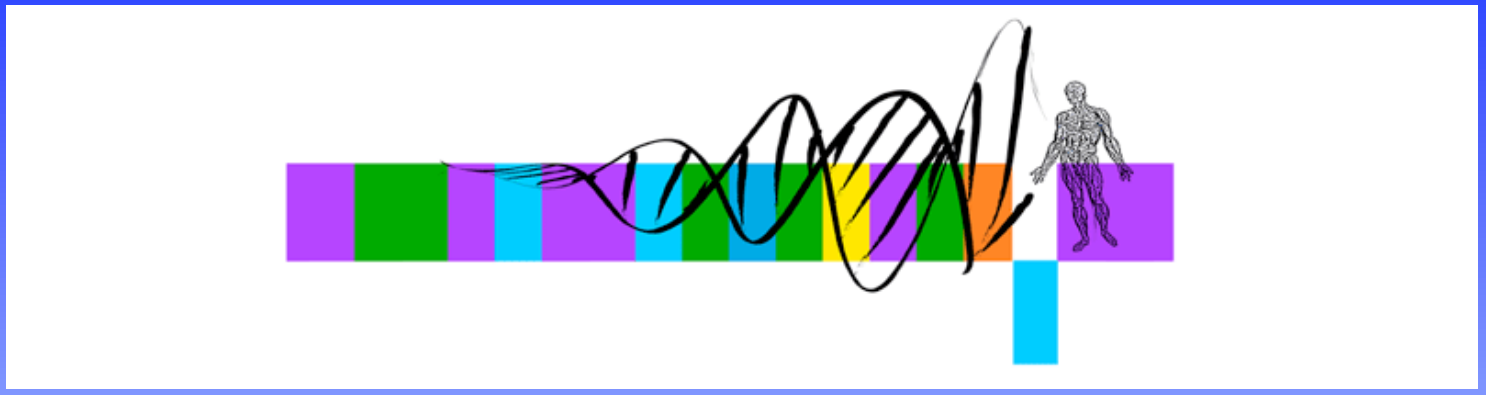


Getting To Know Your Protein

Comparative Protein Analysis:

Part III. Protein Structure Prediction and Comparison

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

Comparative Protein Analysis

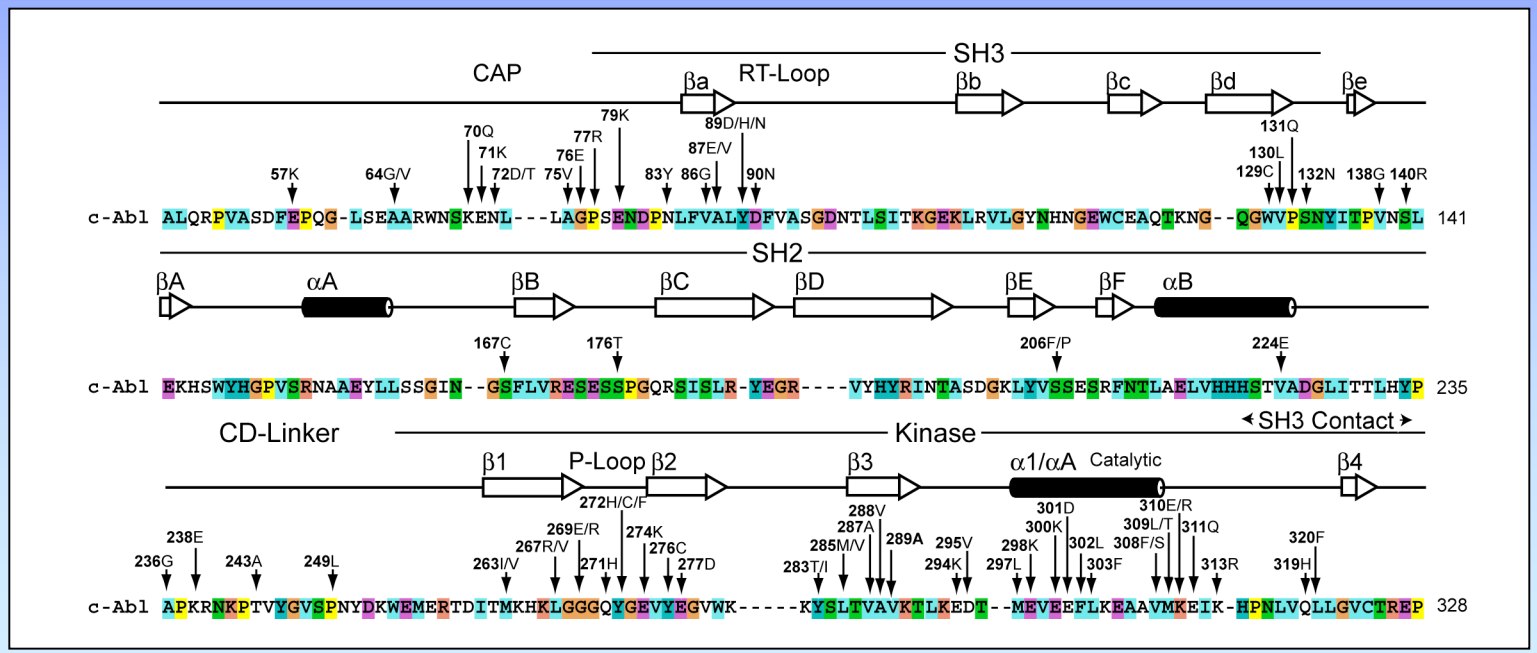
- Global Sequence Comparisons (Trees and MSAs)
  - Bootstrapping
- Localized Sequence Comparisons (Patterns and Profiles)
  - MEME
    http://jura.wi.mit.edu/bio/education/bioinfo2005/proteins/meme.html

Comparative Protein Analysis

- Structural Comparisons
  - Why are protein structure prediction and analysis useful?

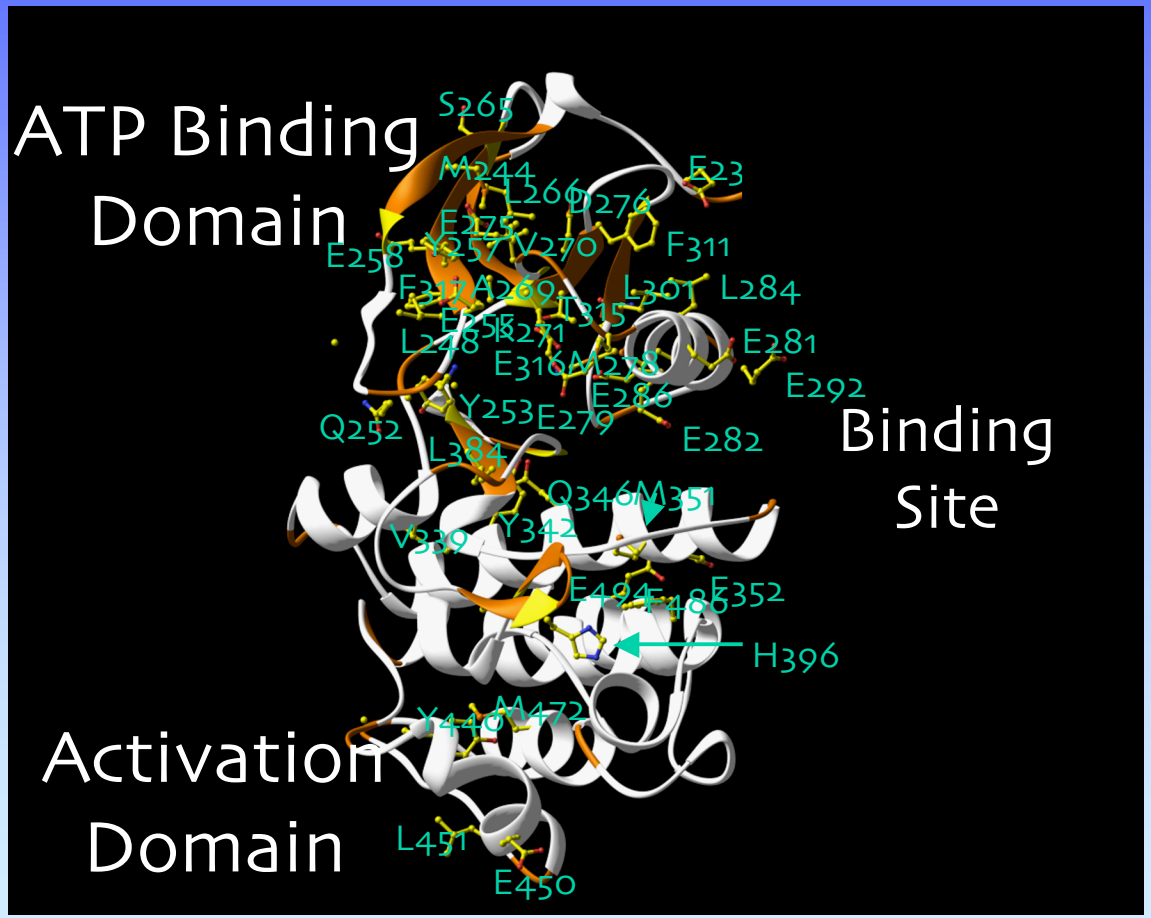
Linear Sequences Contain Densely Encoded Information

- Properties (charge, hydrophobicity)
- Function (mechanisms, contacts)
- Folding (secondary, tertiary structure)



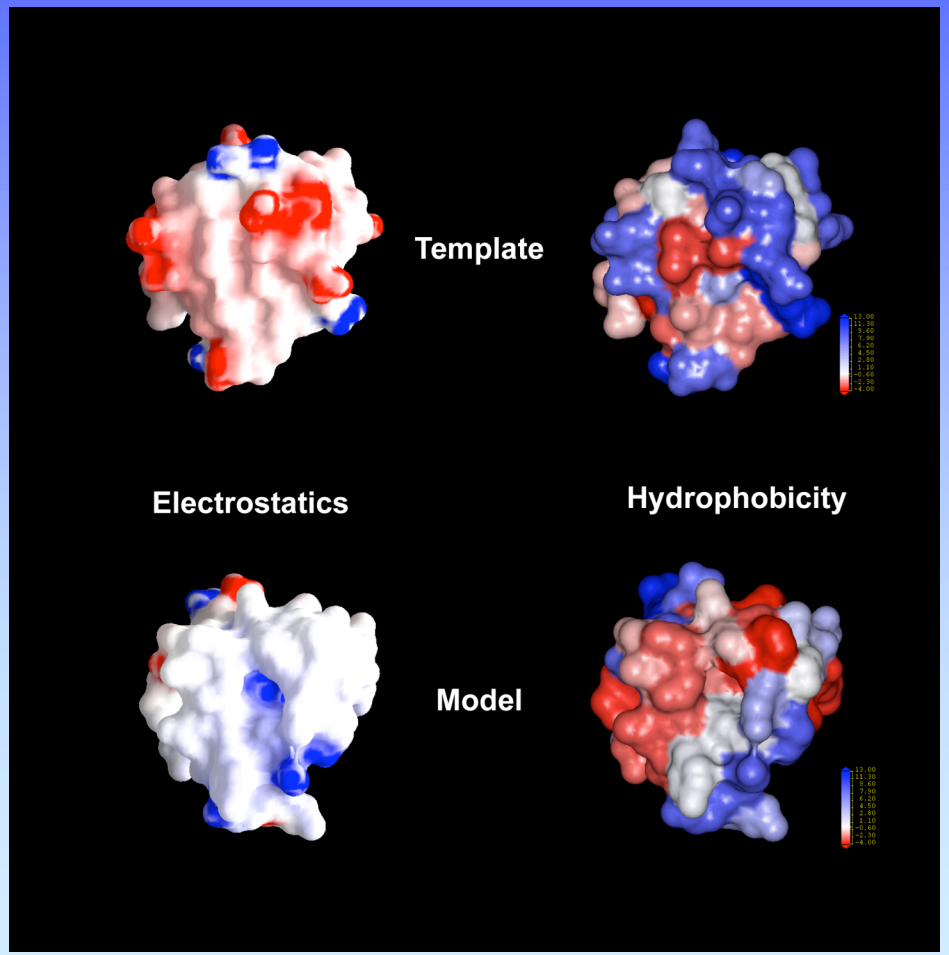
Locating Important AAs

- Identify Mutants
  - Function
    - efficiency
  - Folding
    - misfolding
  - Interactions
  - Localization
    - solubility

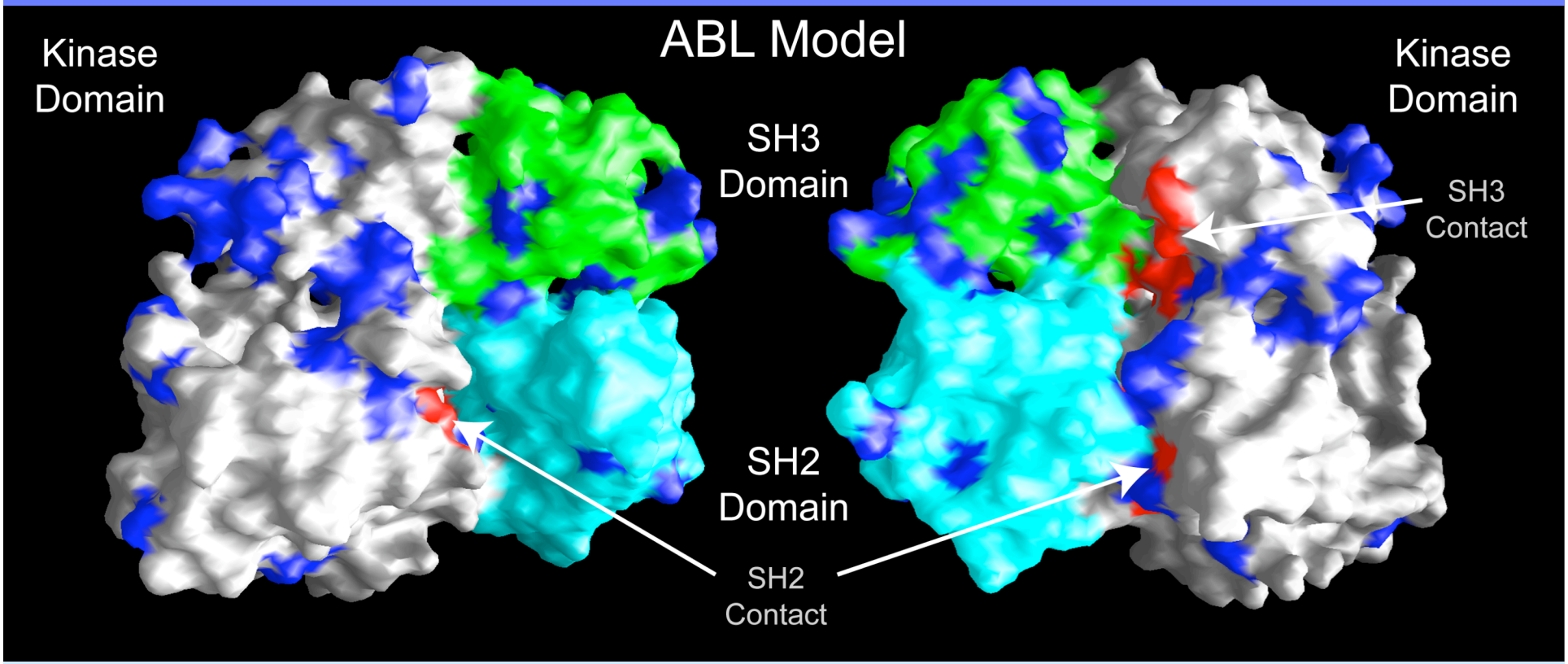


Surface Comparisons

- Topology
- Electrostatics
- Hydrophobicity



Protein Interfaces

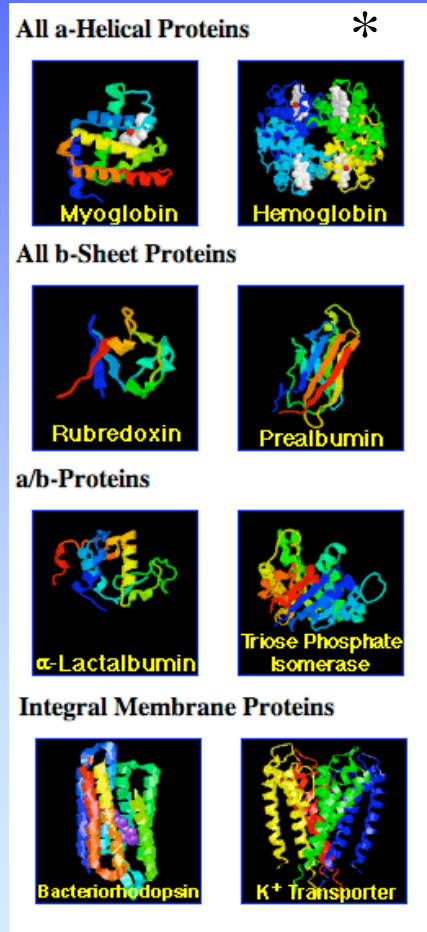


Syllabus

- **Structure Coordinates**
  - **Files & Databases**
- Structure Comparisons
  - Aligning 3D Structures
- Structure Classification
  - Structure Families
- Structure Prediction
  - Specialized Structural Regions
  - Secondary Structure Prediction
  - Tertiary Structure Prediction
- Structure Visualization

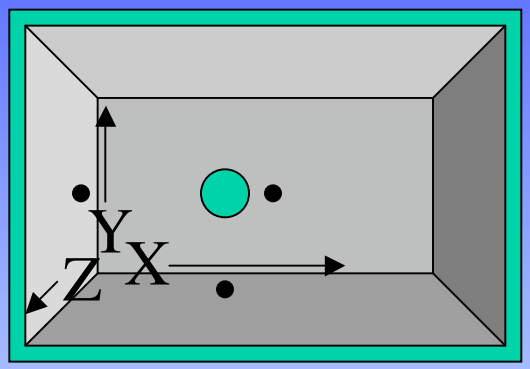
Structure Classification

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



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

Coordinates



Projections of atom on 3 planes

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

Atom	Residue	Chain	X	Y	Z
54	ALA	C	35.4	-9.3	102.5

Coordinate File Formats

- **MMDB** "Molecular Modeling DataBank" Format
  - ASN.1 standard data description language
  - Explicit bond approach - consistent bonding information
- **PDB** "Protein DataBank" Format
  - Column oriented, "flexible format"
  - Chemistry rules approach - connect dots using standard rules to specify bond distances (not consistent among applications)

Example PDB File

tag	Atom#	Atom type	Residue	Chain	Residue#	X	Y	Z	Structure scores	
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

Coordinate Databases

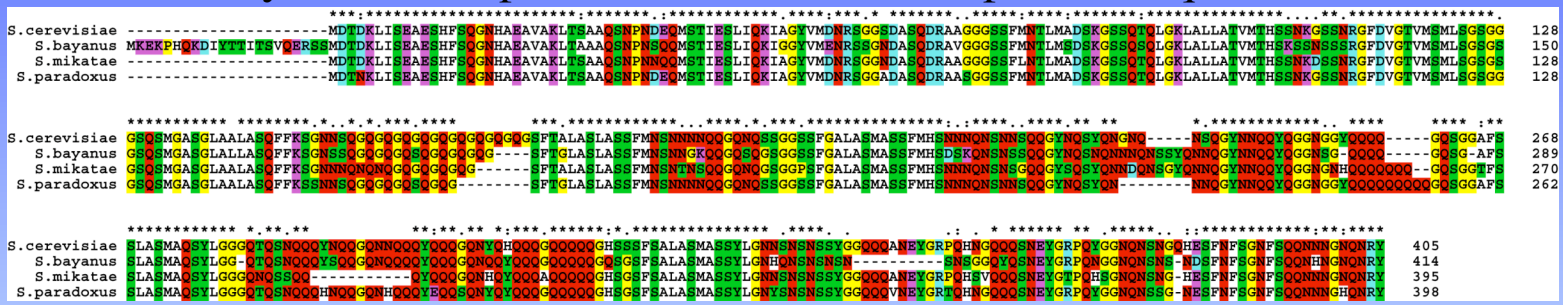
- **RCSB** (Research Collaboratory for Structural Bioinformatics) http://www.rcsb.org/
  - Formally know as the Protein Data Bank at Brookhaven National Laboratories
  - Structure Explorer PDB search engine
    - Text and PDB ID (4 letter code) searching
- **MMDB** (Molecular Modeling Database @NCBI)
  - Compilation of structures represented in multiple formats
  - Provides structure summaries
  - BLAST sequences to search for available structures

Syllabus

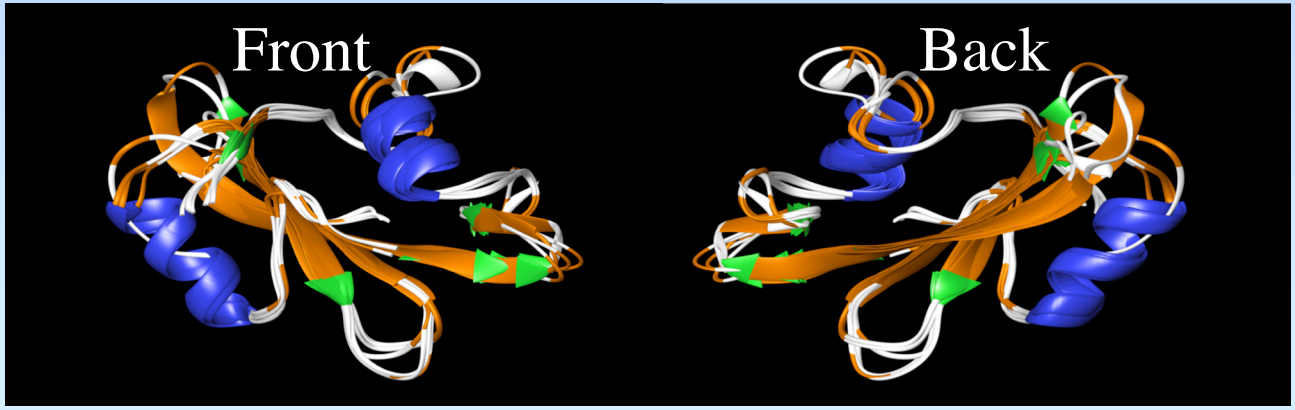
- Structure Coordinates
  - Files & Databases
- **Structure Comparisons**
  - **Aligning 3D Structures**
- Structure Classification
  - Structure Families
- Structure Prediction
  - Specialized Structural Regions
  - Secondary Structure Prediction
  - Tertiary Structure Prediction
- Structure Visualization

Sequence & Structure Homology

- **Sequence Homology**
  - Identify relationships between sets of linear protein sequences



- **Structure Homology**
  - Categorize related structures based on 3D folds
    - 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

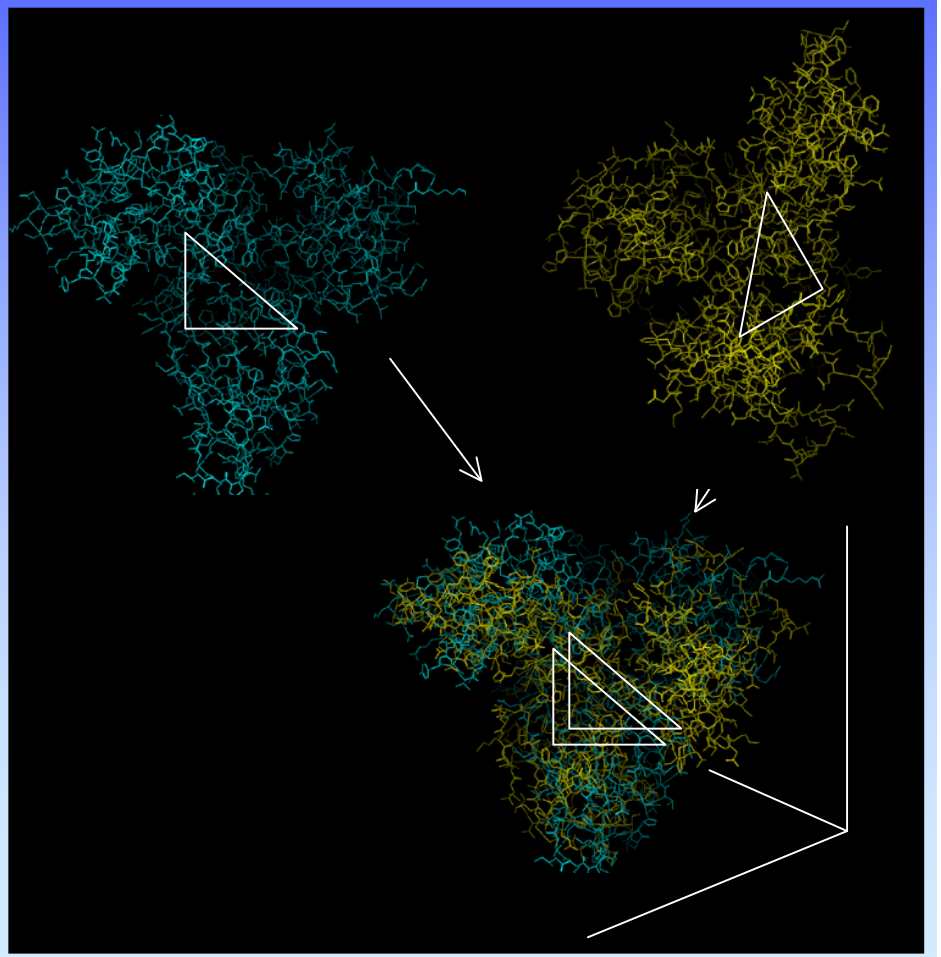
Structural Alignment

- 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

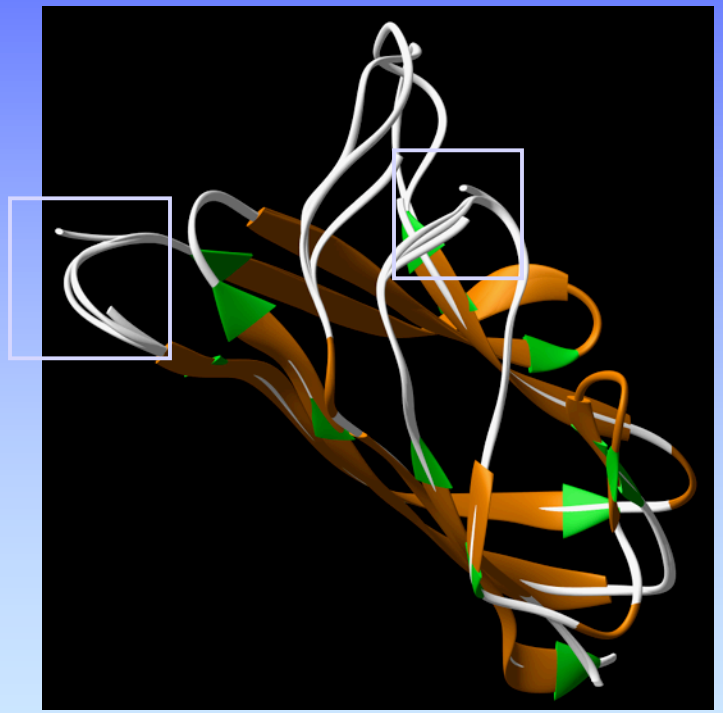
Translation and Rotation

- **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



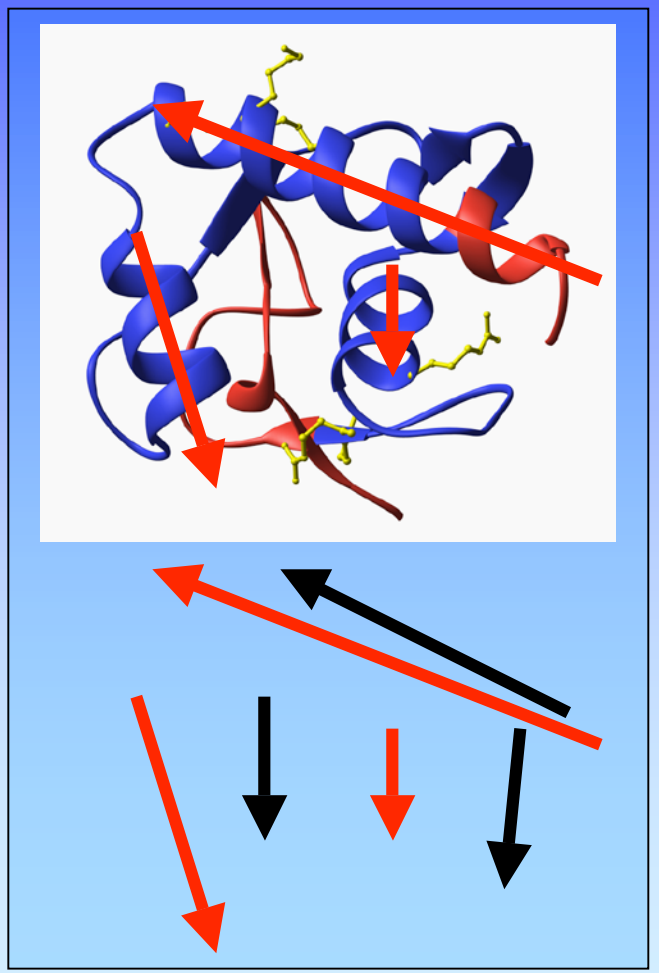
VAST and SARF

- Secondary structure elements can be represented by a vector (Position & length)
- Compare the arrangement of clustered vectors between two structures to identify common folds
- Supplement with information about side chain arrangement (burial/exposure)
- VAST

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

- SARF

http://123d.ncifcrf.gov/



Exhaustive Alignment

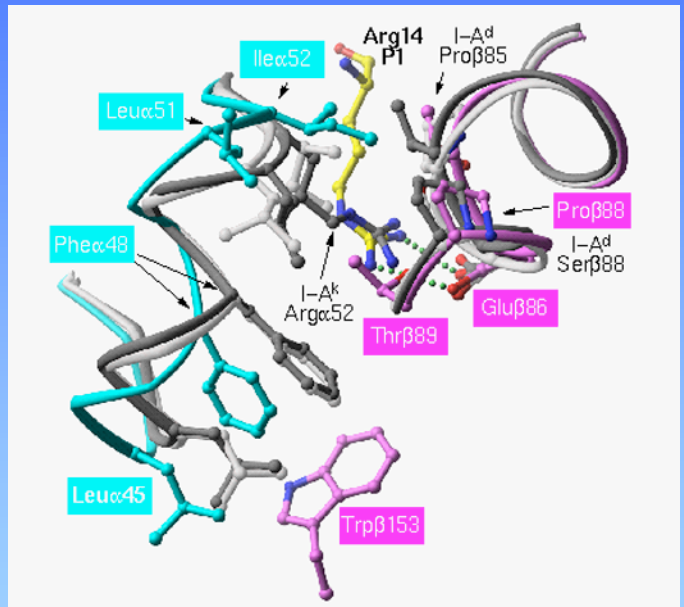
- **Dynamic Programming**
  - Local environment defined in terms of Interatomic distances, bond angles, side chain identity, side chain burial/exposure
  - Align structures by matching local environments - for example, draw vectors representing each Cα-Cβ bond, superimpose vectors
- **Distance Matrix**
  - 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 your structure to PDB structures
- Helps identify potentially biologically interesting similarities not obvious by sequence comparisons

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



$$RMSD = \sqrt{\sum^{N} D_i^2/N}$$

Syllabus

- Structure Coordinates
  - Files & Databases
- Structure Comparisons
  - Aligning 3D Structures
- **Structure Classification**
  - **Structure Families**
- Structure Prediction
  - Specialized Structural Regions
  - Secondary Structure Prediction
  - Tertiary Structure Prediction
- Structure Visualization

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 (3D alignments)
  - Representative members can be used for modeling sequences of unknown structure

Structure Family Databases

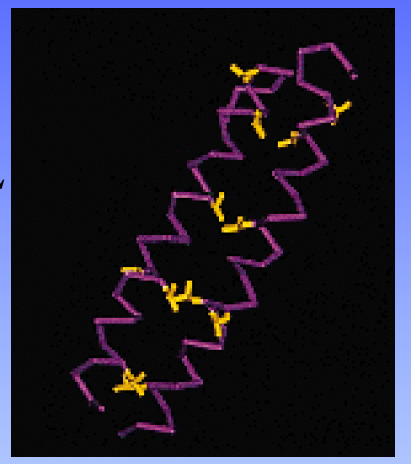
- **SCOP**: Structural Classification Of Proteins
  - based on a definition of structural similarities. Hierarchical levels to reflect evolutionary and structural relationships
  - **http://scop.mrc-lmb.cam.ac.uk/scop**
- **CATH**: Classification by Class, Architecture, Topology, and Homology
  - classified first into hierarchical levels like SCOP
  - **http://www.biochem.ucl.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)
  - **http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml**
- **SARF**
  - categorized on the basis of structural similarity, categories are similar to other dbs
  - **http://123d.ncifcrf.gov/**

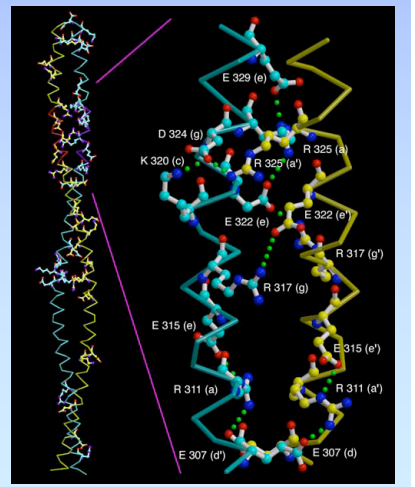
Syllabus

- Structure Coordinates
  - Files & Databases
- Structure Comparisons
  - Aligning 3D Structures
- Structure Classification
  - Structure Families
- **Structure Prediction**
  - **Specialized Structural Regions**
  - **Secondary Structure Prediction**
  - **Tertiary Structure Prediction**
- Structure Visualization

Predicting Specialized Structures

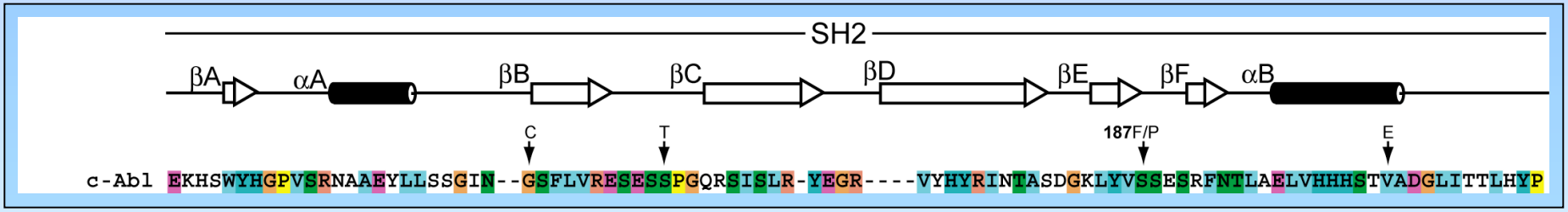
- **Leucine Zippers**
  - Antiparallel α helices held together by interactions between L residues spaced at ever 7th position
  - **2Zip** http://us.expasy.org/tools/
- **Coiled Coils**
  - 2 or three a helices coiled around each other in a left-handed supercoil
  - **Multicoil** http://jura.wi.mit.edu/cgi-bin/multicoil/multicoil.pl
  - **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





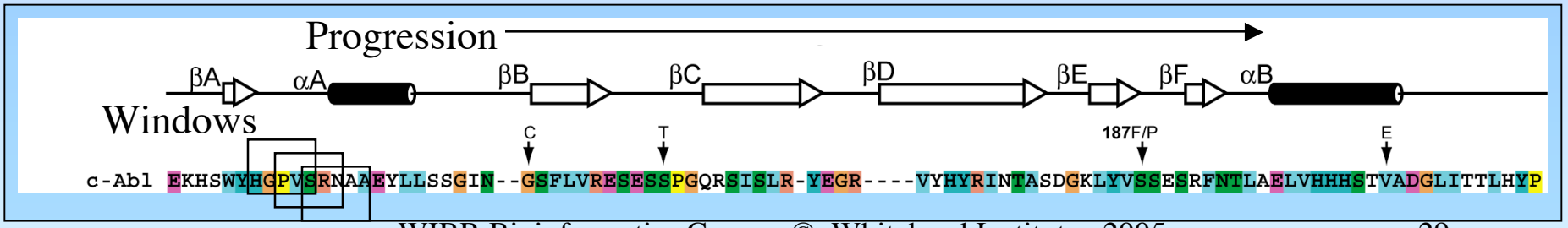
Predicting Secondary Structure

- 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** http://www.cmbi.kun.nl/gv/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 α helices; P,G break helices
- Use a 4-6aa examination window to predict probability of α helix, 3-5aa window for beta strands (as a collection)
  - 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



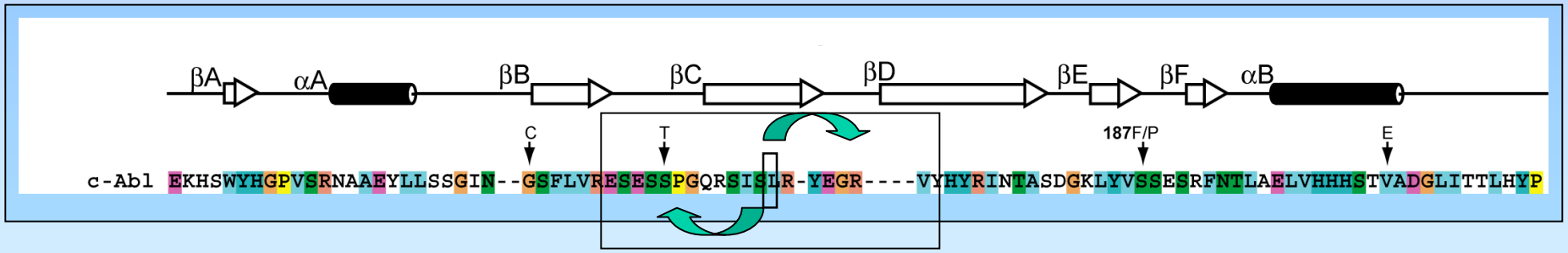
SS Prediction Algorithms Neural Networks

- 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 (segregates sequence)
- 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 (individually)
  - 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

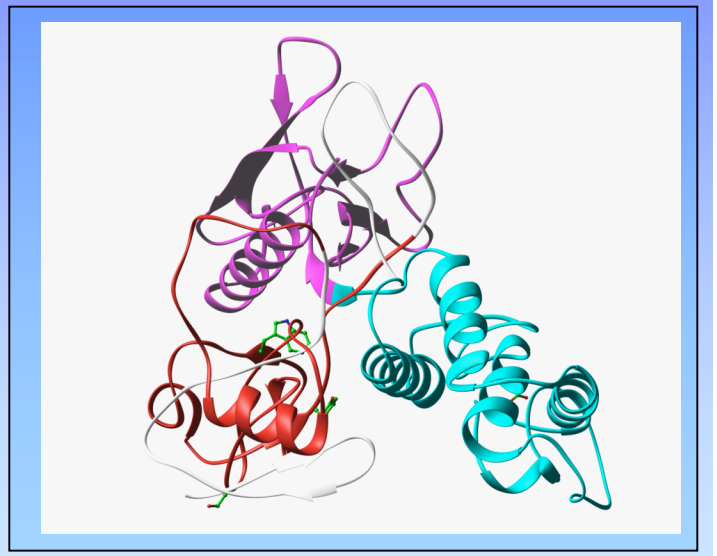


SS Prediction Tools

- **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, predict 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 to use as a **TEMPLATE**

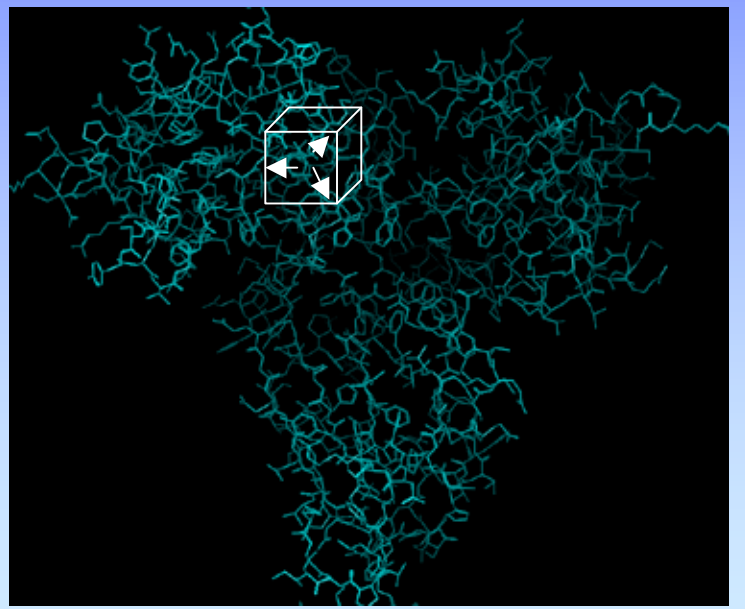


Homology Structure Prediction

- BLAST search PDB sequence database
  - Find structures that have similar sequences to your target protein
- Remember
  - Subtle sequence differences can have a large impact on 3D folding
  - Very different sequences can fold into similar structures!

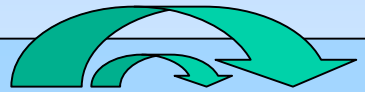
Threading - Approaches

- 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
- **123D** http://123d.ncifcrf.gov/123D+.html



MYNPQGGYQQQFNPQGGRGNYKNFNYNNNLQGYQAGFQPQSQGMSLNDFQKQQKQAAPKPKKTLKLVSSSGIKLANATKK
VGTKPAESDKKEEEKSAETKEPTKEPTKVEEPVKKEEKPVQTEEKTEEKSELPKVEDLKISESTHNTNNANVTSADALIK
EQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLTGSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEER

Model Building

- Perform automated model constructions
  - SWISS-MODEL
    - Compare sequence to ExPdb to find a template (homology)
    - 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

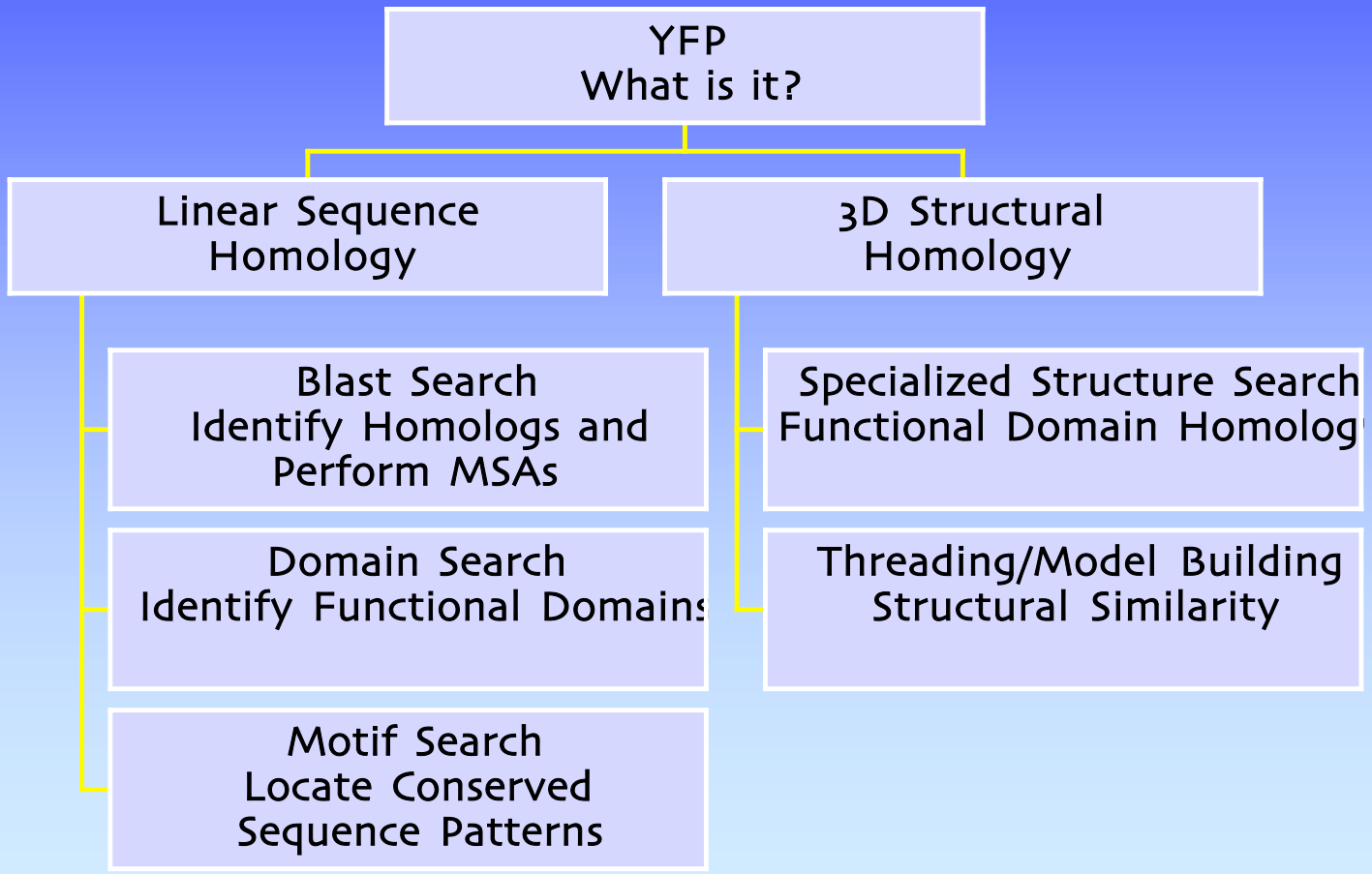
Model Evaluation

- 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)

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/
  - **Cn3D** MMDB viewer (See in 3D) with explicit bonding
    - http://www.ncbi.nlm.nih.gov/Structure
  - **SwissPDB** Viewer (Deep View)
    - http://www.expasy.ch/spdbv/mainpage.html
  - **iMol**
    - http://www.pirx.com/iMol

Pulling It All Together



Structure Resource Examples

- RCSB http://www.rcsb.org/pdb
  - Search for SH2 domain
  - Find coordinates for 1f3j
- MMDB http://www.ncbi.nlm.nih.gov/Structure
  - Search for WD repeat
  - VAST Search
- Dali http://www.ebi.ac.uk/dali
- Prediction
  - Specialized - Multicoil http://jura.wi.mit.edu/cgi-bin/multicoil/multicoil.pl
  - SS (EYA) http://www.compbio.dundee.ac.uk/~www-jpred/submit.html
  - Tertiary (ACRP30) http://123d.ncifcrf.gov/123D+.html
- Model Building
  - Swiss-PDB http://www.expasy.ch/swissmod/SWISS-MODEL.html

Structure Visualization 101

- Deep View Molecular Visualization Tool
  - http://us.expasy.org/spdbv/mainpage.html (it's free!)
  - User friendly interface
  - Analyze several proteins at the same time
  - Structural alignments
  - Amino acid mutations, H-bonds, angles and distances between atoms
  - Integration with Swiss-PDB
  - Reasonable output for figures

Exercises

- RCSB http://www.rcsb.org/pdb
  - Search for Protein Kinase domain
  - Find coordinates for 1iep
- MMDB http://www.ncbi.nlm.nih.gov/Structure
  - Search for telomerase structures
- Dali http://www.ebi.ac.uk/dali
  - Align 2 structures
  - Search for similar structures
- Prediction
  - SS http://www.compbio.dundee.ac.uk/~www-jpred/submit.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.