**Protein threading**

Structure is better conserved than sequence

Structure can adopt a wide range of mutations.

Physical forces favor certain structures.

Number of folds is limited.

- Currently ~700
- Total: 1,000 ~10,000

**TIM barrel**

**Concept of Threading**

- Thread (align or place) a query protein sequence onto a template structure in "optimal" way
- Good alignment gives approximate backbone structure

- **Query sequence**
  
  MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

- **Template set**

**Threading problem**

- Threading: Given a sequence, and a fold (template), compute the optimal alignment score between the sequence and the fold.
- If we can solve the above problem, then
  - Given a sequence, we can try each known fold, and find the best fold that fits this sequence.
  - Because there are only a few thousands folds, we can find the correct fold for the given sequence.
- Threading is NP-hard.

**Components of Threading**

- Template library
  - Use structures from DB classification categories (PDB)
- Scoring function
  - Single and pairwise energy terms
- Alignment
  - Consideration of pairwise terms leads to NP-hardness
  - heuristics
- Confidence assessment
  - Z-score, P-value similar to sequence alignment statistics
- Improvements
  - Local threading, multi-structure threading

**Protein Threading**

- Basic premise
  
  The number of unique structural (domain) folds in nature is fairly small (possibly a few thousand)

- Statistics from Protein Data Bank (~35,000 structures)

  90% of new structures submitted to PDB in the past three years have similar structural folds in PDB
Protein Threading – energy function

MTYKLNLGKTQGETTEAVDAATAEKVFQYANDNGVDGEWTYTE

- how preferable to put two particular residues nearby: $E_p$
- how well a residue fits a structural environment: $E_s$
- alignment gap penalty: $E_g$

Total energy: $E_p + E_s + E_g$

Find a sequence-structure alignment to minimize the energy function

Assessing Prediction Reliability

MTYKLNLGKTQGETTEAVDAATAEKVFQYANDNGVDGEWTYTE

Score = -1500  Score = -900  Score = -1120  Score = -720

Which one is the correct structural fold for the target sequence if any?

The one with the highest score?

Prediction of Protein Structures

- Examples – a few good examples

Prediction of Protein Structures

- Not so good example

Existing Prediction Programs

- PROSPECT
  - https://csbl.bmb.uga.edu/protein_pipeline

- FUGU
  - http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html

- THREADER
  - http://bioinf.cs.ucl.ac.uk/threader/
CASP/CAFASP

- CASP: Critical Assessment of Structure Prediction
- CAFASP: Critical Assessment of Fully Automated Structure Prediction

1. Won't get tired
2. High-throughput

CASP6/CAFASP4

- 64 targets
- Resources for predictors
- No X-ray, NMR machines (of course)
- CAFASP4 predictors: no manual intervention
- CASP6 predictors: anything (servers, google,...)
- Evaluation:
  - CASP6 assessed by experts+computer
  - CAFASP4 evaluated by a computer program.
  - Predicted structures are superimposed on the experimental structures.
  - CASP7 will be held this year (November)

Protein structure databases

- PDB
  - 3D structures
- SCOP
  - Murzin, Brenner, Hubbard, Chothia
  - Classification
    - Class (mostly alpha, mostly beta, alpha/beta (interspersed), alpha+beta (segregated), multi-domain, membrane)
    - Fold (similar structure)
    - Superfamily (homology, distant sequence similarity)
    - Family (homology and close sequence similarity)

The SCOP Database

Structural Classification Of Proteins

FAMILY: proteins that are >30% similar, or >15% similar and have similar known structure/function

SUPERFAMILY: proteins whose families have some sequence and function/structure similarity suggesting a common evolutionary origin

COMMON FOLD: superfamilies that have same secondary structures in same arrangement, probably resulting by physics and chemistry

CLASS: alpha, beta, alpha-beta, alpha+beta, multidomain

Protein databases

- CATH
  - Orengo et al
  - Class (alpha, beta, alpha/beta, few SSEs)
  - Architecture (orientation of SSEs but ignoring connectivity)
  - Topology (orientation and connectivity, based on SSAP = fold of SCOP)
  - Homology (sequence similarity = superfamily of SCOP)
    - S level (high sequence similarity = family of SCOP)
  - SSAP alignment tool (dynamic programming)
Protein databases

- FSSP
  - DALI structure alignment tool (distance matrix)
    - Holm and Sander
- MMDB
  - VAST structure comparison (hierarchical)
    - Madej, Bryant et al

Protein structure comparison

- Levels of structure description
  - Atom/atom group
  - Residue
  - Fragment
  - Secondary structure element (SSE)
- Basis of comparison
  - Geometry/architecture of coordinates/relative positions
  - Sequential order of residues along backbone, ...
  - Physio-chemical properties of residues, ...

How to compare?

- **Key problem**: find an optimal correspondence between the arrangements of atoms in two molecular structures (say A and B) in order to align them in 3D
- Optimality of the alignment is determined using a root mean square measure of the distances between corresponding atoms in the two molecules
- **Complication**: It is not known a priori which atom in molecule B corresponds to a given atom in molecule A (the two molecules may not even have the same number of atoms)

Structure Analysis – Basic Issues

- Coordinates for representing 3D structures
  - Cartesian
  - Other (e.g., dihedral angles)
- Basic operations
  - Translation in 3D space
  - Rotation in 3D space
  - Comparing 3D structures
    - Root mean square distances between points of two molecules are typically used as a measure of how well they are aligned
    - Efficient ways to compute minimal RMSD once correspondences are known (O(n) algorithm)
    - Using eigenvalue analysis of correlation matrix of points
- Due to the high computational complexity, practical algorithms rely on heuristics

Structure Analysis – Basic Issues

- Sequence order dependent approaches
  - Computationally this is easier
  - Interest in motifs preserving sequence order
- Sequence order independent approaches
  - More general
  - Active sites may involve non-local AAs
  - Searching with structural information

Find the optimal alignment
Optimal Alignment

- Find the highest number of atoms aligned with the lowest RMSD (Root Mean Squared Deviation)
- Find a balance between local regions with very good alignments and overall alignment

Structure Comparison

Which atom in structure A corresponds to which atom in structure B?

THESE SENTENCES ALIGN NICELY

THE - SEQUENCE ALIGNED NICELY

Structural Alignment

An optimal superposition of myoglobin and beta-hemoglobin, which are structural neighbors. However, their sequence homology is only 8.5%

Structure Comparison

Methods to superimpose structures by translation and rotation

\[
\begin{align*}
&x_1, y_1, z_1 \\
x_2, y_2, z_2 \\
x_3, y_3, z_3
\end{align*}
\]

Translation

\[
\begin{align*}
&x_1 + d, y_1, z_1 \\
x_2 + d, y_2, z_2 \\
x_3 + d, y_3, z_3
\end{align*}
\]

Rotation

Structure Comparison

Scoring system to find optimal alignment

Answer: Root Mean Square Deviation (RMSD)

\[
RMSD = \sqrt{\frac{\sum d_i^2}{n}}
\]

- \( n \) = number of atoms
- \( d_i \) = distance between 2 corresponding atoms \( i \) in 2 structures

Root Mean Square Deviation

\[
RMSD = \frac{\sum (x_{\text{min}} - x_{\text{max}})^2}{5} - \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5}
\]
RMSD

Unit of RMSD => e.g. Ångstroms
- identical structures => RMSD = “0”
- similar structures => RMSD is small (1 – 3 Å)
- distant structures => RMSD > 3 Å

Pitfalls of RMSD

- all atoms are treated equally
  (e.g. residues on the surface have a higher degree of freedom than those in the core)
- best alignment does not always mean minimal RMSD
- significance of RMSD is size dependent

Alternative RMSDs

- aRMSD = best root-mean-square distance calculated over all aligned alpha-carbon atoms
- bRMSD = the RMSD over the highest scoring residue pairs
- wRMSD = weighted RMSD


Structural Alignment Methods

- Distance based methods
  - DALI (Holm and Sander, 1993): Aligning 2-dimensional distance matrices
  - SSAP (Orengo and Taylor, 1995): Double dynamic programming using intra-molecular distance
  - CE (Shendylov and Bourne, 1998): Combinatorial Extension of best matching regions

- Vector based methods
  - VAST (Madej et al., 1995): Graph theory based SSE alignment;
  - 3dSearch (Singh and Brutlag, 1997) and 3D Lookup (Holm and Sander, 1995): Fast SSE index look-up by geometric hashing.
  - TOP (Lu, 2000): SSE vector superpositioning

- Both vector and distance based
  - LOCK (Singh and Brutlag, 1997): Hierarchically uses both secondary structures vectors and atomic distances.

Basic DP (STRUCTAL)

1. Start with arbitrary alignment of the points in two molecules A and B
2. Superimpose in order to minimize RMSD.
3. Compute a structural alignment (SA) matrix where entry \( (i,j) \) is the score for the structural similarity between the \( i \)th point of A and the \( j \)th point of B
4. Use DP to compute the next alignment.
   - Gap cost = 0
5. Iterate steps 2–4 until the overall score converges
6. Repeat with a number of initial alignments

STRUCTAL

- Given
  2 Structures (A & B),
  2 Basic Comparison Operations
1. Given an alignment optimally
   SUPERIMPOSE A onto B
2. Find an Alignment between A and B based on their 3D coordinates
   \[ S_{ij} = M[1+(d/d_{0})^2] \]
   M and \( d_{0} \) are constants