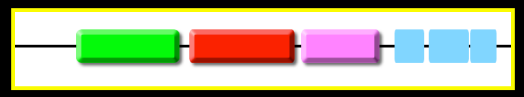
Protein Bioinformatics

Building Useful Relationships
Between Multiple Protein
Sequences and Structures

MSRKGPRAEVCADCSA
PDPGWASISRGVLVCDE
CCSVHRLGRHISIVKHLR
HSAWPPTLLQMVHTLA
SNGANSIWEHSLLDPA
QVQSGRRKAN



Robert Latek, Ph.D.
Bioinformatics Scientist, Biocomputing

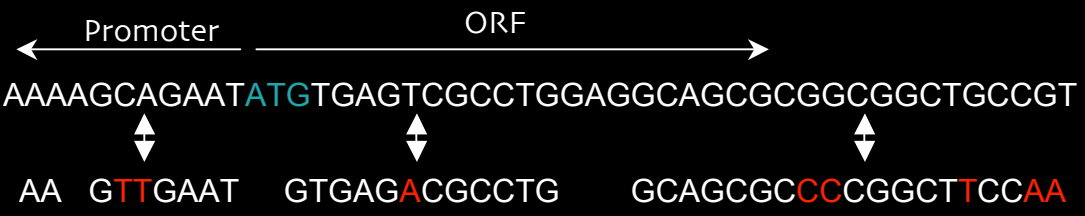
WIBR Biocomputing, © Whitehead Institute, 2002

Overview

- Introduction
- Sequence Comparisons
  - Pairwise Letter by Letter vs. Word by Word
  - Multiple Sequence Alignments (MSA)
- Sequence Domains
  - Phylogenetic Trees and Multiple Sequence Alignments
  - Blocks, Patterns, and Profiles (Motifs)
- Structural Comparisons
  - Analyzing Protein Structures
  - Structure Prediction

Translated Transition

- DNA Bioinformatics
  - Sequence Comparison, Gene & Promoter ID



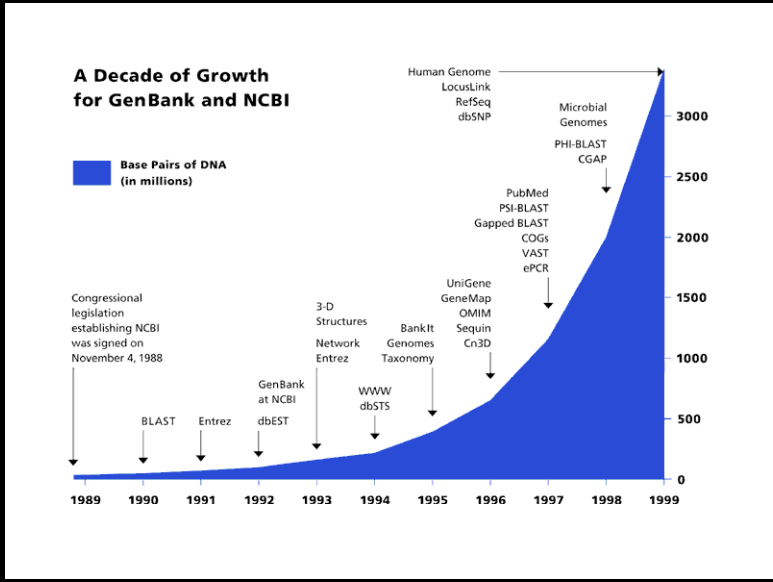
- Protein Bioinformatics
  - Sequence, Function, Structure ID & Comparison

WIBR Biocomputing, © Whitehead Institute, 2002

Bioinformatics Databases

Sequence Database Growth

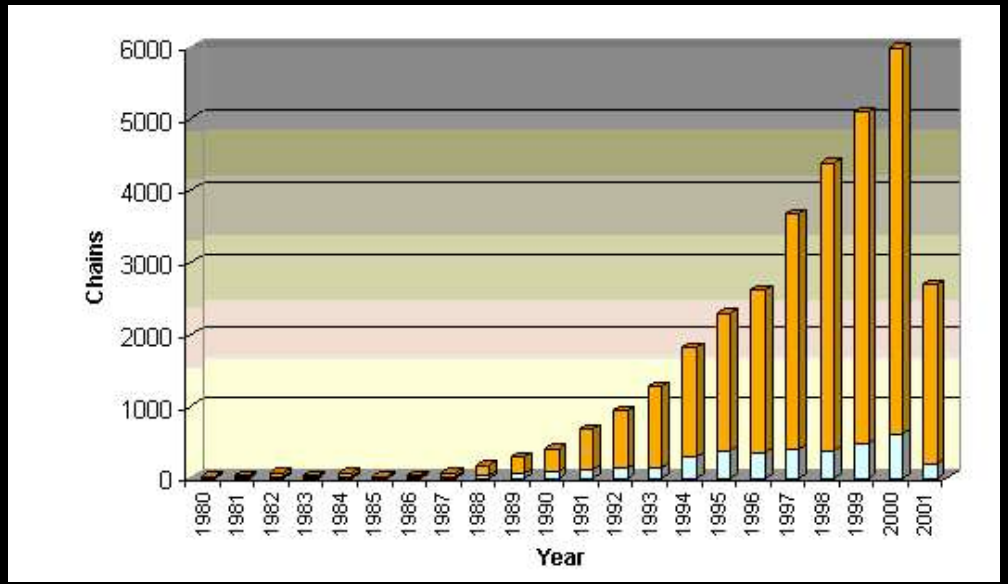
Sequences / NT ( )



(Cooper NCBI)

Protein Structure Database Expansion

Sequences / AA ( )



(PDB)

Bioinformatics Resources

aBi
Atelier BioInformatique
Université Aix-Marseille I

On line analysis tools

er for Biotechnology Information
f Medicine National Institutes of Health
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search Nucleotide for Go

e! project Ensembl
The Wellcome Trust Sanger Institute EBI

SWISS-MODEL

ne Database

Pfam :: HMM Search
Analyze a query sequence using the Pfam HMM d

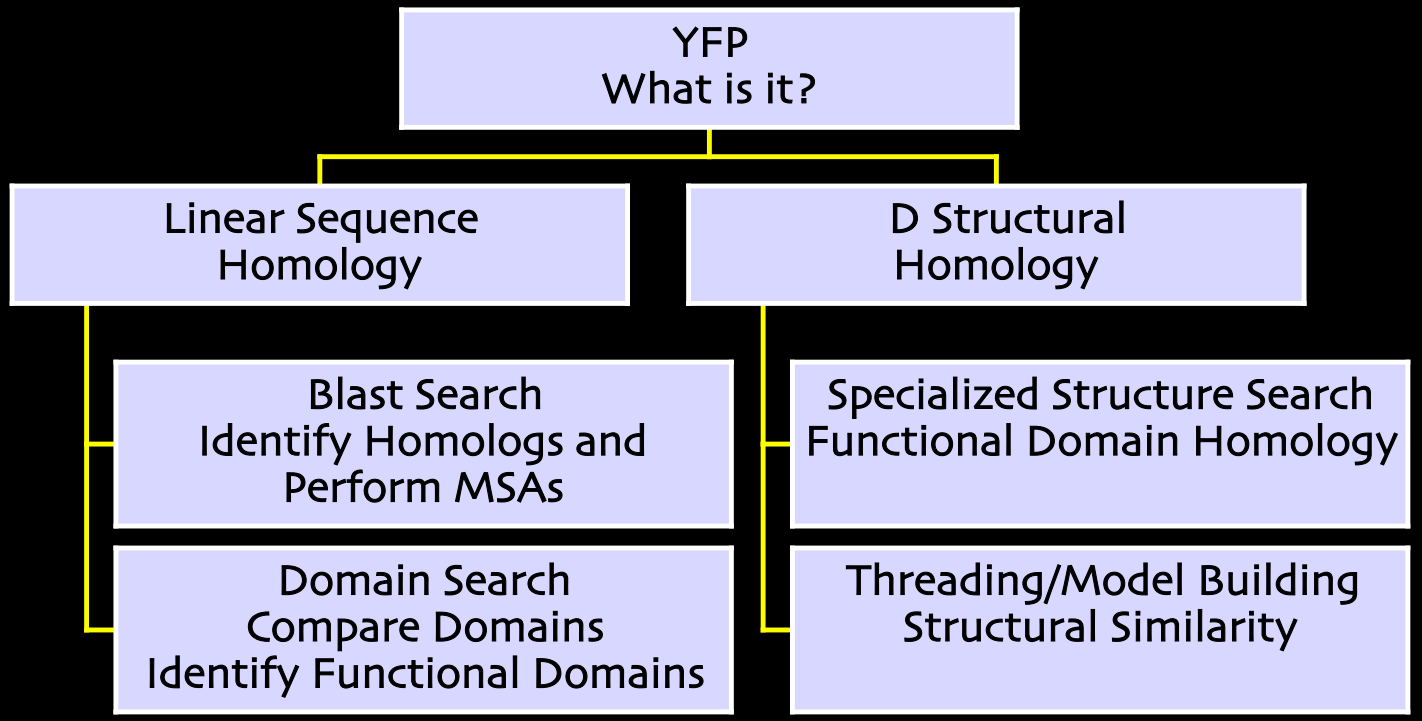
The Protein Domain Datab
ProDor

SWISS-PROT
Protein knowledgebase
TrEMBL
SIB swissprot

SS-PROT

Pole Bio-Informatique Lyonnais
Network Protein Sequence @nalysis
NPS@ is the IBCP contribution to PBIL in Lyon, France

Bioinformatics Approaches

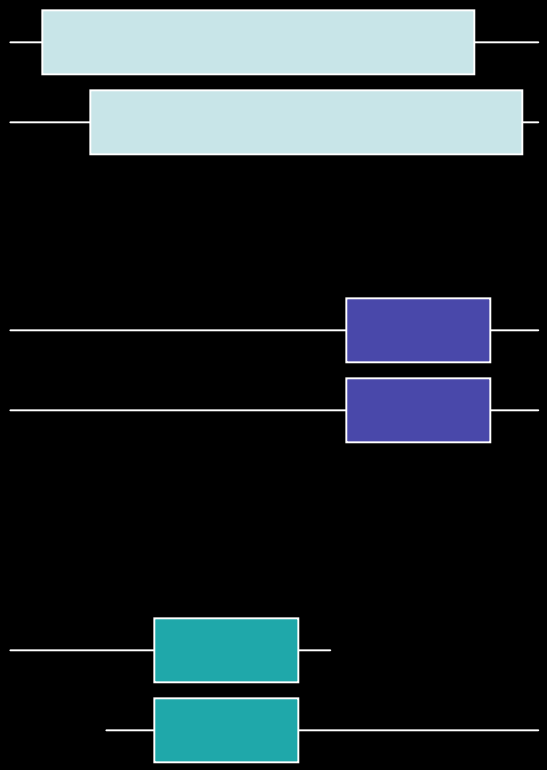


Overview

- Introduction
- Sequence Comparisons
  - Letter by Letter vs. Word by Word
  - Global Multiple Sequence Comparisons
- Local Sequence Domains
- Structural Domains

Protein Modules

- Proteins are derived from a limited number of basic building blocks (Domains)
- Evolution has shuffled these modules giving rise to a diverse repertoire of protein sequences
- As a result, proteins can share a global or local relationship



(Higgins)

Residue by Residue Comparisons

A: V R W A H S G C T R T

B: V R G A H K D R T

GUUAGAUGGGCUCAUAGUGGCUGCACUCGGACC
GUUAGAGGUGCUCATAAAGAUCGGACC

I. V R W A H S G C T R T

V R G A H – – K D R T

GUU AGA UGG GCU CAU AGU GGC UGC ACU CGG ACC
GUU AGA GGU GCU CAU AAA GAU CGG ACC

II. V R W A H S G C T R T

V R G A H – K D – R T

III. V R W A H S G C T R T

V R G A H – K – D R T

Word by Word Comparisons

Query: **GTQITVEDLFYNIATRRKALKN**

GTQ
Word Size = 3
TQI
QIT
ITV -> **LTV,MTV,ISV,LSV,MSV**
TVE **IAV,LAV,MAV,ITL,etc.**
VED
EDL
DLF

Adjustable

2 or 3 for protein (3 default)

> 7 for blastn searches (11 default)

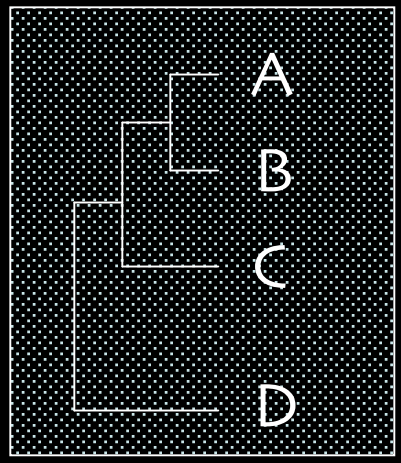
Neighborhood Words

Make table for both query and database

(Cooper NCBI)

Multiple Sequence Comparisons

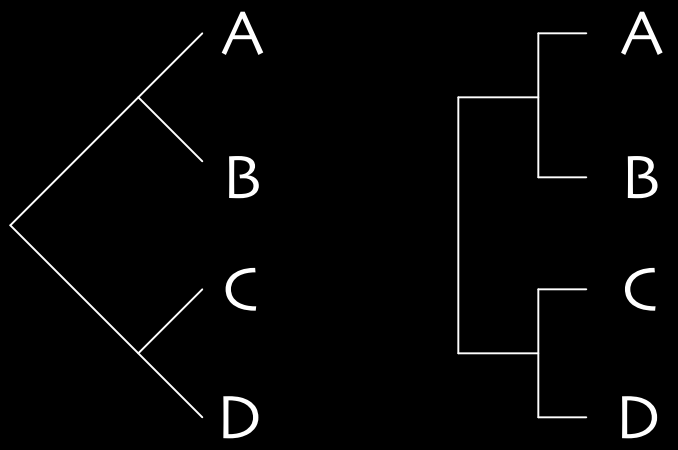
- Phylogenetic Trees
  - Graphs representing the evolutionary history of a gene
  - Relationship of one sequence to other sequences
  - Dissect the appearance order of insertions, deletions, and mutations
- Multiple Sequence Alignments
  - Place residues in columns that are derived from a common ancestral residue



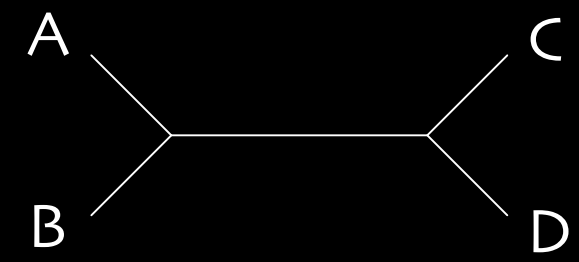

```
A:V W A H S C R
B:V G A H - K R
C:V G A H S K R
D:V G A H - C R
```

Phylogenetic Trees

Rooted



Unrooted



Branches intersect at Nodes

Predict function, observe epidemiology, analyzing changes in viral strains

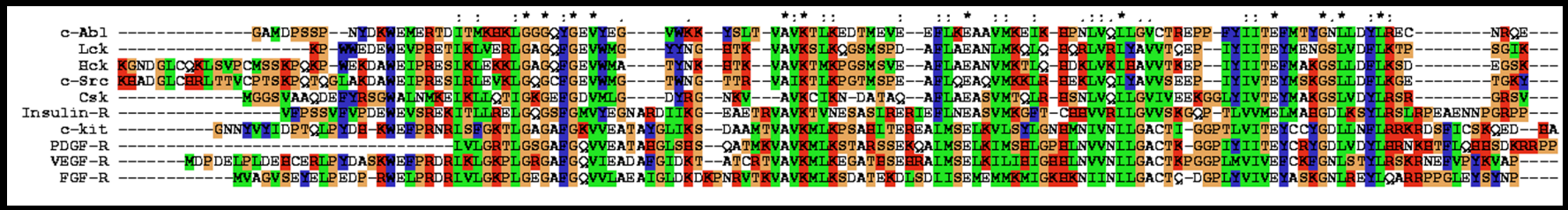
Tree Building

- Building and Drawing Trees

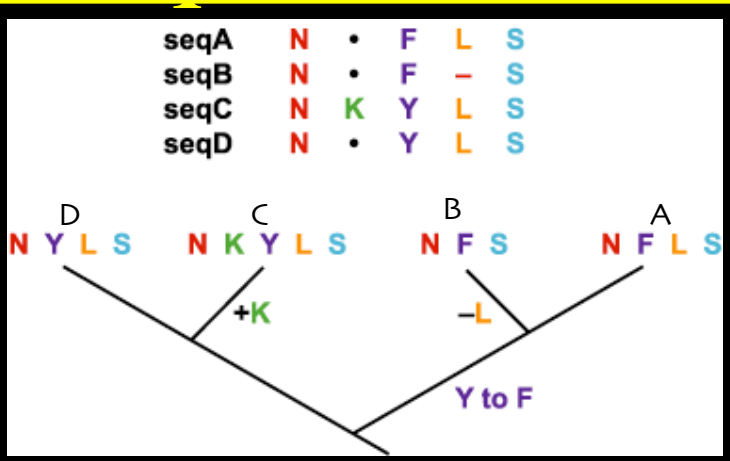
  http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html

Multiple Sequence Alignments

- Global
  - Search for alignments, matching over entire sequences
- Local
  - Examine regions of sequence for conserved segments
- Matches, Mismatches, Gaps



Multiple Sequence Comparisons



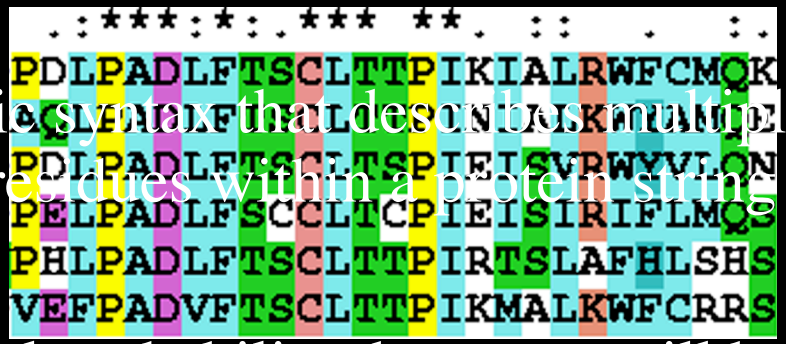
- MSCs Can Reveal Sequence Patterns
  - Demonstration of homology between >2 sequences
  - Identification of functionally important sites & structure prediction
- ClustalX
  - Web-based http://pir.georgetown.edu/pirwww/search/multaln.html
  - Stand-alone
- MSAs Are Related To Trees

Overview

- Introduction
- Sequence Comparisons
- Local Sequence Domains
- Structural Domains

Local Alignments: Domains of Conserved Sequence

- Blocks
  - Represent a conserved region within a MSA that contains matches & mismatches but no gaps
  - Serve as anchors to assist in aligning sequences
  - http://blocks.fhcrc.org/
- Patterns
  - Deterministic syntax that describes multiple combinations of possible residues within a protein string
- Profiles
  - Mathematical probability that an aa will be located at a given position



Pattern Matching

- Find patterns like aa1 xx aa2 xxxx aa3
  - Definition of a non-contiguous motif

```
C Y D     C A F T L R Q S A V M H K H A R E H
C A T Y   C R T A I D T V K N S L K H H S A H
C W D G G C G I S V E R M D T V H K H D T V H
C Y C     C S D H M K K D A V E R M H K K D H
C N M F   C M P I F R Q N S L A R E H E R M H
C L N N T C T A F W R Q K K D D T V H N S L H
```

C x(2,4) C x(3) [LIVMFW] x(8) H x(3,5) H

Define/Search A Motif http://us.expasy.org/tools/scanprosite/

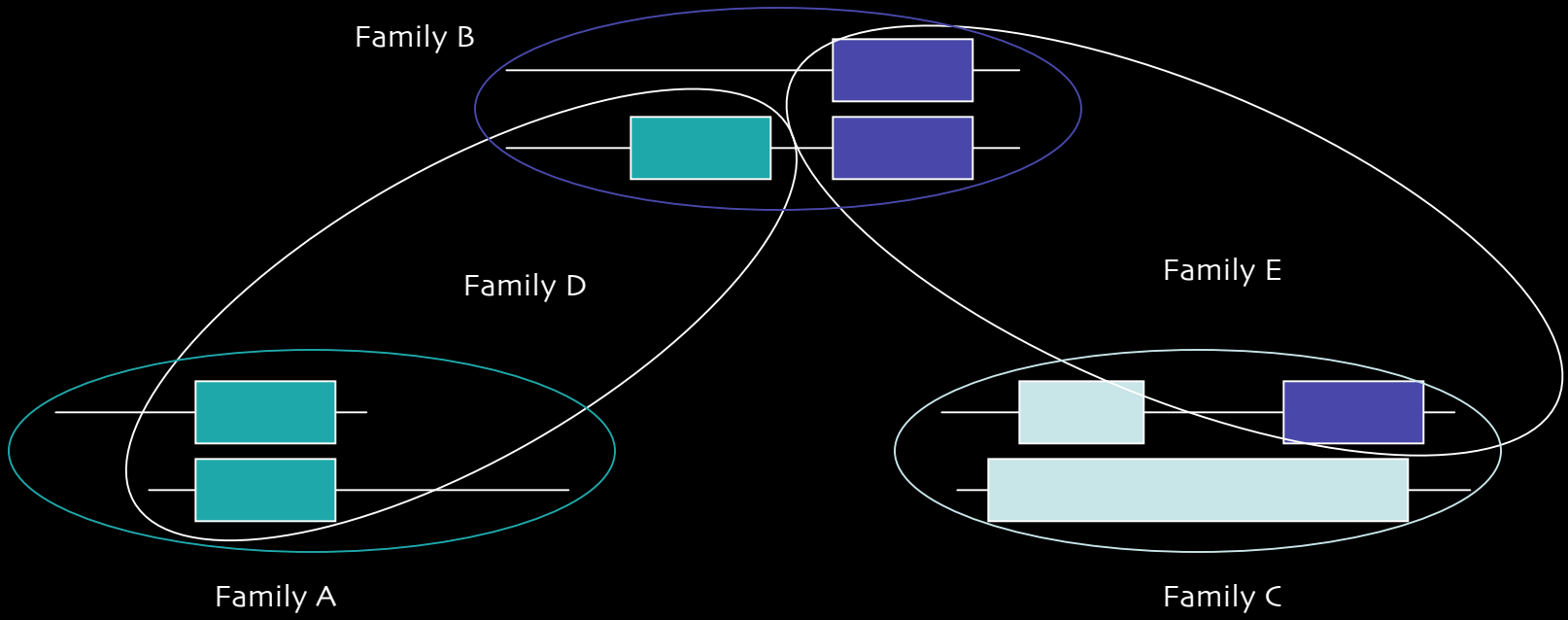
Profile Searching

- PSSM - Position Specific Scoring Matrix
  - –Columns of weights for every aa corresponding to each column of a MSA

http://www.ncbi.nlm.nih.gov/BLAST

CON	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
I	8	-2	5	4	5	5	-4	24	0	15	13	1	1	1	-7	2	22	21	-18	-6
T	13	-5	24	18	-18	19	7	1	7	E	-4	14	11	10	-1	9	29	3	-28	-14
L	5	-5	3	4	13	4	2	8	-4	14	12	8	-5	0	-10	0	10	10	-1	5
S	17	17	13	10	-12	29	-5	-5	6	-14	-9	12	10	0	-2	34	19	1	-8	-15

Protein Families



- Protein Family - a group of proteins that share a common function and/or structure, that are potentially derived from a common ancestor (set of homologous proteins)

Family Databases Resources

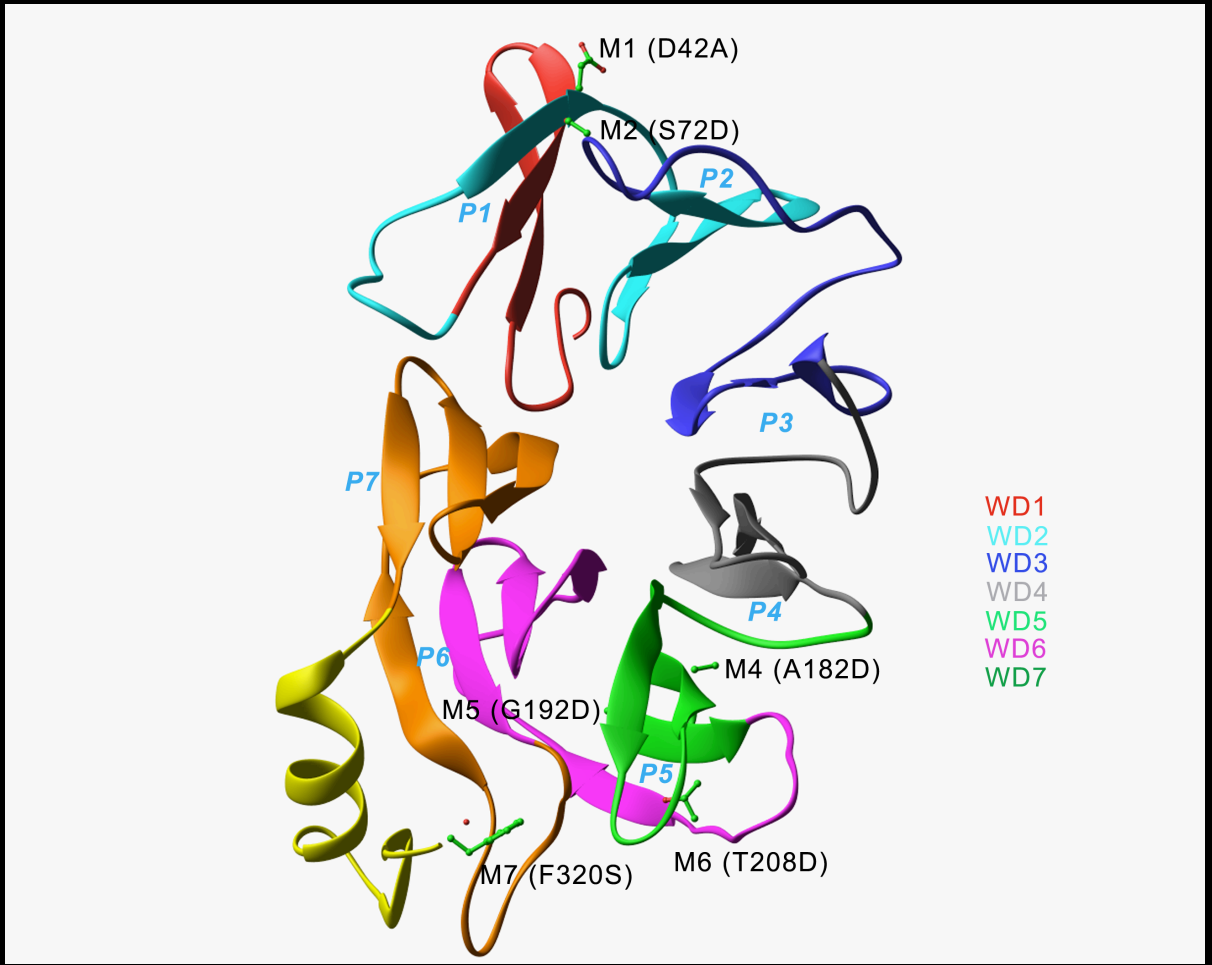
- Curated Databases
  - Proteins are placed into families with which they share a specific sequence pattern
  - Pfam http://pfam.wustl.edu/hmmsearch.shtml/
- Clustering Databases
  - Sequence similarity-based without the prior knowledge of a specific patterns by searching a database against itself
  - ProDom http://prodes.toulouse.inra.fr/prodom/doc/prodom.html
- Derived Databases
  - Pool other databases into one central resource
  - Blocks - Search and Make http://blocks.fhcrc.org/blocks/

Overview

- Introduction
- Sequence Comparisons
- Local Sequence Domains
- Structural Domains
  - Structural Comparisons (Alignment, Substitution Modeling, Domain Identification, Residue Mapping)
  - Structure Prediction

Protein Structure Classification

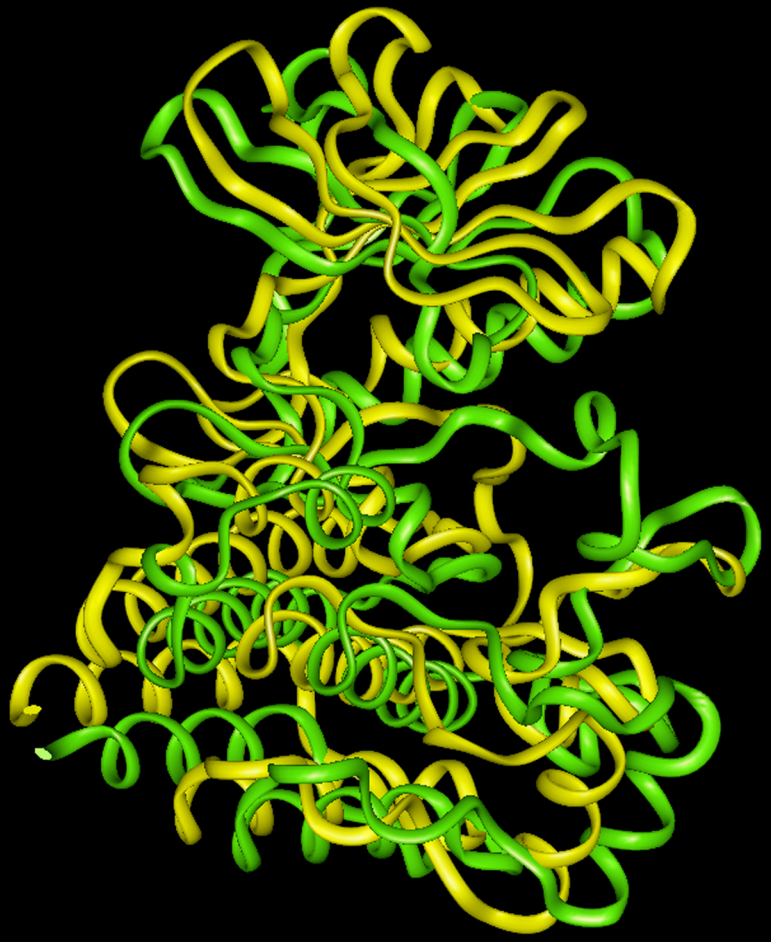
- Proteins can adopt only a limited number of possible 3D conformations
- Protein Structure Classes
- Completely different sequences can fold into similar shapes



Structure Family 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/

Structural Alignments



Superimpose Structures

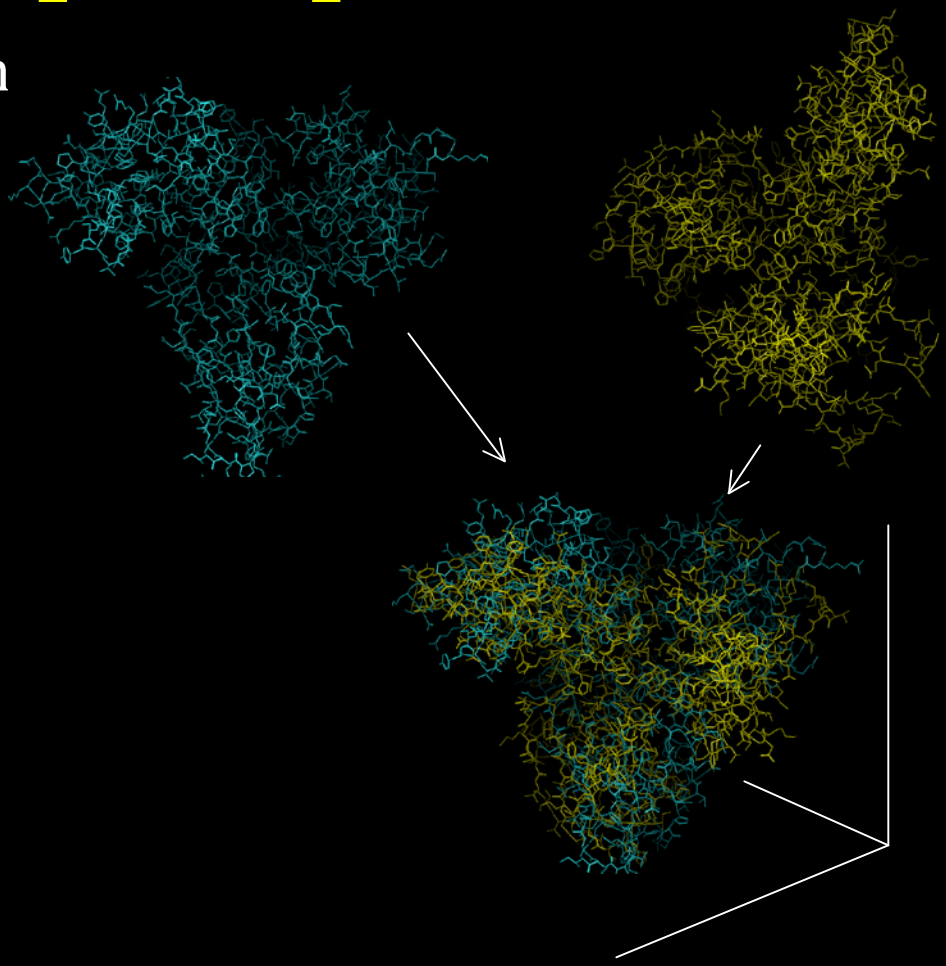
ıdent
. holo

istory,

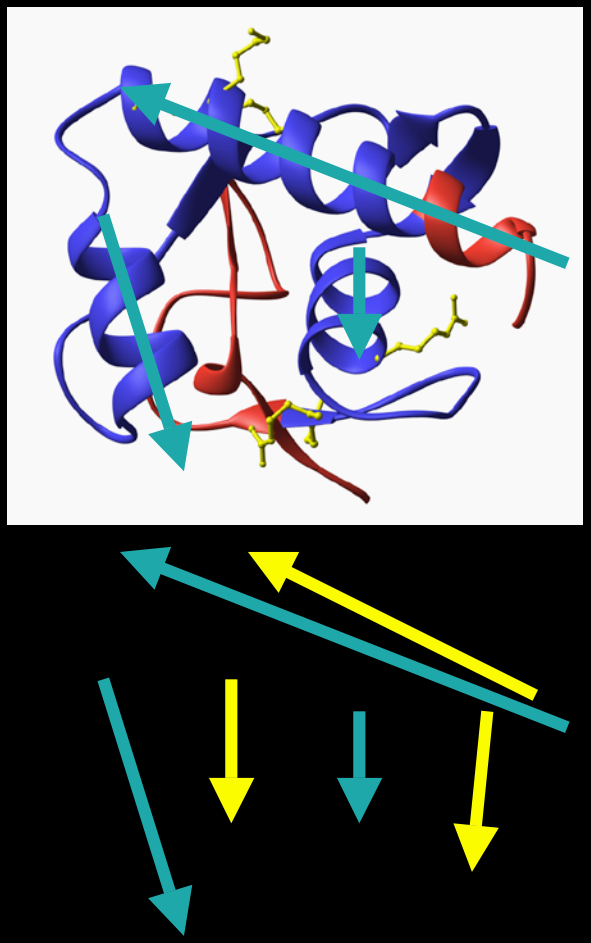
ı protein
ıll
ligand

Structure Superimposition

- Sequence alignment based on LINEAR sequence similarity
- Structure alignment forms relationships in 3D space
  - similarity can be redundant for multiple sequences
- Alignment
  - Translate center of mass to a common origin
  - Rotate to find a suitable superposition
- Considerations
  - What to align and compare?



VAST and SARF



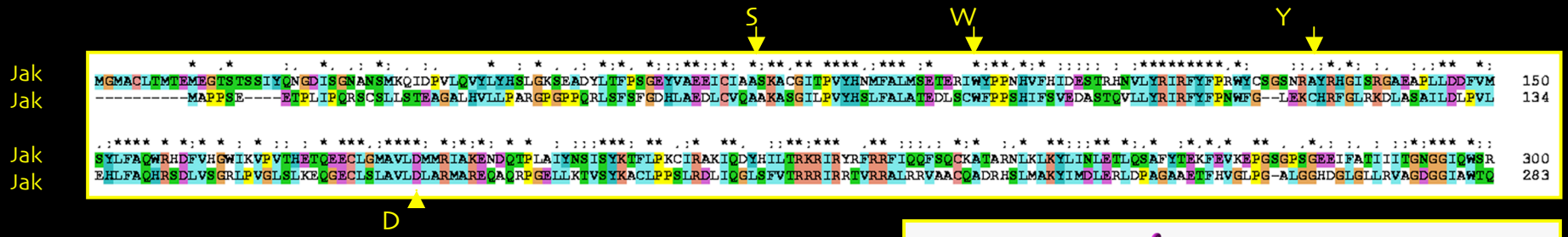
- Implement rapid methods to assign secondary structure
- VAST

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

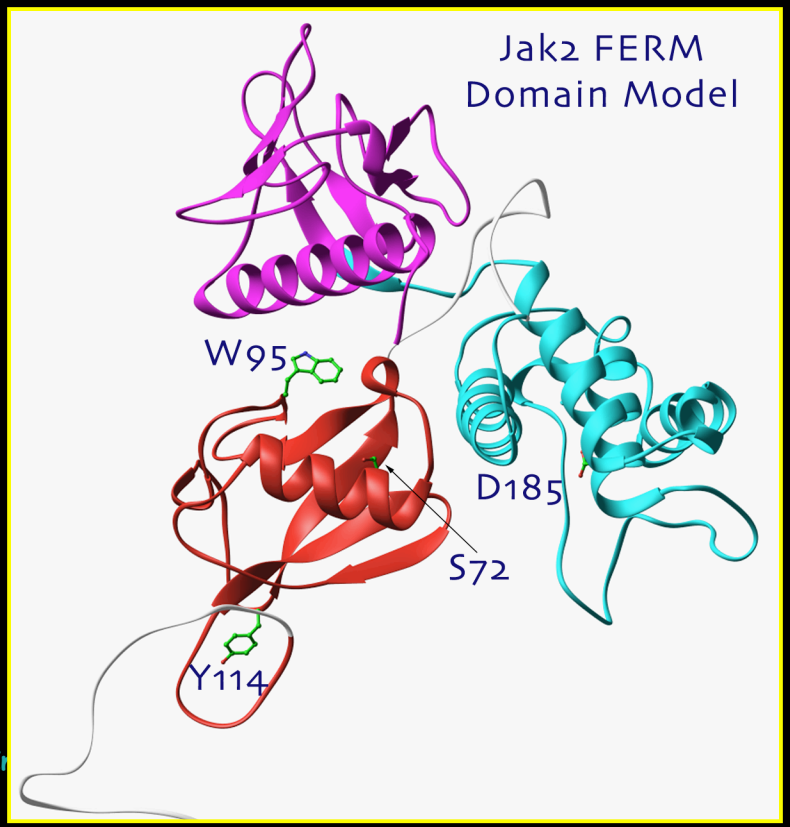
- SARF

http://123d.ncifcrf.gov/

Substitution Modeling



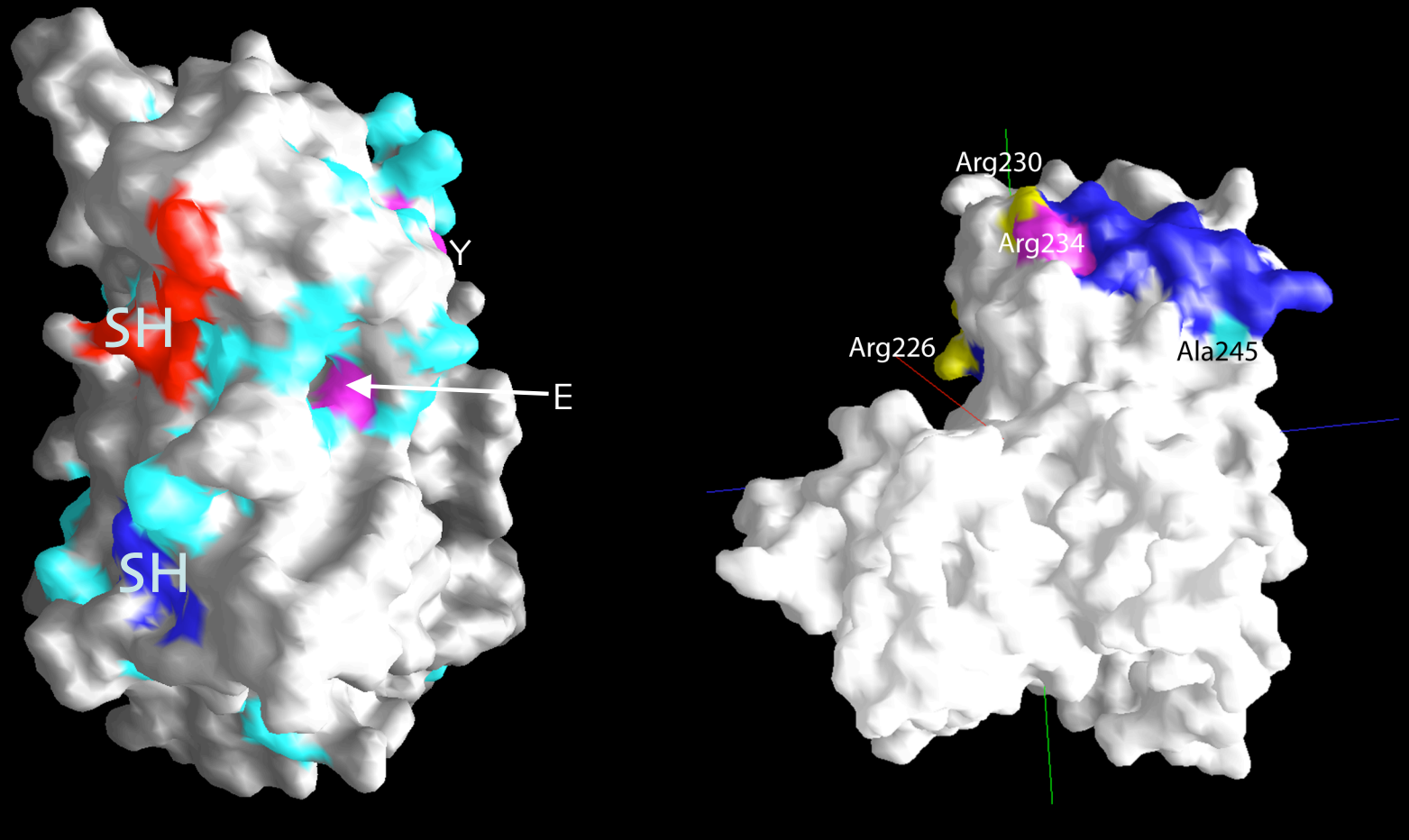
- Mapping Variants
- Construct Design
- Ligand Associations



(L Huang & R Latek)

WIBR Biocomputing, © Whitehead I

Surface Mapping



WIBR Biocomputing, © Whitehead Institute, 2002

Structure Prediction

- Secondary Structure Prediction
- Specialized Structures
- Tertiary Structural Modeling

Secondary Structure Prediction

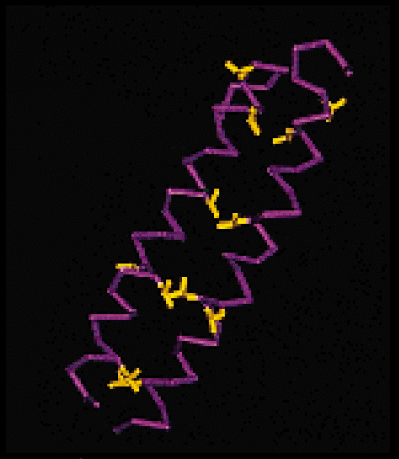
- Recognizing Potential Secondary Structure
  - 50% of a sequence is usually alpha helices and beta sheet structures
  - Helices: 3.6 residues/turn, N+4 bonding
  - Strands: extended conformation, interactions between strands, disrupted by beta bulges
  - Coils: A,G,S,T,P are predominant
  - Sequences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

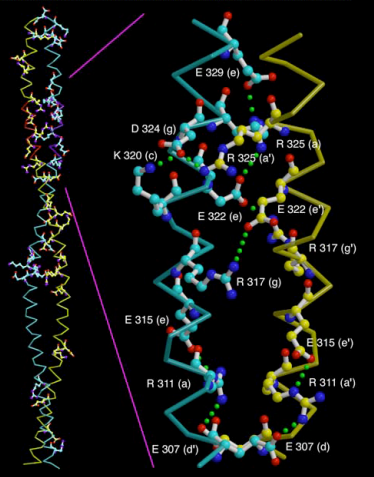
Secondary Structure Prediction Tools

- 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/

Specialized Structures

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





Tertiary Structural Modeling

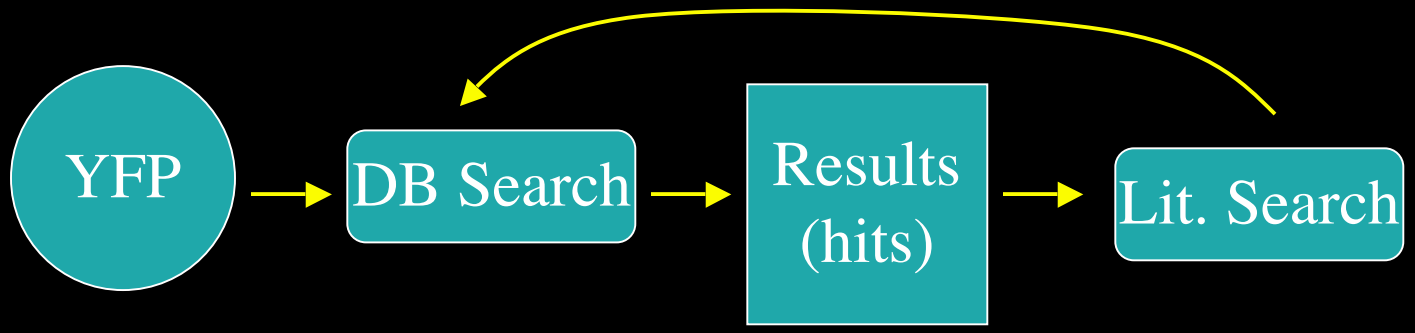
- Goal
  - Build a model to use for comparison with other structures, identify important residues/interactions, determine function
- Challenges
  - Reveal interactions that occur between residues that are distant from each other in a linear sequence
  - Slight changes in local structure can have large 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

Structural Modeling Tools

- 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
- Threading

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

Post-Processing



- Literature Searches
- Do the Results Make Sense?
- Are the Results Consistent?
- Can the Results be Substantiated Biologically?

Questions?

Robert Latek
latek@wi mit edu

PDF Slides Available At:
http://staffa wi mit edu/bio/proteins pdf

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 Ouellette. Wiley Interscience, 2001.

Brooks, Charles L., Department of Molecular Biology. The Scripps Research Institute.

Burkhard P, Kammerer RA, Steinmetz MO, Bourenkov GP, Aebi U. The coiled-coil trigger site of the rod domain of cortexillin I unveils a distinct network of interhelical and intrahelical salt bridges. Structure Fold Des 2000 Mar 15;8(3):223-30
Oxford University Press, 2000.

Bing Hao, Weimin Gong, Tsuneo K. Ferguson, Carey M. James, Joseph A. Krzycki, and Michael K. Chan. A New UAG-Encoded Residue in the Structure of a Methanogen Methyltransferase. *Science* 2002 May 24; 296: 1462-1466.

Gayathri Srinivasan, Carey M. James, and Joseph A. Krzycki**.** Pyrrolysine Encoded by UAG in Archaea: Charging of a UAG-Decoding Specialized tRNA. *Science* 2002 May 24; 296: 1459-1462.