Bioinformatics for Biologists

Comparative Protein Analysis:
Part III. Protein Structure Prediction
and Comparison

Robert Latek, PhD
Sr. Bioinformatics Scientist
Whitehead Institute for Biomedical Research

Protein Structure

Why is protein structure information useful?
Predicting Important AAs

Surface Mapping
Protein Interfaces

Property Comparisons
Syllabus

- **Protein Structure Classification**
- Structure Coordinate Files & Databases
- Comparing Protein Structures
  - Aligning 3D Structures
- Predicting Protein Structure
  - Specialized Structural Regions
  - Secondary Structure Prediction
  - Tertiary Structure Prediction
    - Threading
    - Modeling
- Structure Visualization

Structure Classification

- Proteins can adopt only a limited number of possible 3D conformations
  - Combinations of helices, sheets, loops, and coils
- Completely different sequences can fold into similar shapes
- Protein Structure Classes
  - Class [a]: bundles of helices
  - Class [β]: anti-parallel sheets (sandwiches and barrels)
  - Class [α]/[β]: parallel sheets with intervening helices
  - Class [α] + [β]: segregated helices and anti-parallel sheets
  - Multi-domain
  - Membrane/Cell surface proteins

*http://info.bio.cmu.edu/courses/03231/ProtStruc/ProtStruc2.htm*
Structure Families

- Divide structures into the limited number of possible structure families
  - Homologous proteins can be identified by examining their respective structures for conserved fold patterns
  - Representative members can be used for modeling sequences of unknown structure

Structure Family Databases

- SCOP: Structural Classification Of Proteins
  - based on a definition of structural similarities. Hierarchical levels to reflect evolutionary and structural relationships
  - http://scop.mrc-lmb.cam.ac.uk/scop

- CATH: Classification by Class, Architecture, Topology, and Homology
  - classified first into hierarchical levels like SCOP
  - http://www.biochem.ac.uk/bsm/cath/

- FSSP: Fold classification based on Structure-structure alignment of proteins
  - based on structural alignment of all pair-wise combinations of proteins in PDB by DALI (used to id common folds and place into groups)
  - http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html

- MMDB
  - Aligns 3D structures based on similar arrangements of secondary structural elements (VAST)

- SARF
  - categorized on the basis of structural similarity, categories are similar to other dbs
  - http://123d.ncifcrf.gov/
Syllabus

- Protein Structure Classification
- **Structure Coordinate Files & Databases**
- Comparing Protein Structures
  - Aligning 3D Structures
- Predicting Protein Structure
  - Specialized Structural Regions
  - Secondary Structure Prediction
  - Tertiary Structure Prediction
    - Threading
    - Modeling
- Structure Visualization

Coordinates

- Coordinate Data: location of a molecule’s atoms in space (XYZ triple)
- XYZ triple is labeled with an atom, residue, chain
  - Modified aa are labeled with X, H’s not usually listed

<table>
<thead>
<tr>
<th>Atom</th>
<th>Residue</th>
<th>Chain</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>ALA</td>
<td>C</td>
<td>35.4</td>
<td>9.3</td>
<td>102.5</td>
</tr>
</tbody>
</table>

- Data Representation
  - Chemistry Rules Approach: connect the dots utilizing a standard rules base to specify bond distances (not consistent among applications)
  - Explicit Bonding Approach: explicit bonding information is specified in the file (very consistent)
Coordinate File Formats

- **MMDB** “Molecular Modeling DataBank” Format
  - ASN.1 standard data description language (explicit bond information)
- **mmCIF** “Chemical Interchange Format”
  - (relational db format)
- **PDB** “Protein DataBank” Format
  - Column oriented, “flexible format” (chemistry rules)

```
    ATOM   1432  N   ALA A 259      15.711  12.486  46.370  1.00 28.54
    ATOM   1433  CA  ALA A 259      17.047  12.953  46.726  1.00 27.48
    ATOM   1434  C   ALA A 259      17.029  14.459  46.979  1.00 25.31
    ATOM   1435  O   ALA A 259      17.787  15.207  46.367  1.00 25.19
    ATOM   1436  CB  ALA A 259      18.035  12.617  45.610  1.00 25.32
    ATOM   1437  N   TRP A 260      16.149  14.897  47.875  1.00 23.61
    ATOM   1438  CA  TRP A 260      16.033  16.312  48.210  1.00 21.03
    ATOM   1439  C   TRP A 260      17.121  16.700  49.211  1.00 20.94
    ATOM   1440  O   TRP A 260      17.917  17.601  48.957  1.00 19.84
```

**Example PDB File**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Residue</th>
<th>Atom</th>
<th>Atom Type</th>
<th>Chain</th>
<th>Residue</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Structure scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>1432</td>
<td>N</td>
<td>ALA</td>
<td>A</td>
<td>259</td>
<td>15.711</td>
<td>12.486</td>
<td>46.370</td>
<td>1.00 28.54</td>
</tr>
<tr>
<td>ATOM</td>
<td>1433</td>
<td>CA</td>
<td>ALA</td>
<td>A</td>
<td>259</td>
<td>17.047</td>
<td>12.953</td>
<td>46.726</td>
<td>1.00 27.48</td>
</tr>
<tr>
<td>ATOM</td>
<td>1434</td>
<td>C</td>
<td>ALA</td>
<td>A</td>
<td>259</td>
<td>17.029</td>
<td>14.459</td>
<td>46.979</td>
<td>1.00 25.31</td>
</tr>
<tr>
<td>ATOM</td>
<td>1435</td>
<td>O</td>
<td>ALA</td>
<td>A</td>
<td>259</td>
<td>17.787</td>
<td>15.207</td>
<td>46.367</td>
<td>1.00 25.19</td>
</tr>
<tr>
<td>ATOM</td>
<td>1436</td>
<td>CB</td>
<td>ALA</td>
<td>A</td>
<td>259</td>
<td>18.035</td>
<td>12.617</td>
<td>45.610</td>
<td>1.00 25.32</td>
</tr>
<tr>
<td>ATOM</td>
<td>1437</td>
<td>N</td>
<td>TRP</td>
<td>A</td>
<td>260</td>
<td>16.149</td>
<td>14.897</td>
<td>47.875</td>
<td>1.00 23.61</td>
</tr>
<tr>
<td>ATOM</td>
<td>1438</td>
<td>CA</td>
<td>TRP</td>
<td>A</td>
<td>260</td>
<td>16.033</td>
<td>16.312</td>
<td>48.210</td>
<td>1.00 21.03</td>
</tr>
<tr>
<td>ATOM</td>
<td>1439</td>
<td>C</td>
<td>TRP</td>
<td>A</td>
<td>260</td>
<td>17.121</td>
<td>16.700</td>
<td>49.211</td>
<td>1.00 20.94</td>
</tr>
<tr>
<td>ATOM</td>
<td>1440</td>
<td>O</td>
<td>TRP</td>
<td>A</td>
<td>260</td>
<td>17.917</td>
<td>17.601</td>
<td>48.957</td>
<td>1.00 19.84</td>
</tr>
</tbody>
</table>

Coordinate Databases

- **RCSB** (Research Collaboratory for Structural Bioinformatics) [http://www.rcsb.org/](http://www.rcsb.org/)
  - Formally know as the Protein Data Bank at Brookhaven National Laboratories
  - Structure Explorer PDB search engine
    - Text and PDB ID (4 letter code) searching

- **MMDB** (Molecular Modeling Database @NCBI)
  - Compilation of structures represented in multiple formats
  - Provides structure summaries
  - BLAST sequences to search for available structures
Syllabus

- Protein Structure Classification
- Structure Coordinate Files & Databases
- Comparing Protein Structures
  - Aligning 3D Structures
- Predicting Protein Structure
  - Specialized Structural Regions
  - Secondary Structure Prediction
  - Tertiary Structure Prediction
    - Threading
    - Modeling
- Structure Visualization

Sequence & Structure Homology

- Sequence
  - Identify relationships between sets of linear protein sequences
- Structure
  - Categorize related structures based on 3D shapes
    - Structure families do not necessarily share sequence homology
Structure Comparison

• Compare Structures that are:
  – **Identical**
    • Similarity/difference of independent structures, x-ray vs. nmr, apo vs. holo forms, wildtype vs. mutant
  – **Similar**
    • Predict function, evolutionary history, important domains
  – **Unrelated**
    • Identify commonalities between proteins with no apparent common overall structure - focus on active sites, ligand binding sites

• Superimpose Structures by 3D Alignment for Comparison

Structural Alignment

• Structure alignment forms relationships in **3D space**
  – similarity can be redundant for multiple sequences

• **Considerations**
  – Which atoms/regions between two structure will be compared
  – Will the structures be compared as rigid or flexible bodies
  – Compare all atoms including side chains or just the backbone/Cα
  – Try to maximize the number of atoms to align or focus on one localized region (biggest differences usually in solvent-exposed loop structures)
  – How does the resolution of each structure affect comparison
Translation and Rotation

- **Alignment**
  - Translate center of mass to a common origin
  - Rotate to find a suitable superposition

- **Algorithms**
  - Identify equivalent pairs (3) of atoms between structures to seed alignment
    - Iterate translation/rotation to maximize the number of matched atom pairs
  - Examine all possible combinations of alignments and identify the optimal solution

Alignment Methods

- Initially examine secondary structural elements and Cα-Cβ distances to identify folds and the ability to align
- Gap penalties for structures that have discontinuous regions that do not align (alignment-gap-alignment)
  - Anticipate that two different regions may align separately, but not in the same alignment
- Proceed with alignment method:
  - Fast, Secondary Structure-Based
  - Dynamic Programming
  - Distance Matrix
Fast Alignment by SS

- Secondary structure elements can be represented by a vector starting at the beginning of the element
  - Position & length
- Compare the arrangement of clustered vectors between two structures to identify common folds
- Sometimes supplement vectors with information about the arrangement of the side chains (burial/exposure)
- Significance of alignment
  - Likelihood that a cluster of secondary structural elements would be expected between unrelated structures

VAST and SARF

- Implement automatic methods to assign secondary structure

- VAST
- SARF
  http://123d.ncifcrf.gov/
Exhaustive Alignment

• Dynamic Programming
  – Local environment defined in terms of Interatomic distances, bond angles, side chain identity, side chain burial/exposure
  – Align structures by matching local environments - for example, draw vectors representing each Cα-Cβ bond, superimpose vectors

• Distance Matrix
  – Graphic procedure similar to a dot matrix alignment of two sequences to identify atoms that lie most closely together in a 3D structure (based on Cα distances)
  – Similar structures have super-imposable graphs

DALI Distance Alignment

• DALI - http://www2.embl-ebi.ac.uk/dali/
• Aligns two structures
• Determines if a new structure is similar to one already in database (for classification)
Alignment Quality

- Calculate deviation between two aligned structures
- **RMSD** (Root Mean Square Deviation)
  - Goodness of fit between two sets of coordinates
  - Best if < 3 Å
  - Calculate Cα-Cα distances, sum square of distances, divide by the number of pairs, square root

\[
\text{RMSD} = \sqrt{\frac{\sum D_i^2}{N}}
\]

Syllabus

- Protein Structure Classification
- Structure Coordinate Files & Databases
- Comparing Protein Structures
  - Aligning 3D Structures
- **Predicting Protein Structure**
  - Specialized Structural Regions
  - Secondary Structure Prediction
  - Tertiary Structure Prediction
    - Threading
    - Modeling
- Structure Visualization
Predicting Specialized Structures

• **Leucine Zippers**
  - Antiparallel helices held together by interactions between L residues spaced at every 7th position

• **Coiled Coils**
  - 2 or three helices coiled around each other in a left-handed supercoil
  - Multicoil http://jura.wi.mit.edu/cgi-bin/multicoil/multicoil.pl
  - COILS2 http://www.ch.embnet.org/software/COILS_form.html

• **Transmembrane Regions**
  - 20-30aa domains with strong hydrophobicity
  - PHDhtm, PHDtopology, TMpred (TMbase)
  - http://www.embl-heidelberg.de/predictprotein/predictprotein.html

Predicting Secondary Structure

• Recognizing Potential Secondary Structure
  - 50% of a sequence is usually alpha helices and beta sheet structures
  - Helices: 3.6 residues/turn, N+4 bonding
  - Strands: extended conformation, interactions between strands, disrupted by beta bulges
  - Coils: A,G,S,T,P are predominant
  - Sequences with >45% sequence identity should have similar structures

• Databases of sequences and accompanying secondary structures (DSSP)
SS Prediction Algorithms
Chou-Fasman/GOR

- Analyze the frequency of each of the 20 aa in every secondary structure (Chou, 1974)
- A,E,L,M prefer ⎛ helices; P,G break helices
- Use a 4-6aa examination window to predict probability of ⎛ helix, 3-5aa window for beta strands
  - Extend regions by moving window along sequence
- 50-60% effective (Higgins, 2000)
- GOR method assumes that residues flanking the central window/core also influence secondary structure

SS Prediction Algorithms
Neural Networks

- Examine patterns in secondary structures by computationally learning to recognize combinations of aa that are prevalent within a particular secondary structure
- Program is trained to distinguish between patterns located in a secondary structure from those that are not usually located in it
- PHDsec (Profile network from HeiDelberg)
  - ~ 70% correct predictions
  http://www.embl-heidelberg.de/predictprotein/submit_def.html
SS Prediction Algorithms

Nearest Neighbor

- Generate an iterated list of peptide fragments by sliding a fixed-size window along sequence
- Predict structure of aa in center of the window by examining its k neighbors (Yi, 1993)
  - Propensity of center position to adopt a structure within the context of the neighbors
- Method relies on an initial training set to teach it how neighbors influence secondary structure
- NNSSP http://bioweb.pasteur.fr/seqanal/interfaces/nnssp-simple.html

SS Prediction Tools

- **NNpredict** - 65% effective*, outputs H,E,-
  - http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
- **PredictProtein** - query sequence examined against SWISS-PROT to find homologous sequences
  - MSA of results given to PHD for prediction
  - 72% effective*
  - http://www.embl-heidelberg.de/predictprotein/submit_def.html
- **Jpred** - integrates multiple structure prediction applications and returns a consensus, 73% effective*
  - http://www.compbio.dundee.ac.uk/~www-jpred/submit.html
Tertiary Structure Prediction

- **Goal**
  - Build a model to use for comparison with other structures, identify important residues/interactions, determine function

- **Challenges**
  - Reveal interactions that occur between residues that are distant from each other in a linear sequence
  - Slight changes in local structure can have large effects on global structure

- **Methods**
  - **Sequence Homology** - use a homologous sequence as a template
  - **Threading** - search for structures that have similar fold configurations without any obvious sequence similarity

Threading - Approaches

- Sequence is compared for its compatibility (structural similarity) with existing structures

- Approaches to determine compatibility
  - **Environmental Template**: environment of ea. aa in a structure is classified into one of 18 types, evaluate ea. position in query sequence for how well it fits into a particular type (Mount, 2001)
  - **Contact Potential Method**: analyze the closeness of contacts between aa in the structure, determine whether positions within query sequence could produce similar interactions (find most energetically favorable) (Mount, 2001)
Threading Process

- Sequence moved **position-by-position** through a structure
- Protein fold modeled by pair-wise inter-atomic calculations to align a sequence with the backbone of the template
  - Comparisons between local and non-local atoms
  - Compare position i with every other position j and determine whether interactions are feasible
- Optimize model with pseudo energy minimizations - most energetically stable alignment assumed to be most favorable
- Thread the smallest segment reasonable! Computationally intensive.
- 123D [http://123d.ncifcrf.gov/123D+.html](http://123d.ncifcrf.gov/123D+.html)

Model Building

- Perform automated model constructions
  - SWISS-MODEL
    - Compare sequence to ExPdb to find a template (homology)
    - Define your own templates (from threading)
  - GENO3D
    - PSI-BLAST to identify homologs possessing structures to be used as templates
    - [http://geno3d-pbil.ibcp.fr](http://geno3d-pbil.ibcp.fr)
Model Evaluation

• Manually examine model and alignments
• Find similar structures through database searches
  – DALI
• How does the model compare to other structures with the template family?
• Remember, it’s only a MODEL (but even models can be useful)

Structure Visualization

• Different representations of molecule
  – wire, backbone, space-filling, ribbon
• NMR ensembles
  – Models showing dynamic variation of molecules in solution
• VIEWERS
  – RasMol (Chime is the Netscape plug-in)
    • http://www.umass.edu/microbio/rasmol/index2.html
  – Cn3D MMDB viewer (See in 3D) with explicit bonding
  – SwissPDB Viewer
    • http://www.expasy.ch/spdbv/mainpage.html
  – iMol
    • http://www.pirx.com/iMol
Pulling It All Together

- [YFP What is it?](#)
  - Linear Sequence Homology
  - Blast Search: Identify Homologs and Perform MSAs
  - Domain Search: Identify Functional Domains
  - Motif Search: Locate Conserved Sequence Patterns
- [3D Structural Homology](#)
  - Specialized Structure Search: Functional Domain Homology
  - Threading/Model Building: Structural Similarity

Demonstration

- Thread sequence to identify template
  - Web-based: 123D
    - http://123d.ncifcrf.gov/123D+.html
- Model sequence with template
- Visualization
References