Basics of Protein Bioinformatics and Structural Bioinformatics

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB

“The Ten Most Wanted Solutions in Protein Bioinformatics”

- Ana Tramontano; 2007
- What are the problems:
  - Problem One: The challenge involved with detecting the evolutionary relationship between proteins
  - Problem Two: Detection of local similarities between protein sequences to determine functional assignment
  - Problem Three: Function prediction
  - Problem Four: Protein structure prediction
  - Problem Five: Membrane protein
  - Problem Six: Functional site identification
  - Problem Seven: Protein-protein interaction
  - Problem Eight: Protein-ligand interaction
  - Problem Nine: Protein design (to design completely new proteins)
  - Problem Ten: Protein engineering (to modify the properties of proteins)

Foldit: a multiplayer online game

- Predicting protein structures with a multiplayer online game (Nature Volume: 466, Pages: 756–760, 2010)
- “The integration of human visual problem-solving and strategy development capabilities with traditional computational algorithms through interactive multiplayer games”
  - Top-ranked Foldit players excel at solving challenging structure refinement problems in which substantial backbone rearrangements are necessary to achieve the burial of hydrophobic residues.
  - Players working collaboratively develop a rich assortment of new strategies and algorithms; unlike computational approaches, they explore not only the conformational space but also the space of possible search strategies.

Basics of proteins

- Amino acids
  - 20 amino acids
  - Hydrophobic / hydrophylic
  - Charged / neutral
- Functions
  - Enzymes
  - Structure protein
  - Channel
- Structures
  - Primary structure (sequence; SwissProt)
  - Secondary structure
  - Tertiary structure (PDB)

Amino acids properties

- Why 20aa?
- Reduced alphabet?

---

Amino acid structure

Image from: http://www.chemistrydaily.com/chemistry/upload/c/c5/Amino_acids_2.png

Amino acids properties

Image from: http://www.jalview.org/help/html/misc/properties.gif
Proteins: polypeptides

Resonance forms

Peptide bond

Protein backbone torsion angles

Ramachandran plot

Protein secondary structures

Local structures which are typically recognized by specific backbone torsion angles and specific mainchain hydrogen bond pairing patterns

Protein tertiary structure

- PDB (Protein Data Bank; text files)
  - More than 60k structures as of Nov 2009
- Structure visualization (PyMol)

A PDB example file 1dhy

Protein structure visualization

- PyMol (produces images of high quality)
- Swiss-PDBViewer (DeepView)
- Chimera
- Rasmol
- WebMol
- Jmol
Sequence-structure-function

Many computational problems!!

Multiple domain proteins
- Domain reshuffling
- Proteins, especially Eukaryotic proteins have multiple domains

Protein domain prediction problem
- Sequence based
- Structure based

Protein (domain) classification
- Classification
  - Families
- Sequence-based
- Structure-based
  - SCOP
  - CATH

Protein domain
- Domains are units of:
  ✓ compact structure
  ✓ function and evolution
  ✓ folding

Protein domain prediction problem
- Pfam overview
  - http://pfam.sanger.ac.uk/
  - Pfam is a large collection of MSAs and HMMs covering many common protein domains and families (flat organization; clan)
    - Version 22.0 Jul 2007, 9318 families
    - Version 24.0 Oct 2009, 11912 families
    - E.g., SH2, zf-C3HC4
  - hmmer package
    - Sensitive database searching against Pfam
    - Available at BigRed!!
    - hmmer3
  - Use online Pfam database & do local domain prediction using Himmer
Pfam family example: SH2

- http://pfam.sanger.ac.uk/family?acc=PF00017
- An overview of this domain/family
- Alignment
- Domain architecture
- Species distribution
- Phylogenetic tree
- Other information

SCOP classification

- Structural Classification Of Proteins
- 1.75 release (June 2009)
  - 38221 PDB Entries. 1 Literature Reference. 110800 Domains

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of folds</th>
<th>Number of superfamilies</th>
<th>Number of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha proteins</td>
<td>284</td>
<td>597</td>
<td>851</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>174</td>
<td>314</td>
<td>742</td>
</tr>
<tr>
<td>Alpha and beta proteins (a/b)</td>
<td>347</td>
<td>214</td>
<td>803</td>
</tr>
<tr>
<td>Alpha and beta proteins (a+b)</td>
<td>376</td>
<td>233</td>
<td>399</td>
</tr>
<tr>
<td>Multi-domain proteins (a/b)</td>
<td>96</td>
<td>66</td>
<td>86</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>398</td>
<td>110</td>
<td>125</td>
</tr>
<tr>
<td>Small proteins</td>
<td>96</td>
<td>129</td>
<td>219</td>
</tr>
<tr>
<td>Total</td>
<td>1295</td>
<td>1362</td>
<td>3482</td>
</tr>
</tbody>
</table>

http://scop.mrc-lmb.cam.ac.uk/scop/

SCOP hierarchy

- SCOP classes
  - All alpha proteins
  - All beta proteins
  - Alpha and beta proteins (a/b)
    - Mainly parallel beta sheets (beta-alpha-beta units)
    - Mainly antiparallel beta sheets (segregated alpha and beta regions)
  - Multi-domain proteins (alpha and beta)
    - Folds consisting of two or more domains belonging to different classes
    - Membrane and cell surface proteins and peptides
  - Small proteins
  - Coiled coil proteins
  - Designed proteins

SCOP classification

1dlw
1. Root: scop
2. Class: All alpha proteins
3. Fold: Globin-like
4. Superfamily: Globin-like
5. Family: Truncated hemoglobin
6. Protein: Protozoan/bacterial hemoglobin
7. Species: Ciliate (Paramecium caudatum)

SCOP classification

1dlw
1. Root: scop
2. Class: All alpha proteins
3. Fold: Globin-like
4. Superfamily: Globin-like
5. Family: Truncated hemoglobin
6. Protein: Protozoan/bacterial hemoglobin
7. Species: Ciliate (Paramecium caudatum)

CATH classification

- Hierarchical classification of protein domain structures
  - Four major levels: Class, Architecture, Topology and Homologous superfamily (correspond to SCOP’s class, +, fold, superfamily)
- Not always consistent with SCOP classification

http://www.cathdb.info/

Architecture

CATH architecture describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures; assigned manually
SCOP & CATH

- Both heavily rely on manual inspections
  - “The boundaries and assignments for each protein domain are determined using a combination of automated and manual procedures which include computational techniques, empirical and statistical evidence, literature review and expert analysis” (CATH website)
- Both lag behind the determination of new structures
  - CATH 3.4 release (year?)
  - SCOP 1.75 release (June 2009)
- Automatic classification of structures is still a challenge

Check out the databases online

- SwissProt
- PDB
- SCOP
- CATH
- Pfam

Protein Structural Bioinformatics

- Experimental determination of structures
- Structural comparison
- Protein folding
  - Folding simulations
  - Folding pathway
- Structure prediction
  - Secondary structure
  - Side-chain prediction
  - Tertiary structure prediction
    - Comparative modeling
    - De novo prediction

Anfinsen’s theory of protein folding

“The native conformation is determined by the totality of interatomic interactions and hence, by the amino acid sequence, in a given environment”.

Experimental determinations of protein structures

- X-ray crystallography (need crystals)
- NMR (for small proteins)
- EM (for large complex structures; could be a powerful tool when combined with protein structure models)

Structural genomics

“The Protein Structure Initiative (PSI) is a federal, university, and industry effort aimed at dramatically reducing the costs and lessening the time it takes to determine a three-dimensional protein structure. The long-range goal of the PSI is to make the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences”

-------- from NIGMS PSI web site
Protein structure comparison

- A key approach to protein structural analysis
  - Structure/function relationship
  - Evolution of protein structures
  - Structure classification
  - Distant homology detection
- Specific goal: to detect the largest common substructure between two proteins

Structure comparison: Old problem

- Early programs
  - Introduction of the RMSD measure
- Fully automated servers
  - DALI Distance ALIgnment tool
  - VAST Vector alignment search tool

What’s structure alignment

Simple case – two closely related proteins with the same number of amino acids.

Find a transformation to achieve the best superposition

Minimize \( \text{rmsd} \) of distances \( 1-1, \ldots, 7-7 \)

\[
\text{rmsd} = \sqrt{\frac{1}{N} \sum (x(i) - y(i))^2}
\]

Otherwise, structure comparison is a difficult problem!!!
Distance Matrix

- Similar 3D structures have similar inter-residue distances
  
  Colicin A common core
  
  Myoglobin common core

Proteins are flexible -- so we need flexible structure comparison

TM0293: The closest homolog (17% id) has a nice active site

Protein folding

- Protein folding is the physical process by which a polypeptide folds into its characteristic 3D structure (native structure)
- Related problems
  - Folding pathway
  - What's the intermediate structures
  - Folding speed (contact order)
  - Energy landscapes
  - Misfolding

Protein folding: what we do NOT know

The Levinthal Paradox

"Despite the huge space of possible conformations, proteins fold reliably and quickly to their native conformation"

Energy landscape theory of protein folding

- Interactions between side chains largely favor the molecule's acquisition of the folded state (evolutionary selection).
- Protein can fold to the native state through any of a large number of pathways and intermediates (folding funnel).
Molecular dynamics simulations

\[ \mathbf{F} = m \mathbf{a} = -\nabla U \]

In molecular dynamics one integrates numerically Newton’s equations of motion and thus generates a trajectory for the molecule.

Molecular mechanics force fields

- van der Waals energy
- Electrostatic energy
- Hydrogen bond
- Bond energy
- Bond angle energy
- Dihedral angle energy
- Solvation

A force field is made up by the contributions of many terms that represent the different types of interactions between the atoms of the protein molecule (energy function).

Molecular dynamics simulations of protein folding

Molecular dynamics simulations enable the sampling of the states of proteins and the calculation of possible folding pathways.

Can protein structure be predicted from their sequences?

- Many proteins fold spontaneously to their native structure
- Protein folding is relatively fast (nsec – sec)
- Chaperones speed up folding, but do not alter the structure

The protein sequence contains all information needed to create a correctly folded protein (Anfinsen principle).

Can we model protein structures from their protein sequences?

Protein structure prediction

- Secondary structure
- Side-chain prediction
- Tertiary structure prediction
  - Comparative modeling
  - De novo prediction

Secondary structure prediction

- Easier than 3D structure prediction (more than 40 years of history).
- Accurate secondary structure prediction can be an important information for the tertiary structure prediction
- Protein function prediction
- Protein classification
- Protein alignment (fold recognition) using secondary structure information
Prediction methods

- Statistical method
- Chou-Fasman method, GOR I-IV
- Nearest neighbors
- NNSSP, SSPAL, Fuzzy-logic based method
- Neural network
- PHD (profile-based neural network), Psi-Pred, J-Pred
- Support vector machine (SVM)
- HMM
- Many available as servers
- Standalone software: PSSpred (http://zhanglab.ccmbr.med.umich.edu/PSSpred/)

DEE algorithm for side-chain conformation prediction

DEE (dead end elimination) facilitates the search for the best solution by systematically eliminating high-energy rotamers that can be rigorously excluded from the global minimum energy solution of the system.

Comparing with rotamer \( i \), rotamer \( j \) at a given position may be eliminated if the inequality holds.

\[ E(i) + \sum \min_E(i, j) > E(i) + \sum \min_E(i, j), i \neq j. \]

Tertiary structure prediction based on thermodynamics of protein folding

- Molecular dynamics methods & Brownian dynamics methods
- (1975) Levitt and Warshel used a simplified protein structure representation and successfully folded a small protein (bovine pancreatic trypsin inhibitor, BPTI, 58 amino acid residues) into its native conformation from an open-chain conformation using energy minimization.
- (1998) Duan and Kollman reported a simulation experiment of one small protein (the villin headpiece subdomain, 36 amino acid residues), running on a Cray T3D and then a Cray T3E supercomputer, that took months of computation with the entire machine dedicated to the problem.

Comparative modeling (homology modeling)

- Browne and co-workers (1969) modeled the structure of \( \alpha \)-lactalbumin using the X-ray structure of lysozyme as a template.
- All comparative modeling packages follow similar steps
  - Find template & get sequence-template alignment
    - Sequence alignment
    - Sequence-structure alignment
  - "Transfer" the coordinates from the templates to the sequence (backbone & sidechain)
  - Predict the structure of missing loops & sidechains
- Packages:
  - Modeller (Sali)
  - Rosetta (Baker)
Finding templates
- Sequence based (pairwise sequence, profile-profile sequence alignment)
- Fold recognition or threading
  - Threading: aligning a protein sequence with one or more protein structures
  - Residue-structure environment compatibility (3D-profile) (123D, FUGUE)
  - Statistical potential model (GenTHREADER, PROSPECT)

Packages and servers for modeling
- Modeller (Sali)
- ROSETTA (ROBETTA) (David Baker)
  - http://zhang.bioinformatics.ku.edu/I-TASSER
- I-TASSER (Yang Zhang)

Modeling with cryo-EM
- Model a sequence using both template and cryo-EM data
  - Build models
  - Fit models into cryo-EM maps
  - Refine models with loop modeling
  - Flexible cryo-EM fitting with Flex-EM
- Modeller package provides this functionality (Flex-EM)
- Gorgon, an iterative molecular modeling system

Flex-EM
- It includes a rigid fitting stage followed by a refinement stage. Rigid fitting can be performed with Mod-EM or any other rigid fitting methods. The refinement stage starts with the components rigidly fitted in the approximate positions in the map. Two methods are available: conjugate gradients minimization (CG) and simulated annealing molecular dynamics (MD).
- The atomic positions are optimized with respect to a scoring function that includes the crosscorrelation coefficient between the structure and the map as well as stereochemical and nonbonded interaction terms.

An initial structure (white) and Flex-EM refined structure (purple) into EM map (image from: http://salilab.org/modeller/ncmi_2008/flexible.html)

De novo protein structure prediction
- Also called “Ab initio protein structure prediction” and “Free-modeling”
- Fragment assembly based methods are the most successful ones
- David Baker’s Rosetta (Robetta)
  - Use segments to narrow the conformational search space
  - Based on the assumption that short sequence segments have strong local structural biases
  - Assembly of segments into structures
  - Conformational space search (Monte Carlo & other minimization methods) + energy calculation
- Unfortunately there is no Baker’s Algorithm
- Zhang and Skolnick’s TASSER (fragments are from the threading results)
- Zhang’s QUARK (ranked no 1 in CASP9 & CASP10)

CASP 8 in numbers
- Number of human expert groups registered 113
- Number of prediction servers registered 122
- Total number of targets released (human/server targets) 128 (57)

Modeling of complex structures
- Very difficult
- Docking
- Integration of high-resolution structures/models and the Electron Microscopy (EM) density map.
  - The basic idea is to fit known high-resolution structures into low-resolution structures of large complexes that are determined by EM to get refined structure of large complexes.
  - Solved the structures of large biological machines/macromolecular complexes, such as viruses, ion channels, ribosomes and proteasomes.
  - Predicted models of the individual proteins may instead be used in fitting.
**Blind test of modeling & beyond**

- **CASP**: Critical Assessment of Techniques for Protein Structure Prediction
- **CAPRI** (Critical Assessment of Predicted Interactions)
- **Design?**
  - Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology (JMB, 2011)
  - A total of 28 research groups took up the challenge of determining what is missing: what distinguish between structures of 87 designed complexes (very favorable computed binding energies but which do not appear to be formed in experiments) and 120 naturally occurring
  - The community found that electrostatics and solvation terms partially distinguish the designs from the natural complexes, largely due to the nonpolar character of the designed interactions.

**CASP10 results**

- "We compare results of the community efforts in modeling protein structures in the tenth CASP experiment, with those in earlier CASPs, particularly in CASP5, a decade ago. There is a substantial improvement in template based model accuracy as reflected in more successful modeling of regions of structure not easily derived from a single experimental structure template, most likely reflecting intensive work within the modeling community in developing methods that make use of multiple templates, as well as the increased number of experimental structures available. Deriving structural information not obvious from a template is the most demanding as well as one of the most useful tasks that modeling can perform. Thus this is gratifying progress. By contrast, overall backbone accuracy of models appears little changed in the last decade. This puzzling result is explained by two factors - increased database size in some ways makes it harder to choose the best available templates, and the increased intrinsic difficulty of CASP targets, as experimental work has progressed to larger and more unusual structures. There is no detectable recent improvement in template free modeling, but again, this may reflect the changing nature of CASP targets."