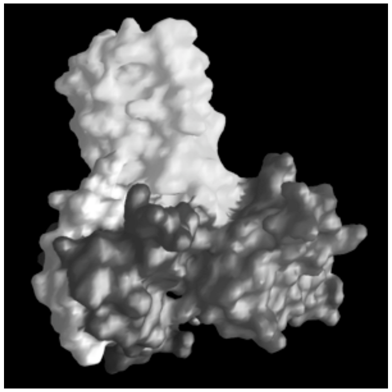
Bioinformatics

Proteins III. -
Comparing and Predicting Protein Structures



Robert Latek, Ph.D.
Bioinformatics, Biocomputing

Proteins I.-III. - Syllabus

- Proteins I.
  - Phylogenetic Trees
  - Multiple Sequence Alignments
- Proteins II.
  - Searching for Homologous Sequences
  - Working with Protein Structure Information
- Proteins III.
  - Comparing Protein Structures
  - Predicting Protein Structures

Last Week

- Searching for Homologous Sequences
  - Patterns
  - Profiles
  - Database Searches
- Working with Protein Structure Information
  - Coordinate Files, Classification, Databases

Structure Databases

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

- CATH: Classification by Class, Architecture, Topology, and Homology
  - http://www.biochem.ucl.ac.uk/bsm/cath/

- FSSP: Fold classification based on Structure-Structure alignment of Proteins
  - DALI pair-wise alignments
  - http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html

- SARF: Spatial Arrangement of Backbone Fragments
  - http://123d.ncifcrf.gov/

Proteins II. - Syllabus

- Comparing Protein Structures
  - Aligning 3D Structures
    - Comparison of Identical, Homologous, Unrelated Structures
    - Translation & Rotation Methods
- Predicting Protein Structure
  - Secondary Structure Prediction
  - Specialized Structural Regions
  - Tertiary Structure Prediction
    - Threading
    - Modeling
- Structure Viewers

Sequence & Structure Homology

- Multiple Sequence Alignments
  - – Identify relationships between sets of proteins
- Structure Classification Databases
  - – Categorize related structures
    - Structure families do not necessarily share sequence homology

Structure Comparison

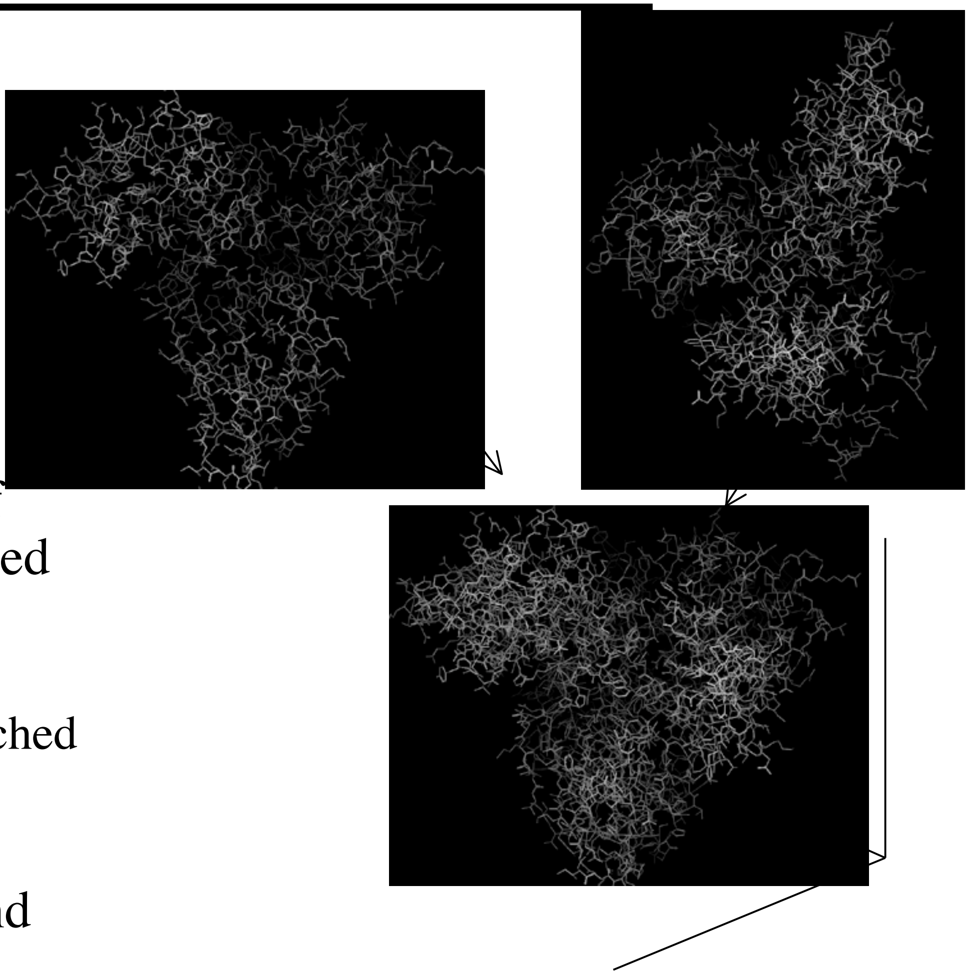
- Compare Structures that are:
  - Identical
    - Similarity/difference of independent structures, x-ray vs. nmr, apo vs. holo forms, wildtype vs. mutant
  - Similar
    - Predict function, evolutionary history, important domains
  - Unrelated
    - Identify commonalities between proteins with no apparent common overall structure - focus on active sites, ligand binding sites
- Superimpose Structures by 3D Alignment

Structural Alignment

- Sequence alignment based on LINEAR sequence similarity
- 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

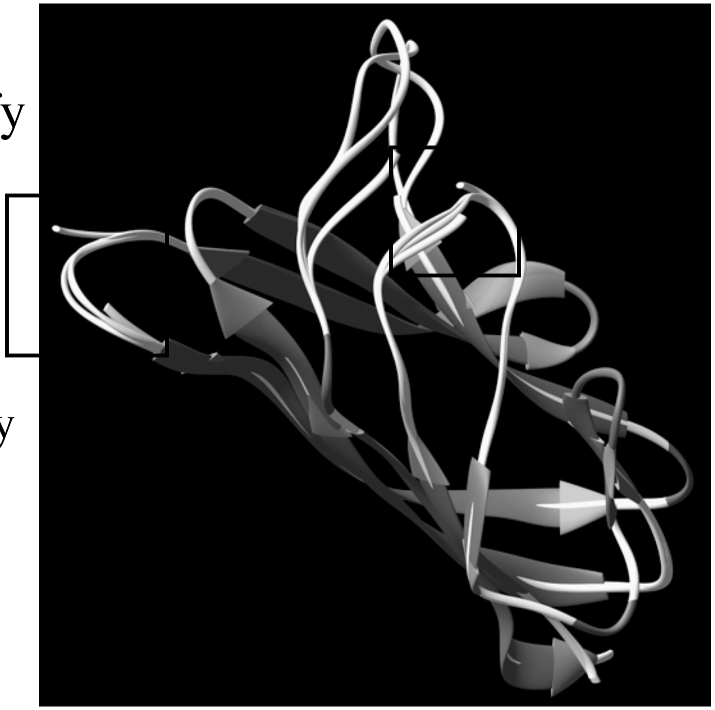
Translation and Rotation

- Alignment
  - Translate center of mass to a common origin
  - Rotate to find a suitable superposition
- Methods
  - 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 C□-C□ 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:
  - Dynamic Programming
  - Distance Matrix
  - Fast, Secondary Structure-Based



Alignment by 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 C□-C□ bond, superimpose vectors

Alignment by 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 C□ distances)
- Similar structures have super-imposable graphs
- Similarity score based on the degree by which matched elements can be superimposed (approximate solution)
- DALI (Distance ALIgnment tool)
  - Used to build FSSP database of structure comparisons

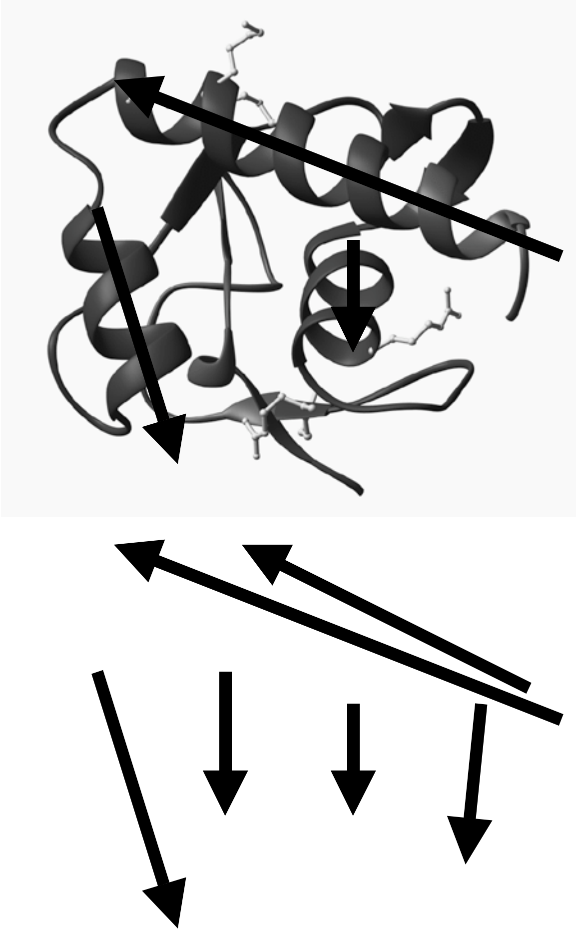
DALI Distance Alignment Method

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

Fast Alignment by Secondary Structure

- 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

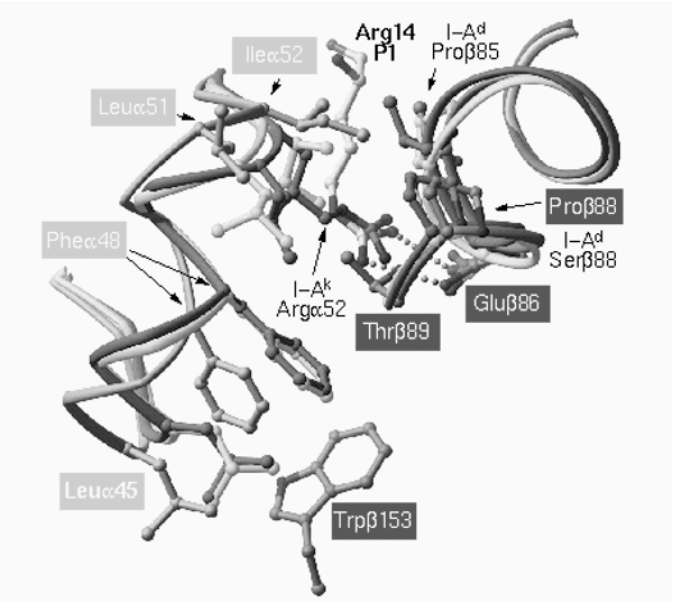
http://www.ncbi.nlm.nih.gov:80/Structure/VAST/vastsearch.html

- SARF

http://123d.ncifcrf.gov/

Quality of Alignments

- 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



$$\text{RMSD} \quad \sqrt{\sum_{i}^{N} D_i / N}$$

Proteins II. - Syllabus

- Comparing Protein Structures
  - Aligning 3D Structures
    - Comparison of Identical, Homologous, Unrelated Structures
    - Translation & Rotation Methods
- Predicting Protein Structure
  - Secondary Structure Prediction
  - Specialized Structural Regions
  - Tertiary Structure Prediction
    - Threading
    - Modeling
- Structure Viewers

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)

Secondary Structure Prediction Methods

- Methods
  - Chou-Fasman/GOR
  - Neural Networks
  - Nearest Neighbor

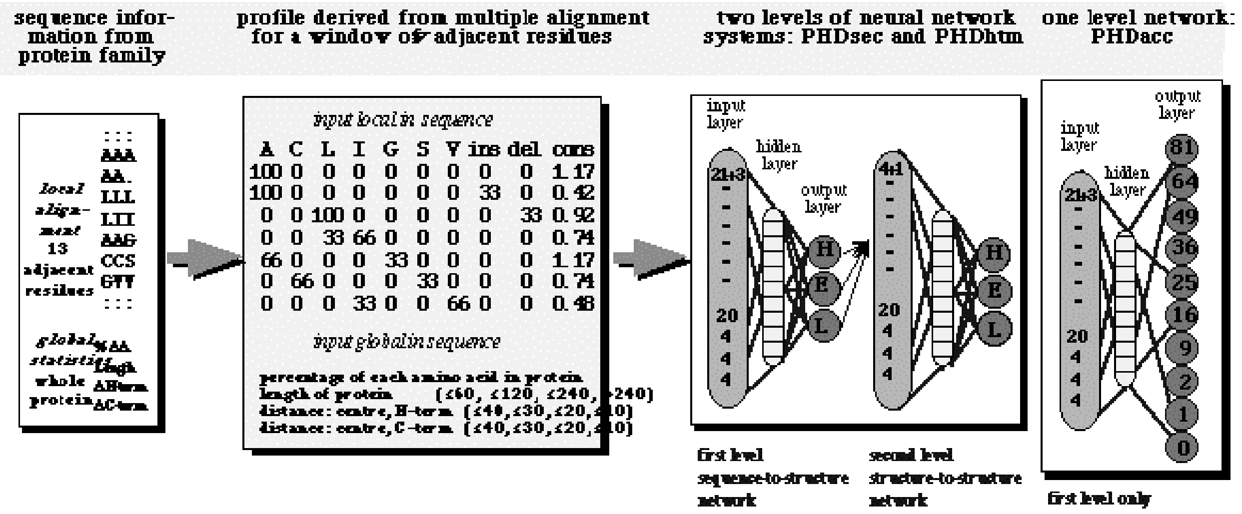
Chou-Fasman/GOR Prediction

- 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

Neural Network Prediction

- 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

PHD Neural Network



A	C	L	I	G	S	V	ins	del	cons
100	0	0	0	0	0	0	0	0	1.17
100	0	0	0	0	0	0	33	0	0.42
0	0	100	0	0	0	0	0	33	0.92
0	0	33	66	0	0	0	0	0	0.74
66	0	0	0	33	0	0	0	0	1.17
0	66	0	0	0	33	0	0	0	0.74
0	0	0	33	0	0	66	0	0	0.48

(Rost Burkhard Meth In Enzym)

http://www.embl-heidelberg.de/predictprotein/submit_def.html

Nearest Neighbor Prediction

- 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://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html

Secondary Structure 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://jura.ebi.ac.uk:8888/

*(Higgins)

Predicting Specialized Structures

- Leucine Zippers
  - Antiparallel □ helices held together by interactions between L residues spaced at ever 7th position
- Coiled Coils
  - 2 or three a helices coiled around each other in a left-handed supercoil
  - Multicoil http://gaiberg.wi.mit.edu/cgi-bin/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

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 interatomic 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.

Threading Tools

- 123D http://123d.ncifcrf.gov/123D+.html
- Prospect (Fladda)

Model Building

- Perform automated model constructions
  - SWISS-MODEL
    - Compare sequence to ExPdb to find a homolog
    - Define your own templates (from threading)
    - http://www.expasy.ch/swissmod/SWISS-MODEL.html
  - GENO3D
    - PSI-BLAST to identify homologs possessing structures to be used as templates
    - 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)

Proteins I.-III. Summary

- Proteins I.
  - Phylogenetic Trees
  - Multiple Sequence Alignments
- Proteins II.
  - Searching for Homologous Sequences
  - Working with Protein Structure Information
- Proteins III.
  - Comparing Protein Structures
  - Building Structural Models

Visualizing Structural Information

- Hand edit files
- 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.htm
  - Cn3D MMDB viewer (See in 3D) with explicit bonding
    - http://www.ncbi.nlm.nih.gov/Structure
  - SwissPDB Viewer
    - http://www.expasy.ch/spdbv/mainpage.html

Demo

- Identify segment to thread and model
  - Align homologs, define important domains
- Thread sequence to identify template
  - Web-based: 123D http://123d.ncifcrf.gov/123D+.html
  - Locally: Prospect (on Fladda)
    - http://compbio.ornl.gov/structure/prospect/
    - % bsub prospect -top 20 yfs.seq
- Model sequence with template
  - http://www.expasy.ch/swissmod/SWISS-MODEL.html
- Visualization

References

Bioinformatics: Sequence and genome Analysis. David W. Mount. CSHL Press, 2001.
Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Andreas D. Baxevanis and B.F. Francis Ouellete. Wiley Interscience, 2001.
Bioinformatics: Sequence, structure, and databanks. Des Higgins and Willie Taylor. Oxford University Press, 2000.
Chou, P.Y. and Fasman, G. D. (1974). Biochemistry, 13, 211.
Yi, T-M. and Lander, E.S.(1993) J. Mol. Biol., 232,1117.