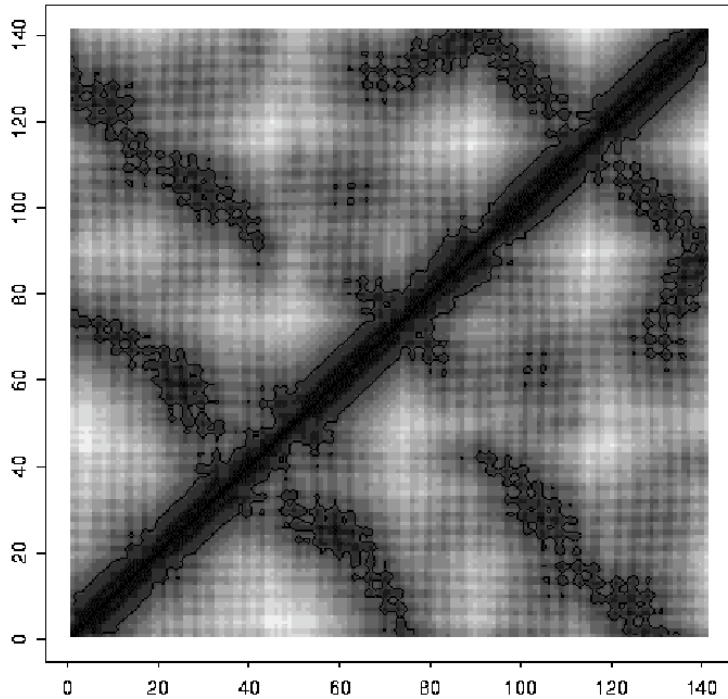


- Algorithm:
  1. Make an initial guess for the superposition
  2. Calculate all pairwise CA-CA distances and generate a scoring matrix.
  3. Find optimal alignment according to this scoring matrix by dynamic programming.
  4. Re-superimpose structures using this alignment
  5. Repeat step 2-4 until convergence.
- No guarantee of optimal solution, final result depends on the initial alignment selected.
- Structural: Subbiah et al, 1993 Curr. Biol 3:141)

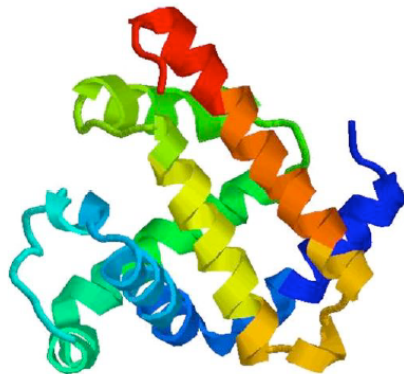
# Structural Alignment

- Many methods other than dynamic programming are used.
- Most methods use some sort of heuristics to speed things up and make good initial guesses:
  - Sheba Sequence alignment
  - Mammoth Local structure alignment
  - VAST aligns secondary structure element vectors
  - DALI Distance matrix alignment

# Distance Matrix Alignment

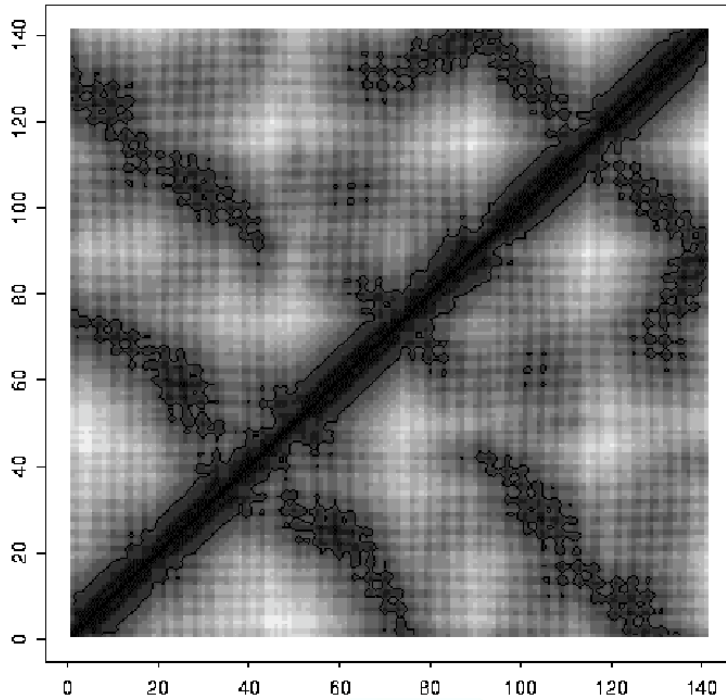


Myoglobin

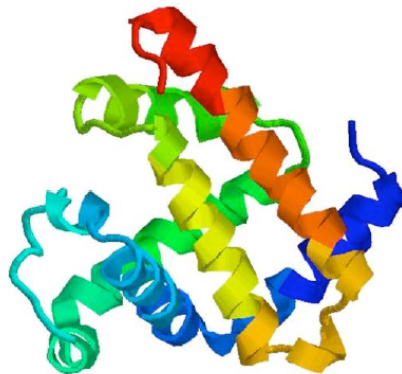


- Matrix of all pair-wise distances
- Characteristic patterns:
  - Main diagonal runs correspond to helix (i.e local contacts)
  - Hairpins - start on main diagonal, run perpendicular
  - Parallel pairs run parallel to main diagonal
  - Others are long range contacts.
- Converts 3D alignment problem to a 2D problem.
  - Find best subset of rows and columns such that the distance matrices of two proteins are optimally similar

# Contact Map Comparison

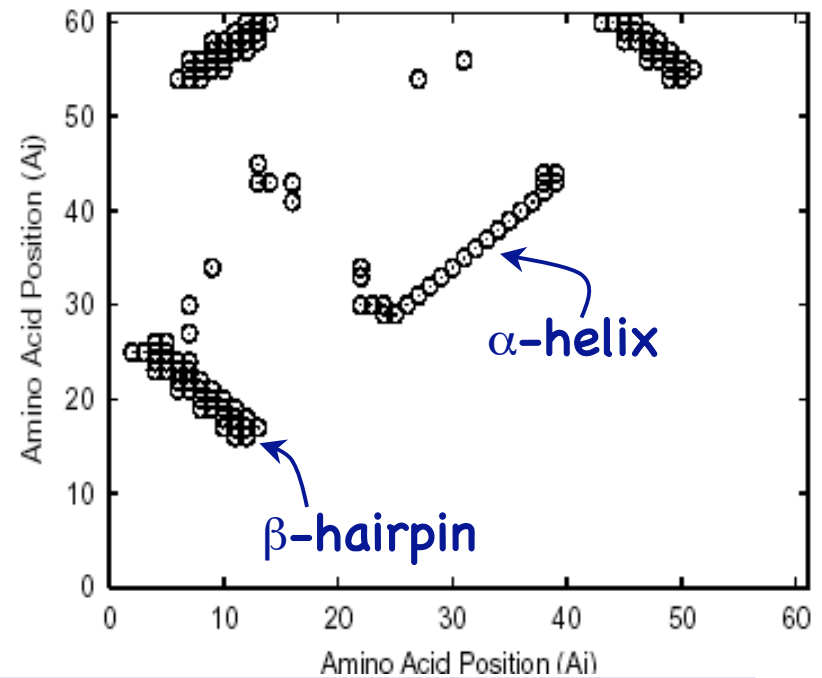


Myoglobin



Protein G

//-strands

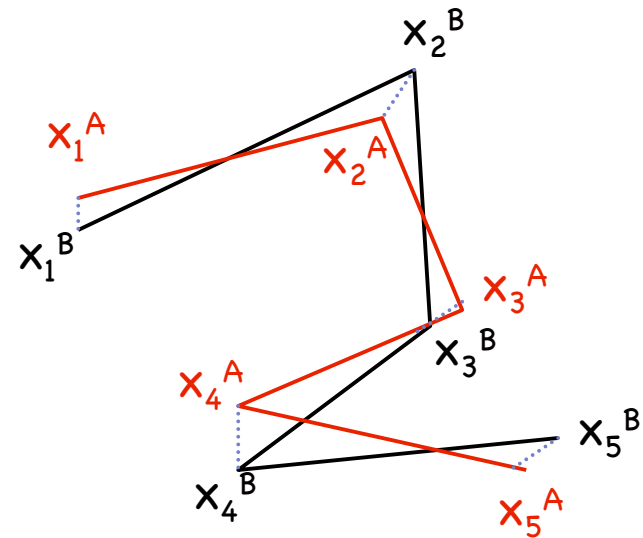


# Similarity Measures: RMSD

- RMSD = root mean square deviation

$$\sqrt{\langle \|x_i^A - x_i^B\|^2 \rangle}$$

1. Superimpose optimally
2. Pair up residues
3. Calculate RMSD

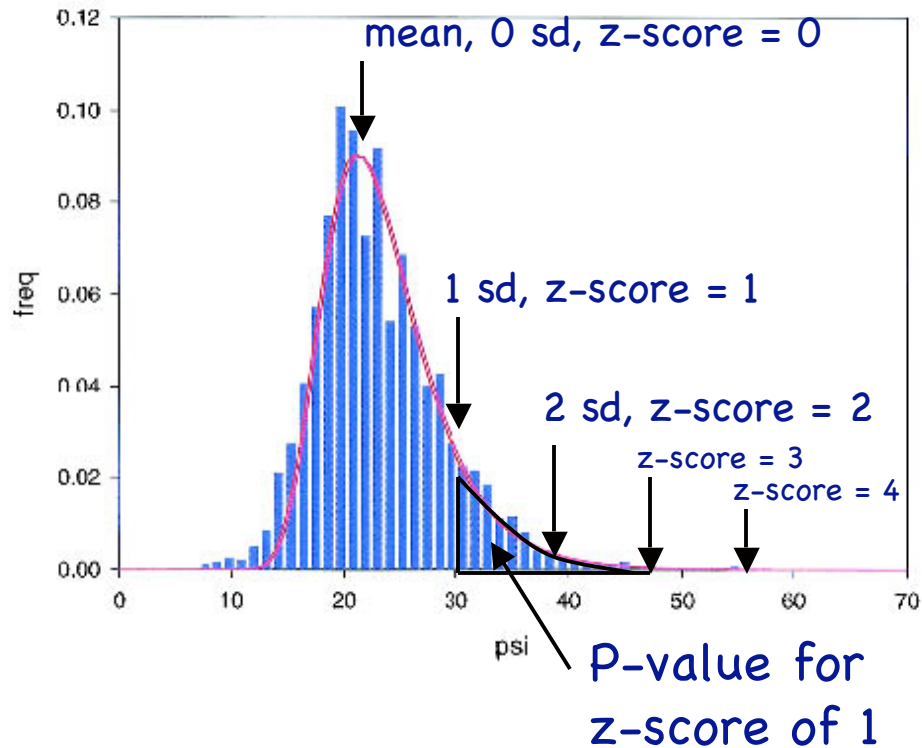


Sensitive to outliers

Depends on number of pairs compared

A better measure is the significance of this RMSD for similar sized matches

# Z-scores & P-values



Histogram of scores for random matches

- Z-score: # of standard deviations above the mean:
  - $\pm 1$  sd  $\sim 66\%$
  - $\pm 2$  sd  $\sim 95\%$
  - If we have a histogram, we can just count; Or integrate a function fitted to the histogram.
- P-value
  - Probability of obtaining  $\geq$  this score under the null model (normally distributed data -- "by chance")



# Case Study -- PAZ domain of Pf\_Ago

- Ji-Joon Song [Song 2004, Science, PMID: 15284453](#) asserts on p. 1435 of the above paper that the following are functionally equivalent:
- What do you think?
- 1si2:224 is 1st residue

Pf_Ago 1u04	hAgo1 1si2/1si3
Y212	Y309(Y90)
Y216	Y314(Y95)
H217	H269(H49)
Y190	Y277(Y57)

# In class example - the PAZ domain

- 1u04 - PAZ domain, chain A 152-275
- 1si2 - PAZ domain, chain A 4-128
- <http://www.ncbi.nlm.nih.gov/structure/>,  
1U04 pdb code, PAZ domain, show structure
- Dalilite at: <http://www.ebi.ac.uk/DaliLite/>
  - Align 1U04 with 1SI2:A (h\_Ago).
  - What is the Z-score? Is it significant?
  - What is the RMSD? Is this a reasonable alignment
  - How many residues aligned?

# In class exercise – Pymol

- Load both molecule 1 and molecule 2 (aligned) into Pymol
  - Action-preset-pretty for both molecules
  - For 1u04, delete everything you don't need
    - select-rename object 1u04 (if needed)
    - chain B; select-remove atoms
    - chain A and resi 1-151; select-remove atoms
    - chain A and resi 276-770; select-remove atoms
    - color red
  - Load 1si2, color it yellow
    - chain B is a small RNA; show spheres, chain B; color blue
  - select 1u04 and resi 212; show as sticks
    - repeat for 216, 217, 190
  - select 1si2 and resi 309; show as sticks
    - repeat for 314, 269, 277
  - shift click on 1u04 chain

# In class exercise summary

- Structural alignments can be very useful
- Helical structures can be troublesome
  
- How should alignments handle out-of-order substitutions?
  
- Is J.J. Song correct? Is Dali? Vast??