Animation using pdb2mgif

Homology Modeling: Concepts and Protocols

Sarangan *Ravi*chandran, Ph.D.
Advanced Biomedical Computing Center
NCI-Frederick, Frederick, MD 21702

http://ncisgi.ncifcrf.gov/~ravichas/HomMod
05/23/2005

Advanced Biomedical Computing Center

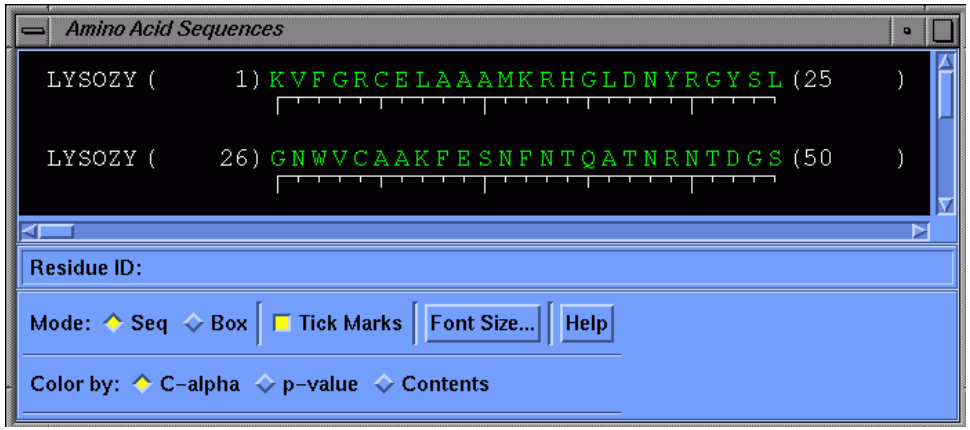
- Web: http://www.abcc.ncifcrf.gov
- Scientific Applications Page
- New Account
- Contacting the Help Desk

Overview

- Basics of Homology Modeling
- Hands-on exercise
  - Homology Modeling using InsightII
- What I will not talk about!
  - Alternatives to Comparative (homology) modeling
  - Basics of protein structure (primary, secondary…)
  - Theory behind sequence alignment (pair-wise and Multiple) and scoring matrices
  - Theory behind InsightII (homology) modules

Overview of Homology Modeling

Raw Sequence

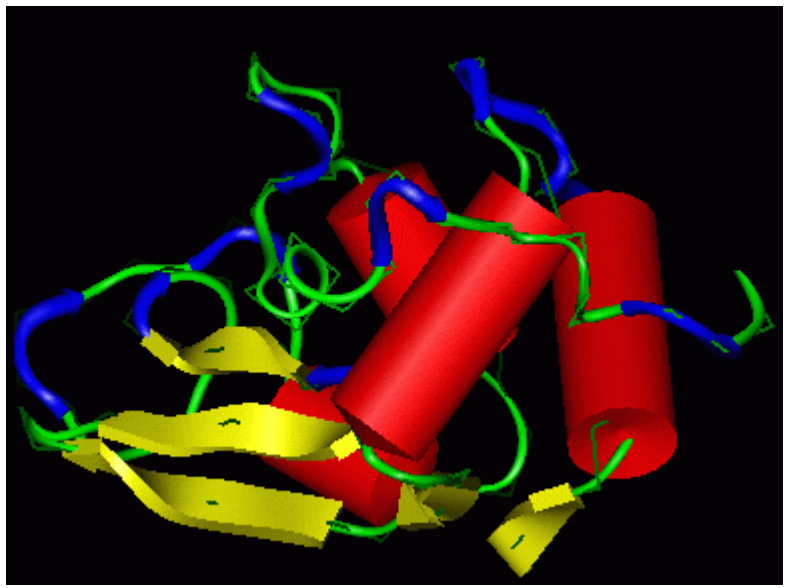


Experiments

X-ray, NMR, e-Diffraction

Physicochemical Simulations

Comparative Modeling Knowledge-Based



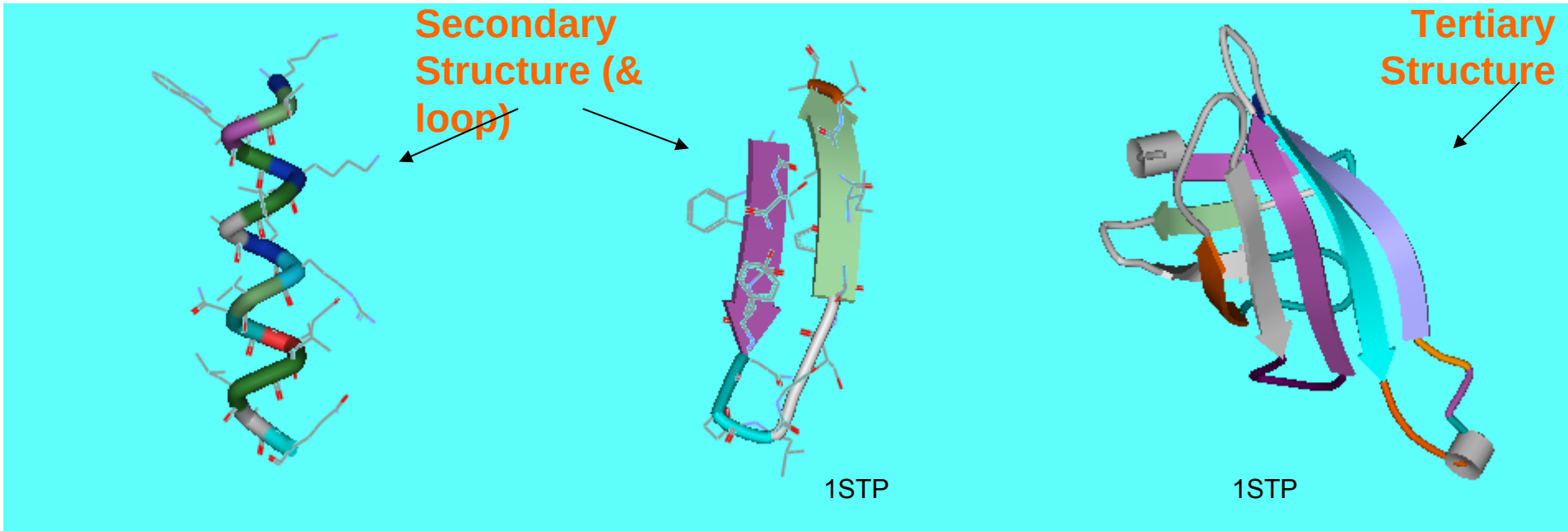
Modeling

3D-Structure

Accelrys: Homology

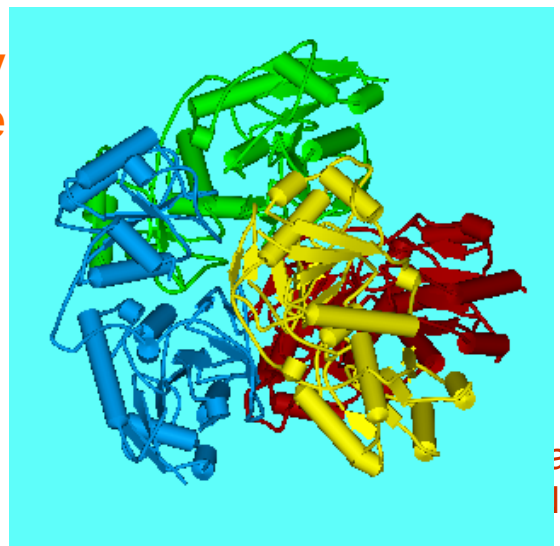
MAPRRVRSFL RGLPALLLLL LFLGPWPAAS HGGKYSREKN

Primary



Chunk From 1WSA

Quarternary Structure

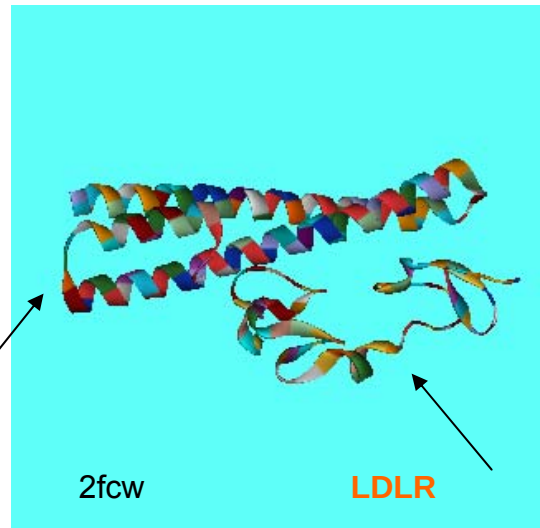


1WSA (A,B,C,D)
06/14/06

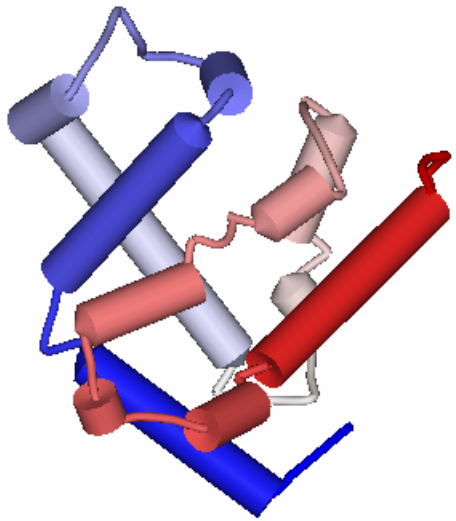
Quinternary (Protein-Protein/Protein-DNA)

RAP

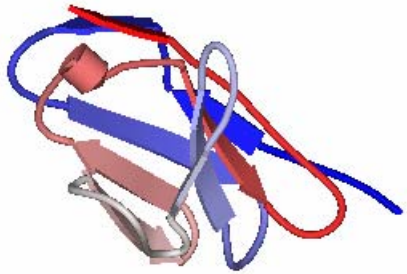
andran, Ph.D., ABCC,
ICI-Frederick



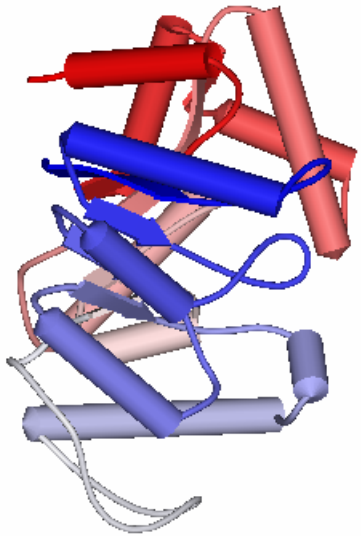
All Alpha: 2mm1
Myoglobin Mutant



All Beta: 1bww_b
BENCE-JONES
IMMUNOGLOBULIN REI
VARIABLE PORTION



Alpha+Beta: 1luc_b
Bacterial Luciferace



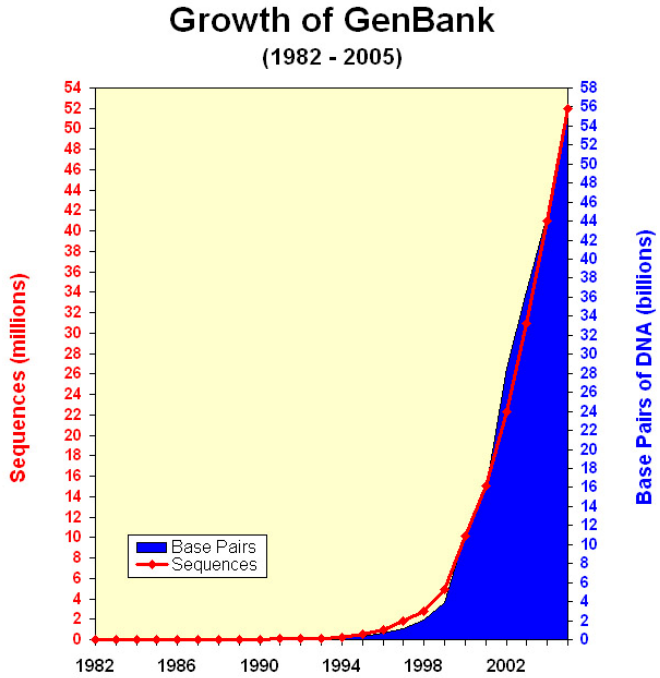
α/β(mainly parallel β with intervening α sheets

α + β (segregated α & anti-parallel β)

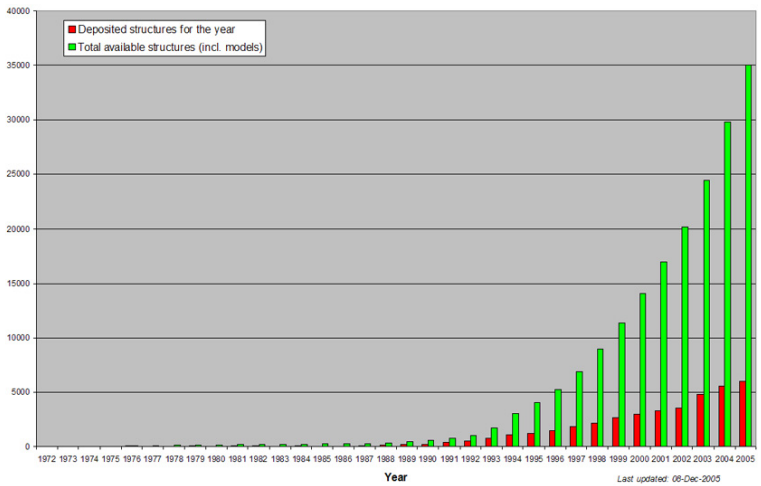
Multi-domain (α & β)

Why Homology Modeling?

- Rate of structure solving through NMR or X-ray is slow compared to the deposition of DNA and Protein sequences
- Crystallization is the bottle-neck (time in months). No generic recipe for crystallization
  - UniProtKB/SP as of 06/01/06 222,289 entries
    - (Release 42.4 had 138,347 entries 11-14-03)
  - PDB as of 05/30/06 has 36,837 structures
    - 23,188 (11/11/03)
    - 28,110 (05/10/05)



NCBI



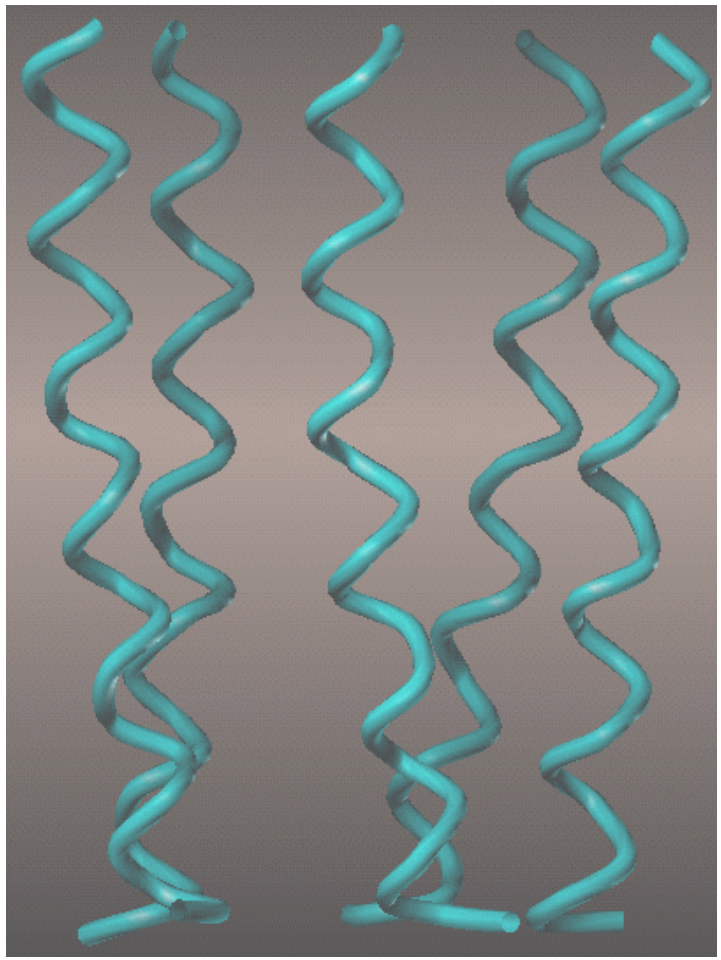
Why Homology Modeling?

- Membrane proteins are difficult to crystallize
  - 30% of proteome of living things
- Knowledge of 3D structure is essential for the understanding of the protein function
  - 3D structure shows us far apart AAs can be closer in the structure
- Structural information enhances our understanding of protein-protein or protein-DNA interactions

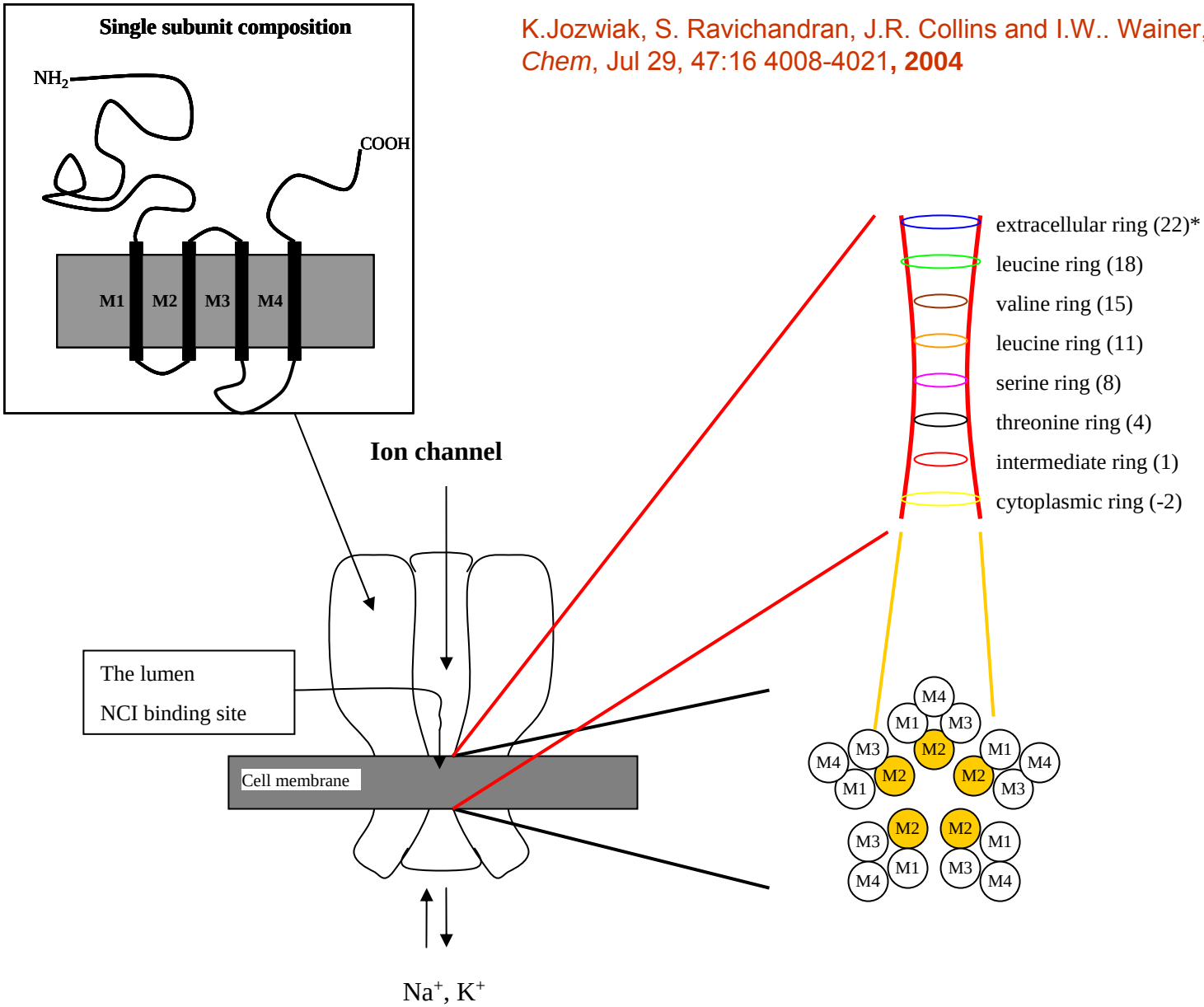
K.Jozwiak, S. Ravichandran, J.R. Collins and I.W.. Wainer, *J. Med. Chem*, Jul 29, 47:16 4008-4021**, 2004**

Applications of Homology Modeling

- Potassium Channel proteins
  - Trans-membrane region-no 3D structure available
  - Used Homology Modeling to build a model for the channel protein
  - Used QSAR to model the binding of inhibitors
  - Docking to study the drug-receptor interaction



K.Jozwiak, S. Ravichandran, J.R. Collins and I.W.. Wainer, *J. Med. Chem*, Jul 29, 47:16 4008-4021**, 2004**



Homologous Proteins

- Homologous Proteins (Orthologous: Same Function on different species or Paralogous: Related but different function on same species):
  - "Having a common evolutionary origin"
    - Evolved (mutations etc.) evolutionarily from a common ancestor
- Many of the essential proteins (key regulators) present in humans are also present in other living organisms (eg. Rat, bacteria)

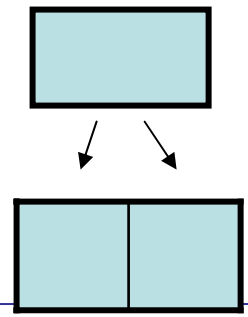
- These essential proteins have to conserve their functionality throughout evolution
  - DNA polymerases
    - DNA replication
      - Necessary for all organisms
  - MHC Major Histocompatibility Complex
    - Antigen presentation to trigger an immune response
      - Present in higher Eukaryotes, rats and humans

How to identify homologous proteins? Can we exploit sequence information?

Homologous Proteins

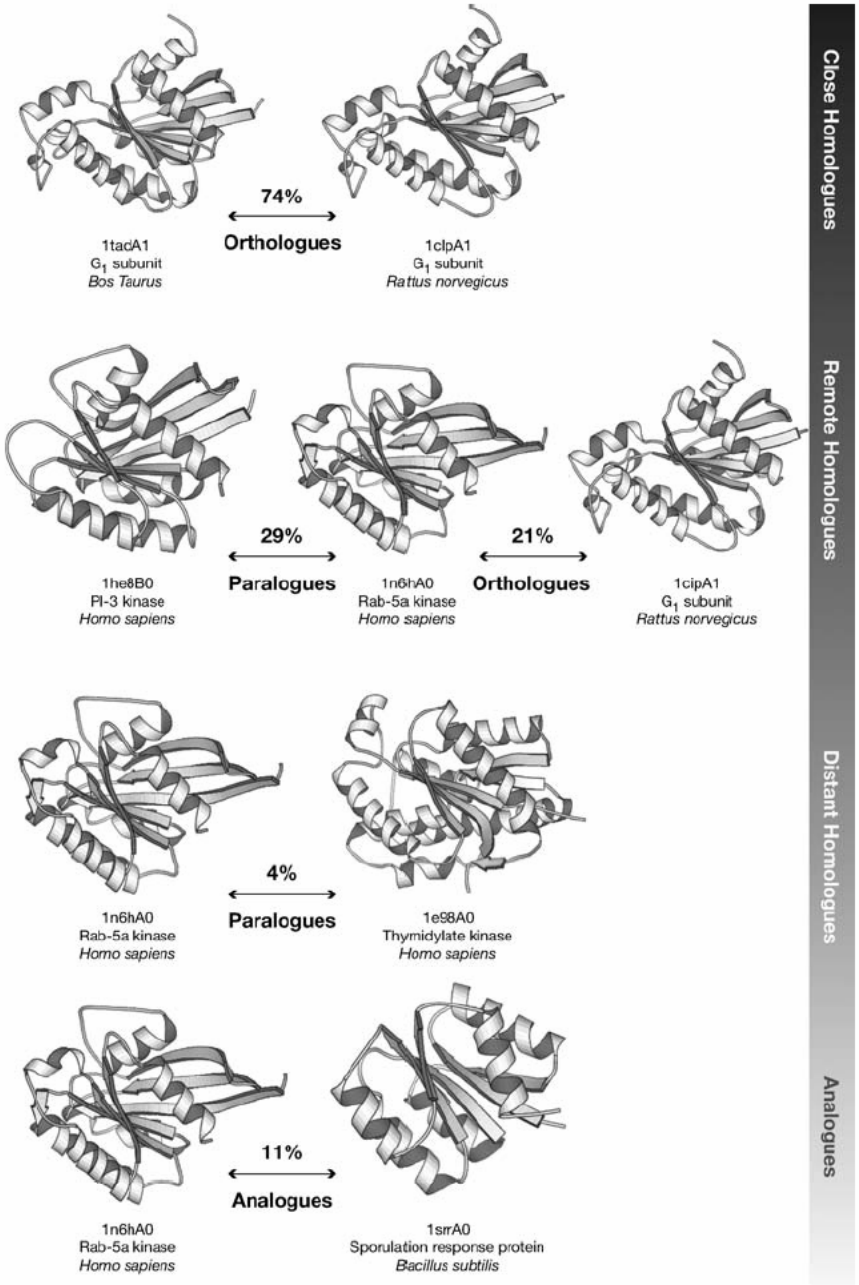
- Orthologous (different Species)
  - ortho (exact)
  - Result of speciation
  - Example:
  - Carrier Proteins (Lipocalin)-Transport Vitamin A in Serum
    - RBP Gene (human)
    - RBP Gene (Rat)

- Paralogous (same Species)
  - para(parallel)
  - Result of Gene Duplication
  - Example:
    - alpha,beta hemoglobins



Homology and Convergence

- Homologous proteins: Proteins that are related by divergence (not convergence) from a common ancestor
  - Convergence (nature reinvented them for different purposes)
  - Similar Fold (analogues) but no significant sequence similarity (same fold, different function, ancestrol origin unknown)
    - β-barrel (Soluble serine proteases and integral membrane porins) [can be the result of stable architecture]
  - Share catalytic site (analogues) but no sequence/structure similarity
    - His-Asp-Ser triads seen in Subtilisin (αβα sandwich) and Chymotrypsinogen (2 domain β-barrel)



Protein Families and their evolution-A structural perspective, C.A. Orengo and J.M. Thornton, Annu. Rev. Biochem, **2005**, 74, 867-900

Sequence Dissimilarity & Structural Similarity

What we already know about homologous proteins

- Core region is pretty much conserved (main secondary structural features)
- Most dissimilarity is observed in the surface (loop) regions
- Within homologous proteins secondary-structures can move relative to each other or even disappear but neither order nor orientation will differ (α becoming β etc.)
- Sequence similarity is less conserved compared to Structural similarity
  - Far diverged proteins has very little sequence similarity

ClustalW 1.83
www.ebi.ac.uk

```
1OG5_B|PDBID|CHAIN|SEQUENCE      MAKKTSSKGR-PPGPTPLPVIGNILQIGIKDISKSLTNLSKVYGPVFTLY 49
1NR6_A|PDBID|CHAIN|SEQUENCE      MAKKTSSKGKLPPGPTPFPIIGNILQIDAKDISKSLTKFSECYGPVFTVY 50
                                 **********: ******:*:*******. ********::*: ******:*

1OG5_B|PDBID|CHAIN|SEQUENCE      FGLKPIVVLHGYEAVKEALIDLGEEFSGRGIFPLAERANRGFGIVFSNGK 99
1NR6_A|PDBID|CHAIN|SEQUENCE      LGMKPTVVLHGYEAVKEALVDLGEEFAGRGSVPILEKVSKGLGIAFSNAK 100
                                 :*:** **************:******:*** .*: *:. .:*:**.***.*

1OG5_B|PDBID|CHAIN|SEQUENCE      KWKEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTF 149
1NR6_A|PDBID|CHAIN|SEQUENCE      TWKEMRRFSLMTLRNFGMGKRSIEDRIQEEARCLVEELRKTNASPCDPTF 150
                                 .***:***********************:***************:********

1OG5_B|PDBID|CHAIN|SEQUENCE      ILGCAPCNVICSIIFHKRFDYKDQQFLNLMEKLNENIEILSSPWIQVYNN 199
1NR6_A|PDBID|CHAIN|SEQUENCE      ILGCAPCNVICSVIFHNRFDYKDEEFLKLMESLHENVELLGTPWLQVYNN 200
                                 ************:***:******:**:***.*:**:*:*.:**:*****

1OG5_B|PDBID|CHAIN|SEQUENCE      FPALLDYFPGTHNKLLKNVAFMKSYILEKVKEHQESMDMNNPQDFIDCFL 249
1NR6_A|PDBID|CHAIN|SEQUENCE      FPALLDYFPGIHKTLLKNADYIKNFIMEKVKEHQKLLDVNNPRDFIDCFL 250
                                 ********** *:.****. ::*.:*:*******: :*:***:*******

1OG5_B|PDBID|CHAIN|SEQUENCE      MKMEKEKHNQPSEFTIESLENTAVDLFGAGTETTSTTLRYALLLLLKHPE 299
1NR6_A|PDBID|CHAIN|SEQUENCE      IKMEQENN---LEFTLESLVIAVSDLFGAGTETTSTTLRYSLLLLLKHPE 297
                                 :***:*::    ***:***  :. ****************:*********

1OG5_B|PDBID|CHAIN|SEQUENCE      VTAKVQEEIERVIGRNRSPCMQDRSHMPYTDAVVHEVQRYIDLLPTSLPH 349
1NR6_A|PDBID|CHAIN|SEQUENCE      VAARVQEEIERVIGRHRSPCMQDRSRMPYTDAVIHEIQRFIDLLPTNLPH 347
                                 *:*:************:*********:*******:**:**:******.***

1OG5_B|PDBID|CHAIN|SEQUENCE      AVTCDIKFRNYLIPKGTTILISLTSVLHDNKEFPNPEMFDPHHFLDEGGN 399
1NR6_A|PDBID|CHAIN|SEQUENCE      AVTRDVRFRNYFIPKGTDIITSLTSVLHDEKAFPNPKVFDPGHFLDESGN 397
                                 *** *::****:***** *: ********:* ****::*** *****.**

1OG5_B|PDBID|CHAIN|SEQUENCE      FKKSKYFMPFSAGKRICVGEALAGMELFLFLTSILQNFNLKSLVDPKNLD 449
1NR6_A|PDBID|CHAIN|SEQUENCE      FKKSDYFMPFSAGKRMCVGEGLARMELFLFLTSILQNFKLQSLVEPKDLD 447
                                 ****.**********:*****.** ***************:*:***:**:**

1OG5_B|PDBID|CHAIN|SEQUENCE      TTPVVNGFASVPPFYQLCFIPVHHHH 475
1NR6_A|PDBID|CHAIN|SEQUENCE      ITAVVNGFVSVPPSYQLCFIPIHHHH 473
                                  *.***** **** *******:****
```

76% identity
RMSD 1.4 Ang

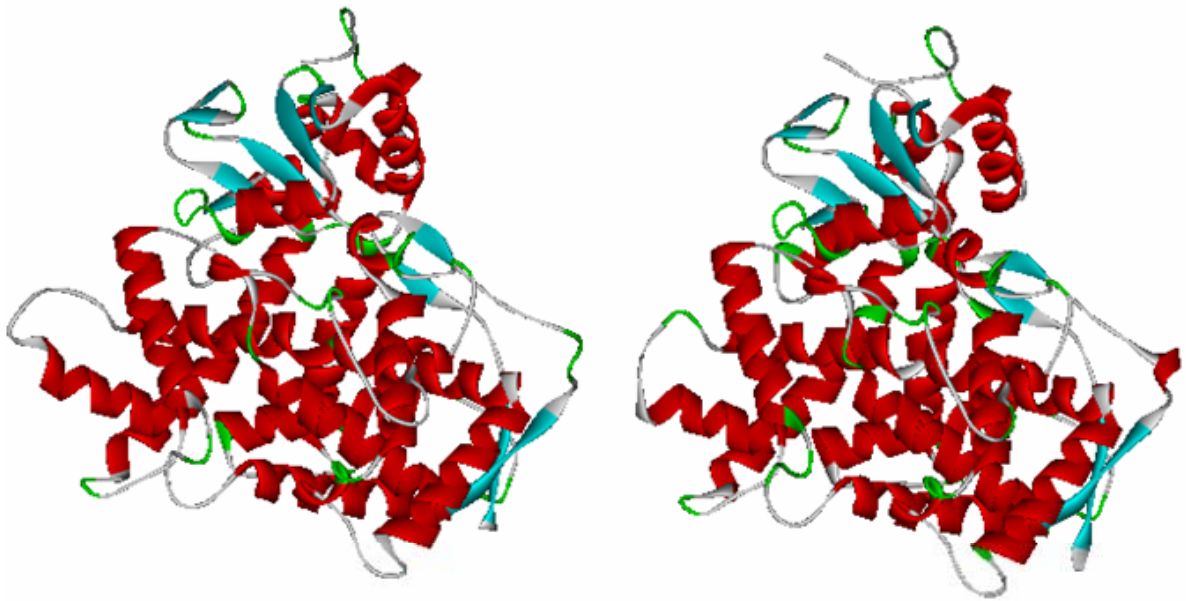
CYP 450 (All Alpha)

1nr6

Science, 15:808 Nayeem et al, Protein (2006)

Structural Overlay done using DeepView

1og5_B



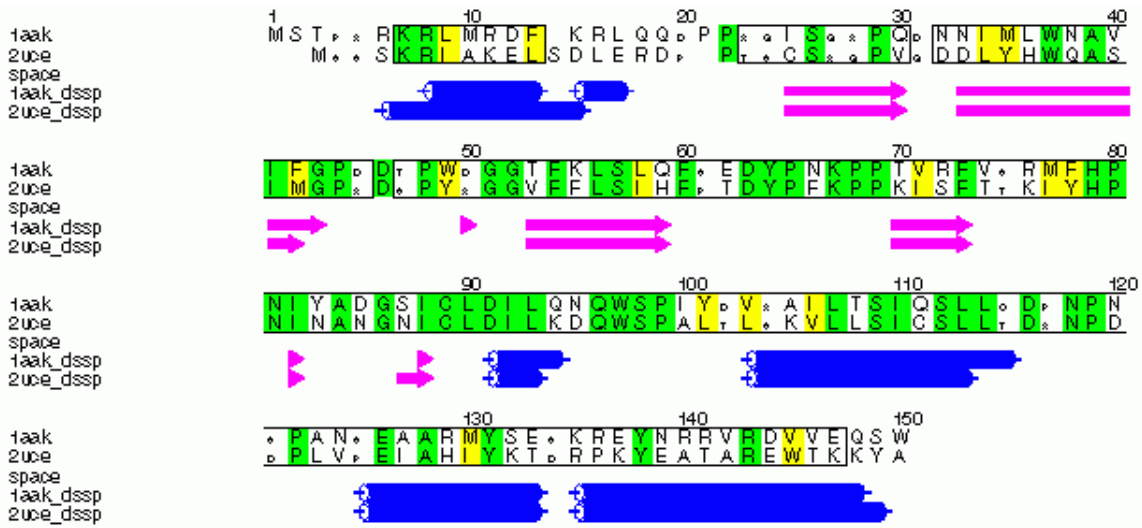
Comparing Homologous enzymes

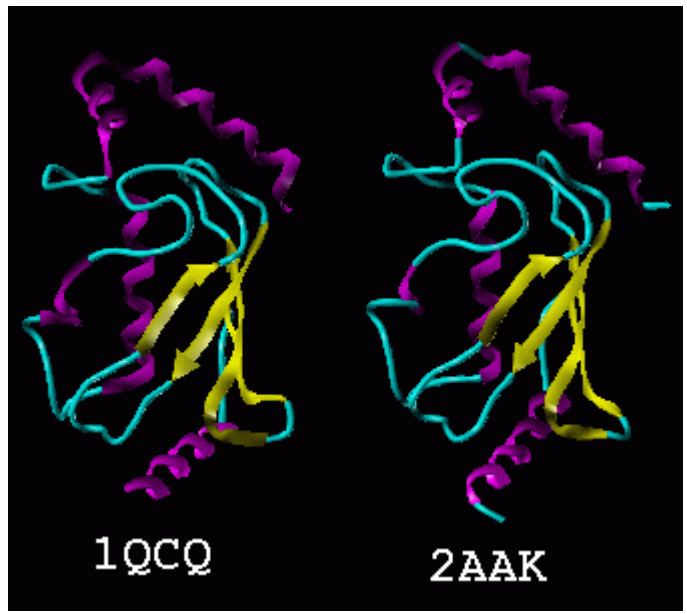
Family: Ubiquitin Conjugating enzyme

1QCQ: Arabidopsis Thaliana 2AAK: Baker's Yeast

Sequence Identity 43%

Topological Similarity



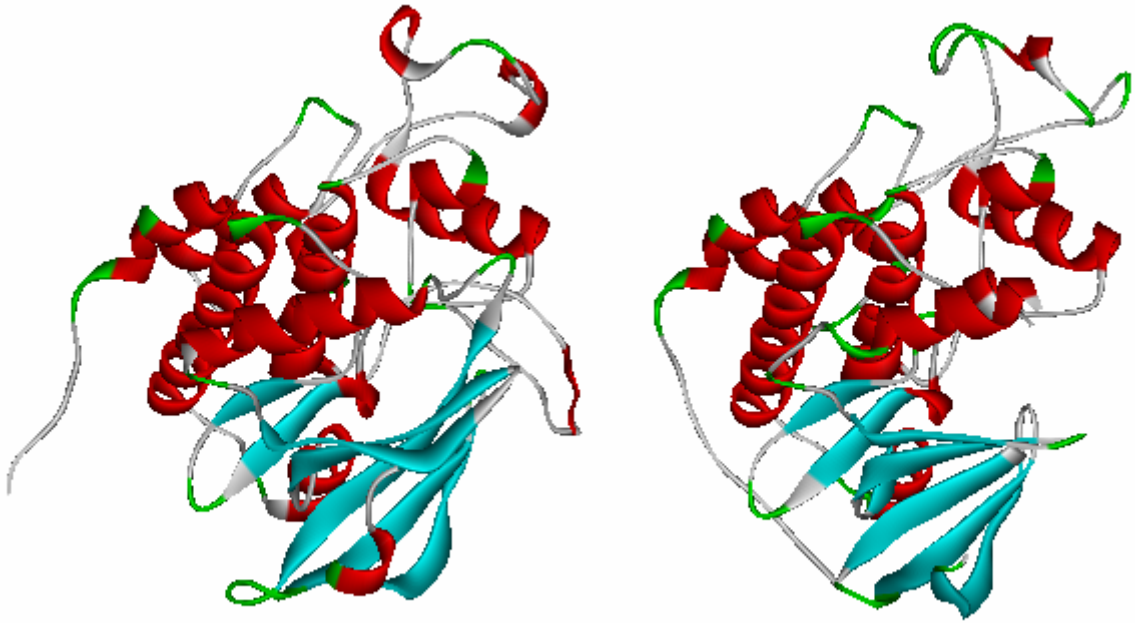


Using PredictProtein

Russell *et al,* JMB, **269,** 423-439 **1997**

Cyclin Dependent
Kinase (Alpha + Beta)

42% identity
RMSD 1.8 Ang



1blx

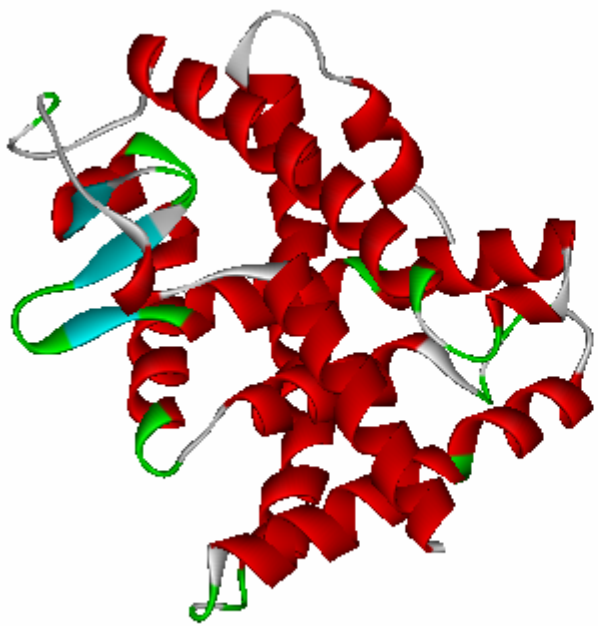
1ckp

Structural Overlay done
using DeepView

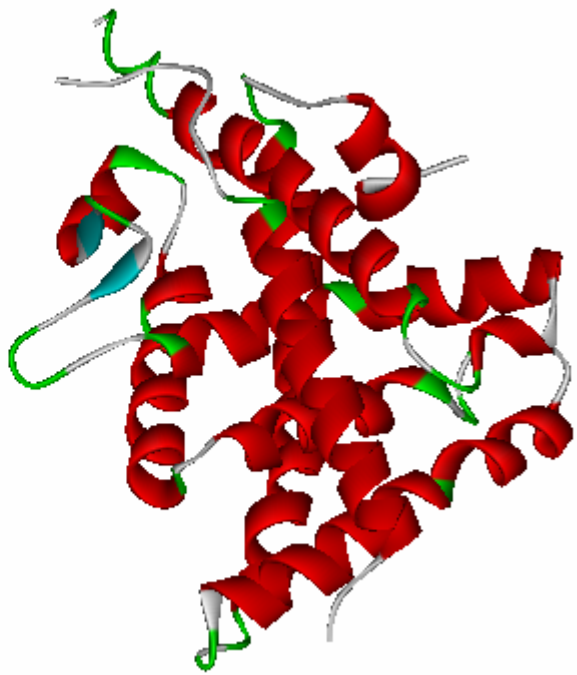
Nayeem et al, Protein Science, 15:808 (2006)

Nuclear Hormone Receptor (All Alpha)

38% identity
RMSD 1.3 Ang



2LBD



1P8D

Structural Overlay done using DeepView

Nayeem et al, Protein Science, 15:808 (2006)

```
1CYD_C|PDBID|CHAIN|SEQUENCE      MKLNFSGLRALVTGAGKGIGRDTVKALHASGAKVVAVTRTNSDLVSL--- 47
1A27__|PDBID|CHAIN|SEQUENCE      -----ARTVVLITGCSSGIGLHLAVRLASDPSQSFKVYATLRDLKTQGRL 45
                                       :    .*:**...***  .  .   * :. ::  .  *   *   ** :

1CYD_C|PDBID|CHAIN|SEQUENCE      -----AKECP--GIEPVCVDLGDWDATEKALG--GIGPVDLLVNNAALVI 88
1A27__|PDBID|CHAIN|SEQUENCE      WEAARALACPPGSLETLQLDVRDSKSVAAARERVTEGRVDVLVCNAGLGL 95
                                      *   **   .:*.: :*: * .:.   *       * **:** **.* :

1CYD_C|PDBID|CHAIN|SEQUENCE      MQPFLEVTKEAFDRSFSVNLRSVFQVSQMVARDMINRGVPGSIVNVSSMV 138
1A27__|PDBID|CHAIN|SEQUENCE      LGPLEALGEDAVASVLDVNVVGTVRMLQAFLPDMKRRGS-GRVLVTGSVG 144
                                 : *:   : ::*.    :.**: ...:: * .   ** .**   * :: ..*:

1CYD_C|PDBID|CHAIN|SEQUENCE      AHVTFPNLITYSSTKGAMTMLTKAMAMELGPHKIRVNSVNPTVVLTDMGK 188
1A27__|PDBID|CHAIN|SEQUENCE      GLMGLPFNDVYCASKFALEGLCESLAVLLLPFGVHLSLIECGPVHTAFME 194
                                 . : :*    .*.::* *:   * :::*: * *. :::. ::    * * : :

1CYD_C|PDBID|CHAIN|SEQUENCE      KVSADPEFARKLKERHPLRKFAEVEDVVNSIL------------FLLSDR 226
1A27__|PDBID|CHAIN|SEQUENCE      KVLGSPEEVLDRTDIHTFHRFYQYLAHSKQVFREAAQNPEEVAEVFLTAL 244
                                 ** ..** .  .  .: *.:::* :      :.::              .:*:

1CYD_C|PDBID|CHAIN|SEQUENCE      SASTSGGGILVDAGYLAS--------------------------- 244
1A27__|PDBID|CHAIN|SEQUENCE      RAPKPTLRYFTTERFLPLLRMRLDDPSGSNYVTAMHREVFGDVPA 289
                                  *...     :.    :*.
```

ClustalW 1.83
www.ebi.ac.uk

19% identity
RMSD
1.7 Ang

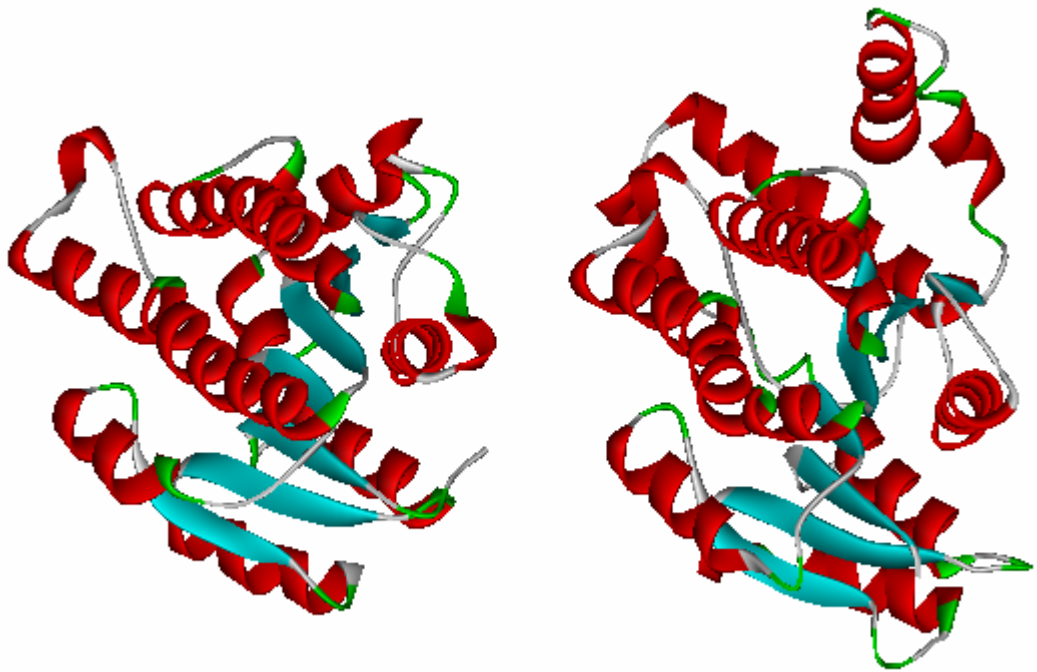
1A27

Nayeem et al, Protein Science, 15:808 (2006)

Short Chain
Dehydrogenase
(Alpha + Beta)

Structural Overlay done
using DeepView

1CYD_C



Sequence Dissimilarity & Structural Similarity

Doolittle's Rule of thumb:

- Sequences longer than 100 aa and has more than
  - 25% identity (with appropriate gaps): Very likely related
  - 15-25% identity: May still be related
  - < 15% probably not
    - How do we make sure that the alignment in the <15% (twilight zone) is biologically meaningful
      - » Random Shuffling-Random mutations and comparison with original score to make sure that the alignment is not random

Homology Modeling: Terminology & Basic Assumptions

Terminology:

- Protein sequence we are modeling is called the *Target*
- Homologous protein used in the modeling is called the *Template*

Basic Assumptions

- Similar sequences have similar conformations
- Core regions provide excellent template for modeling the target protein. If the Core regions share 50% identity, then the two proteins can almost always be superimposed with an Root Mean Squared Deviation of 1 Å or less

3D Structure Database

- PDB (Future formats mmCIF and ASN.1 (MMDB))
  - Brookhaven National Laboratories
  - Research Collaboratory for Structural Bioinformatics (RCSB)-Collaborative effort NIST, Rutgers and San Diego Super Computing Facility
    - http://www.rcsb.org
  - Publically available 3-D structures of Proteins, Proteins + Nucleic Acids (DNA), Proteins complexed with metals and inhibitor
  - Experimental methods: X-ray and NMR

3D-Structural Database of biomolecules (PDB) http://www.rcsb.org

PDB: *50 structures* (1975)

NMR: 3776
X-ray: 24334
Total: 28110

May 10, 2005

- NMR (eary 1960s)
  - Dynamic
    - **Multiple Models (Each conformation is a model)**
  - Aqueous environment
  - Limitations
    - Size of molecule
      - < 30kD
- Examples
  - 1DV0, 1UBA

- X-ray (1958)
  - Static
    - Only one model
  - Crystal
  - Limitations
    - Not limited by size
- Examples
  - 7LYZ , 2SRC

NMR: 5409
X-ray: 31223
e^- micros: 125
Total: 36837

Jun 01, 2006

Anatomy of PDB file

Atom #

Residue

Residue #

```
ATOM      1  N   MET B   1      52.127  -7.410  40.963  1.00 52.97
ATOM      2  CA  MET B   1      51.096  -6.600  40.340  1.00 52.20
ATOM      3  C   MET B   1      51.305  -6.416  38.831  1.00 51.74
ATOM      4  O   MET B   1      52.405  -6.086  38.379  1.00 50.94
ATOM      5  CB  MET B   1      51.012  -5.245  41.044  1.00 52.27
ATOM      6  CG  MET B   1      50.777  -5.353  42.542  1.00 50.84
ATOM      7  SD  MET B   1      49.355  -6.390  42.934  1.00 51.51
ATOM      8  CE  MET B   1      48.078  -5.169  43.363  1.00 47.75
ATOM      9  OXT MET B   1      50.932  -9.312  40.816  1.00 53.41
ATOM     11  N   ASN B   2      50.235  -6.630  38.064  1.00 50.80
ATOM     12  CA  ASN B   2      50.271  -6.496  36.613  1.00 49.26
ATOM     13  C   ASN B   2      50.332  -5.038  36.246  1.00 48.37
ATOM     14  O   ASN B   2      50.120  -4.673  35.089  1.00 50.24
ATOM     15  CB  ASN B   2      49.016  -7.074  35.977  1.00 49.62
ATOM     16  CG  ASN B   2      48.753  -8.479  36.395  1.00 51.39
ATOM     17  OD1 ASN B   2      49.628  -9.339  36.316  1.00 51.26
ATOM     18  ND2 ASN B   2      47.531  -8.701  36.861  1.00 54.69
...........
HETATM 2462 ZN    ZN   909      45.731   9.445  45.851  0.54 77.21
HETATM 2463  C1  RET B 978      33.234   8.591  25.798  1.00 34.05
HETATM 2464  C2  RET B 978      31.995   8.387  24.968  1.00 33.95
HETATM 2465  C3  RET B 978      32.242   8.645  23.513  1.00 33.45
HETATM 2466  C4  RET B 978      32.720  10.104  23.258  1.00 33.69
HETATM 2467  C5  RET B 978      33.717  10.591  24.302  1.00 34.43
HETATM 2468  C6  RET B 978      33.938   9.900  25.443  1.00 34.95
HETATM 2469  C7  RET B 978      34.915  10.362  26.451  1.00 36.11
```

Temp Factor

X Y Z Occ

Chain ID

Only portion of the file is shown

PDB

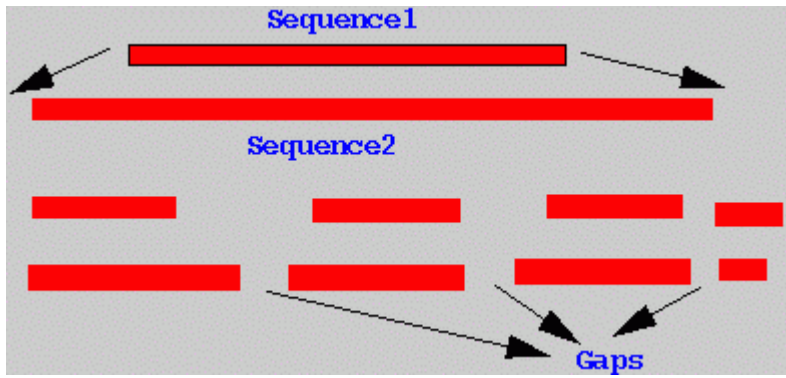
- X-ray structures do not include H-atoms (ex 1qpc)
  - Careful about acidic/basic residues, Water (only oxygen atoms)
- Resolution
  - Smaller the better (1.5Å is better than 2.5Å)
- As a result of proteolytic activity some information about structure could be lost
  - Flexible loops etc.
- Some proteins/enzymes fulfill their function only as dimers or trimers

Database mining

- Why Sequence Comparison?
- Template sequences: Global Alignment
  - Overall alignment sequence homologs with known 3-D str.
- Sequences of both template & target: Local Alignment
  - Best for searching local domains
    - Gaps cannot be introduced endlessly
      - Biologically meaningless

GA_CGGATTAG
GATCGGAATAG

Local Alignment
Smith-Waterman



Global Alignment
Needleman-Wunsch

Scoring Schemes

- Scheme based on Identity
- ” based on Chemical Similarity
- ” based on Genetic Code
- ” based on Observed Mutations

Example of Identity Scoring Scheme

Sequence 1 GACGGATTAG; Sequence 2 GATCGGAATAG

Total Score

9X1+1X(-1)+1X(-2) = 6

Dynamic Programming

Global alignment

G	A	-	C	G	G	A	T	T	A	G
G	A	T	C	G	G	A	A	T	A	G
1	1	-2	1	1	1	1	-1	1	1	1

PAM BLOSUM

- PAM (Dayhoff et al 1988)
  - PAM1 is the matrix obtained by comparing sequences differ by no more than 1%
  - Higher PAMX are extrapolated from PAM1)
  - PAM250: Observed difference(80%) Evolutionary distance (250)
- Limitation: Matrices are derived from alignments of sequence that are 85% identity
  - Difficult to use in Twilight Zone

- Blosum (BLOcks SUbstitution Matrix) Heinkoff & Heinkoff (1992)
  - Derived from BLOCKS database
  - Distant relationships explained better than PAM
  - Blosum62 obtained from sequence BLOCKS clustered at >= 62% identity (default Blast)
  - BlosumX are observed from actual alignments not extrapolated
  - Sequences are very similar use higher Blosum (low PAM)

NCBI

BLOSUM64

Sean R. Eddy, Many References

20 x 20 matrix

MSA

log-odd Score values

$$S(a,b) = 1/\lambda \log(p_{ab}/f_a f_b)$$

p_{ab} Prob. of a & b aligned (correlated)
$f_a f_b$ How the two residues are uncorrelated
What is the probability of observing ab in comparison to chance ?
λ Scaling factor

Calculation of λ

$$\sum_{a,b} f_a f_b e^{\lambda S(a,b)} = 1$$

W/W +11

$p_{ab} = 0.0065$
$f_a = 0.013$, $\lambda = 0.347$

$S(a,b) = \lambda(0.0065/0.013*0.013)$
$= +10.5$

L/L = +4.0

$p_{ab} = 0.0371$
$f_a = 0.099$ $\lambda = 0.347$

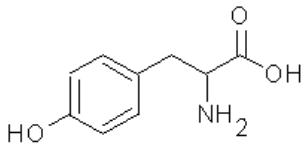
$\lambda(0.0371/0.099*0.099) = +3.8$

PAM250 Matrix (mutation values) (identities at 20% level)

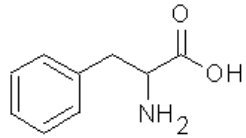
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	.
Ala	2	-2	0	0	-2	0	0	1	-1	-1	
Arg	-2	6	0	-1	-4	1	-1	-3	2	-2	
Asn	0	0	2	2	-4	1	1	0	2	-2	
Asp	0	-1	2	4	-5	2	3	1	1	-2	
Cys	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	
Gln	0	1	1	2	-5	4	2	-1	3	-2	
Glu	0	-1	1	3	-5	2	4	0	1	-2	
Gly	1	-3	0	1	-3	-1	0	5	-2	-3	
His	-1	2	2	1	-3	3	1	-2	6	-2	
Ile	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	
.											

Cystein: Well-known for S-S linkage Important for structure

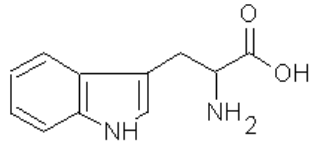
Tryptophan : Highly conserved-Hydrophobic core residue-Important for the structure-difficult to mutate. Trp->Phe, Trp->Tyr (aromatic acids are the next choice to replace Trp)



Tyrosine (Tyr)



Phenylalanine (Phe)



Tryptophan (Trp)

Unitary Matrix

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0		0	1

Searching for Templates

- Do a Blast/Psi-Blast/Fasta or use programs within GCG (Align, gap, bestfit, etc.) for sequence alignment & Family identification.
  - Proteins(Blast); DNA(FASTA)
  - Restrict search only to PDB database why PDB?
- Potentially suitable templates
  - Blast Score < 0.001 (protein), <=10^(-6) (nucleotide)
  - Safe threshold is > 25-30% identity
  - In the Twilight Zone (< 25%) How to proceed?
    - Randomization of sequences and realignment

Pearson

BLOSUM62: Based on 62% identity

Merge all aligining sequences have 62% or more into one block. For sequences less than 62% identities are weighted more heavily

- Usually more than one protein is chosen as templates?
  - Avoid biasing, to model variants (loops etc), side chain conformations
  - Final model will be done using one representative template (called reference)

Structurally Conserved Region (SCR) Modeling

- After identifying template(s), the next task is to identify the SCR

```
  L (     1) KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFE-SNFNTH (41
IHL (     1) KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFE-SNFNTQ (41
HEL (     1) KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFE-SNFNTQ (41
GHL (    A0)GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFE-SNFNTG (A41
LZ1 (     1) KVFERCELARTLKRLGMDGYRGISLANWMCLAKWE-SGYNTR (41
ALC (     1) kqftkcelsqnl--ydidgygrialpelic-tmfhtsgydtq (39
```

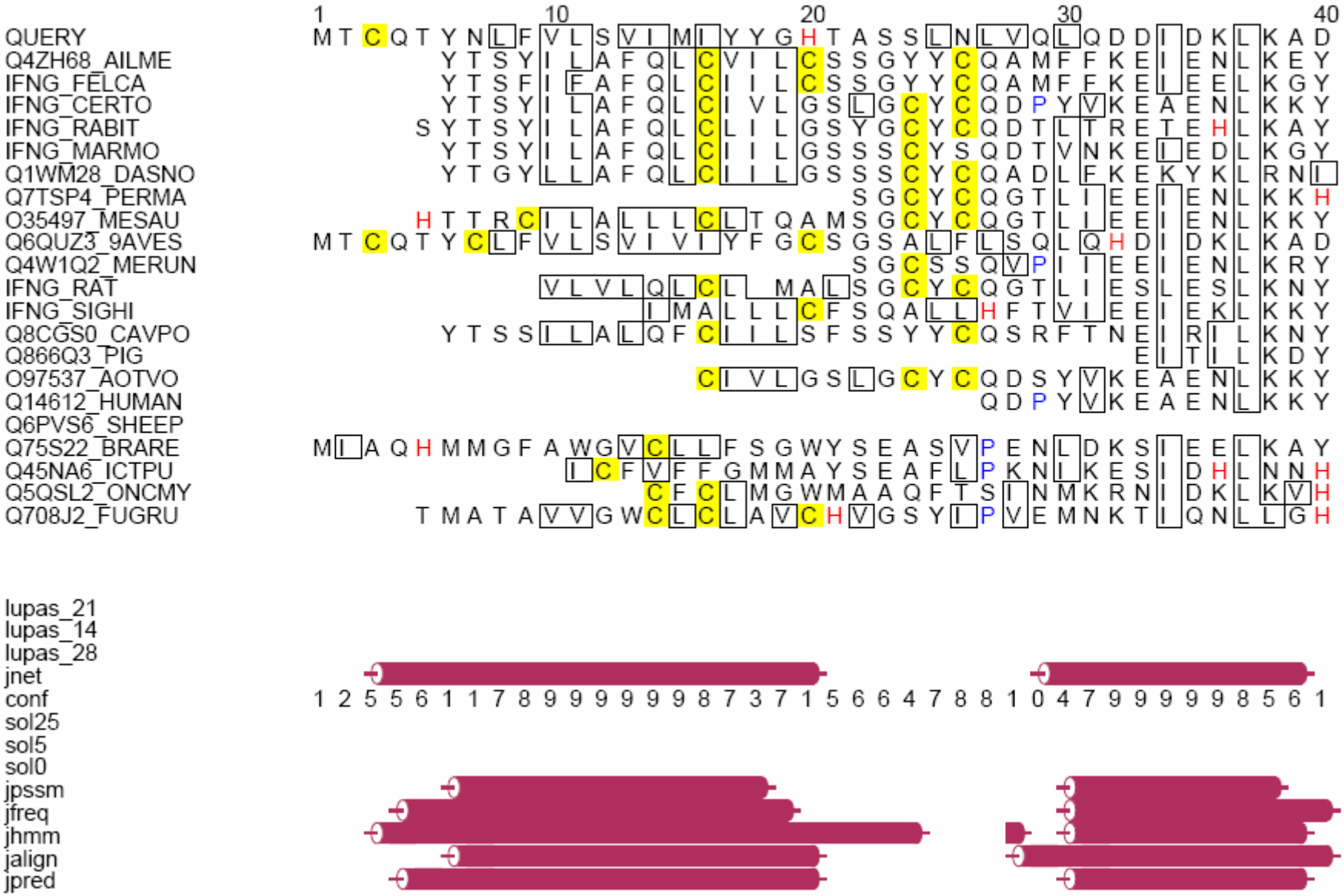
- What are SCRs?
  - Inner core (not the surface exposed loops)
  - How do we identify them?
    - Multiple structure Alignments, secondary structure elements
- Assigning SCR within the family of homologous proteins
  - Align them based on their backbone using Least-square fitting methods
    - Optimizing the alignment using matching points from the conserved secondary structural elements
      - At this point you should see the structures are aligned in the secondary structural areas and non-conservation in the surface loops etc.

SCRs

- What happens when we have only one template?
  - Use the secondary structure information given in the protein
    - Residues involved in Salt-bridges, H-bonds, di-sulphide bonds etc. are conserved in a protein family
      - Residues involved in active site
  - Use Theoretical methods to predict
    - Multiple Sequence Alignments

Multiple Sequence Alignment & Secondary Structure Prediction

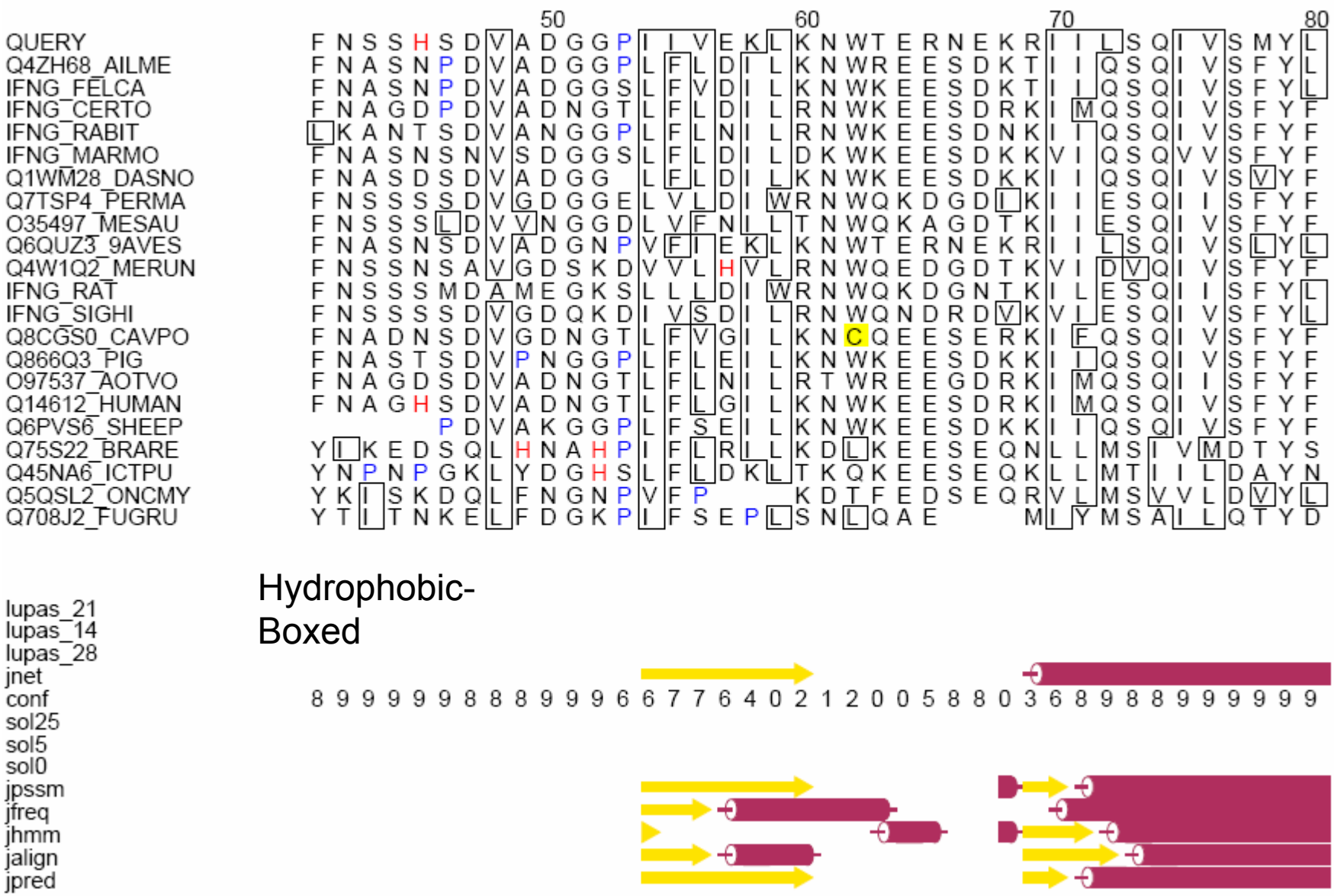
Query: IFN-Gamma Chicken Seq.



Made with JPRED Server

Multiple Sequence Alignment & Secondary Structure Prediction

Query: IFN-Gamma Chicken Seq.



Hydrophobic-Boxed

NCI-Frederick

Made with JPRED Server

SCRs

- The next step is to align the Structurally aligned templates with the unknown sequence
- What is known?
  - SCRs also show strong sequence homology
    - Loop regions of aligned sequences also show high similarity but structurally variant
  - No gaps are allowed within the SCR regions
    - Special sequence alignment algorithm used which discourages gaps within SCR

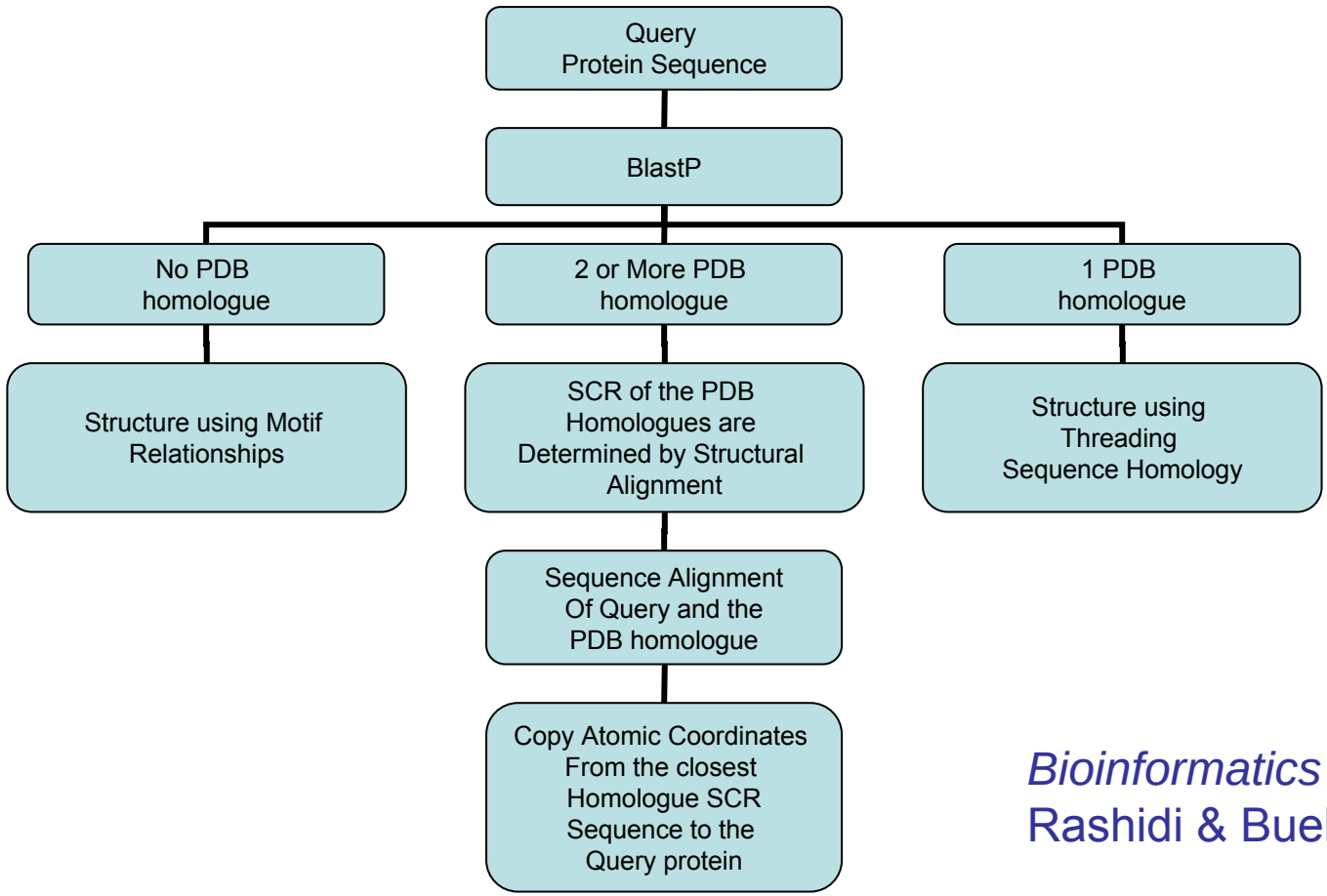
Structurally Variable Region (SVR) Modeling (3 methods)

- If the reference protein has similar loops then it can be copied
- Perform a database (derived from PDB) search for structures with loops
  - Criterion is the conserved residues flanking the loop area and the # of loop residues
    - Software usually keep a loop database derived from PDB.
- *de novo* method of building and constrained minimization
  - If the number of residues in the template and the reference differ
- MM calculations carried out as the last step to relieve steric clashes

Modeling Side Chains

- Given that each side chain can be in one of many different conformations—Multiple minima problem
- Following options are generally used:
  - If the residues are same/similar
    - Copy the same conformation (why?—scoring matrix scores)
  - If they are different
    - Use built-in libraries based on known info (PDB)
    - Random conformations without any collisions
- Residues in the border (SCR,SVR) have to be dealt carefully

Overview of Homology Modeling



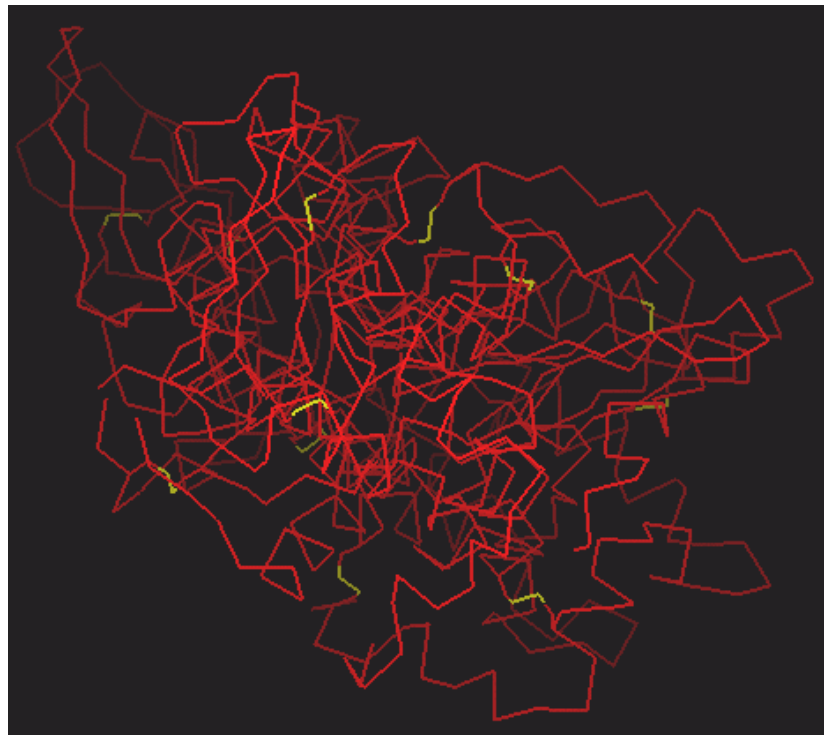
Bioinformatics Basics:
Rashidi & Buehler

Homology Modeling By Example

Homology Module of InsightII

Template Alignment

- 5 template lysozyme proteins (only α-C shown) structurally <u>uncorrected</u> multiple sequence alignment
- Reference Red
- Query Sequence violet



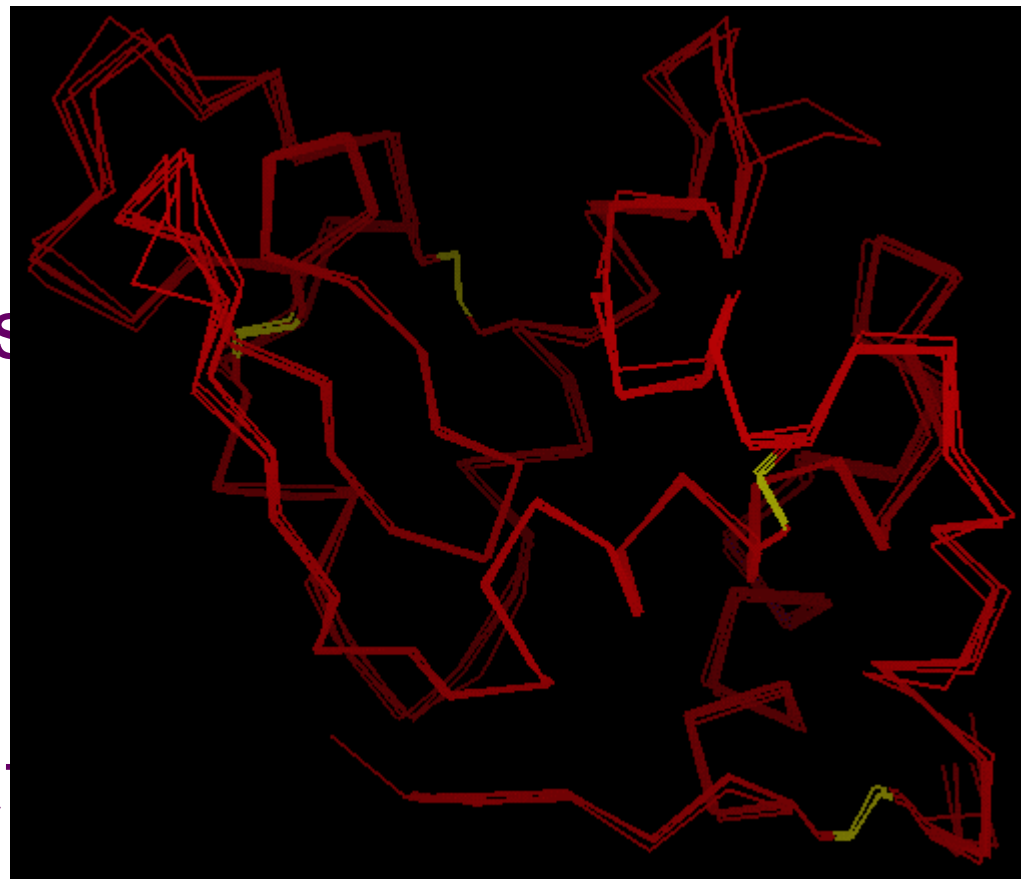
```
  L (     1) KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHA (42
IHL (     1) KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFESNFNTQA (42
HEL (     1) KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQA (42
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTG (A41
LZ1 (     1) KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRA (42
ALC (     1) kqftkcelsqnlydidgygrialpelictmfhtsgydtqaiv (42
```

Studying the corrected template alignment

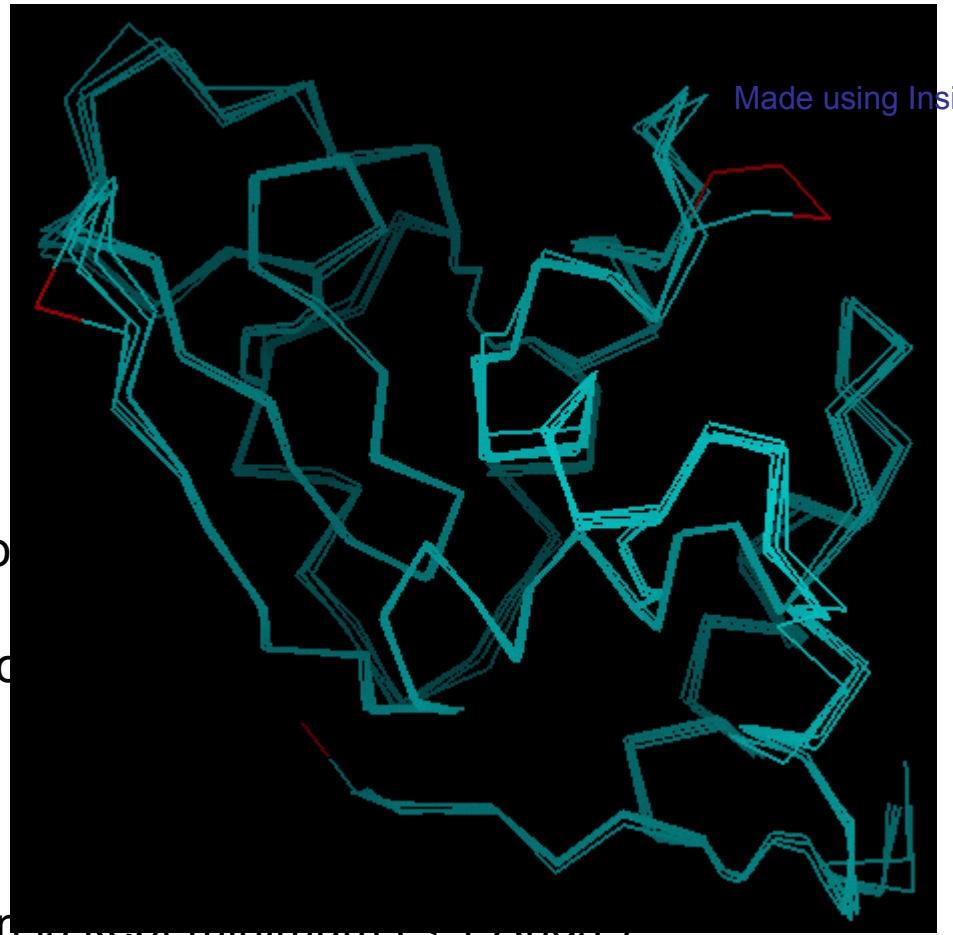
- Look at cys

  How about the Structural Cons veration?

- Which regions structural varia



Structurally corrected MSA



Made using InsightII, Accelrys

	A	B
A	0	0.5
B	0.5	0

Do you see the location o the variable region (core c surface)

RMS deviation is kept minimum (< 1 Angs.)
Structurally corrected MSA

Target Core Modeling

- Target sequence is aligned with the template or Structurally Corrected Multiple Sequence alignment (in case of templates)
  - Which residues can be aligned to the conserved block region of the multiple sequence alignment of the reference protein so that one can copy the coordinates from the reference to the sequence
    - Do a sequence alignment using a chosen matrix, gap penalty etc. of the reference with the model sequence

Target Core Modeling

- *Target sequence* is now aligned with the template or Structurally Corrected Multiple Sequence alignment (in case of templates)

Made using InsightII, Accelrys

```
  L (     1)  KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTH (41
IHL (     1)  KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFESNFNTQ (41
HEL (     1)  KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQ (41
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTG (A41
LZ1 (     1)  KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTR (41
ALC (     1) kqftkcelsqnlydidgygrialpelictmfhtsgydtqaiv (42
```

Alignments will be carried out between SCRs. Special algorithms (not Needleman-Wunsch) can treat each SCR.

Sequence Alignment

Before Aligning the model sequence to the template

```
  L (     1)  KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTH (41
IHL (     1)  KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFESNFNTQ (41
HEL (     1)  KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQ (41
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTG (A41
LZ1 (     1)  KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTR (41
ALC (     1) kqftkcelsqnlydidgygrialpelictmfhtsgydtqaiv (42
```

Are these insertions reasonable?

Gap insertion, conserved region split

After Aligning the model sequence to the template

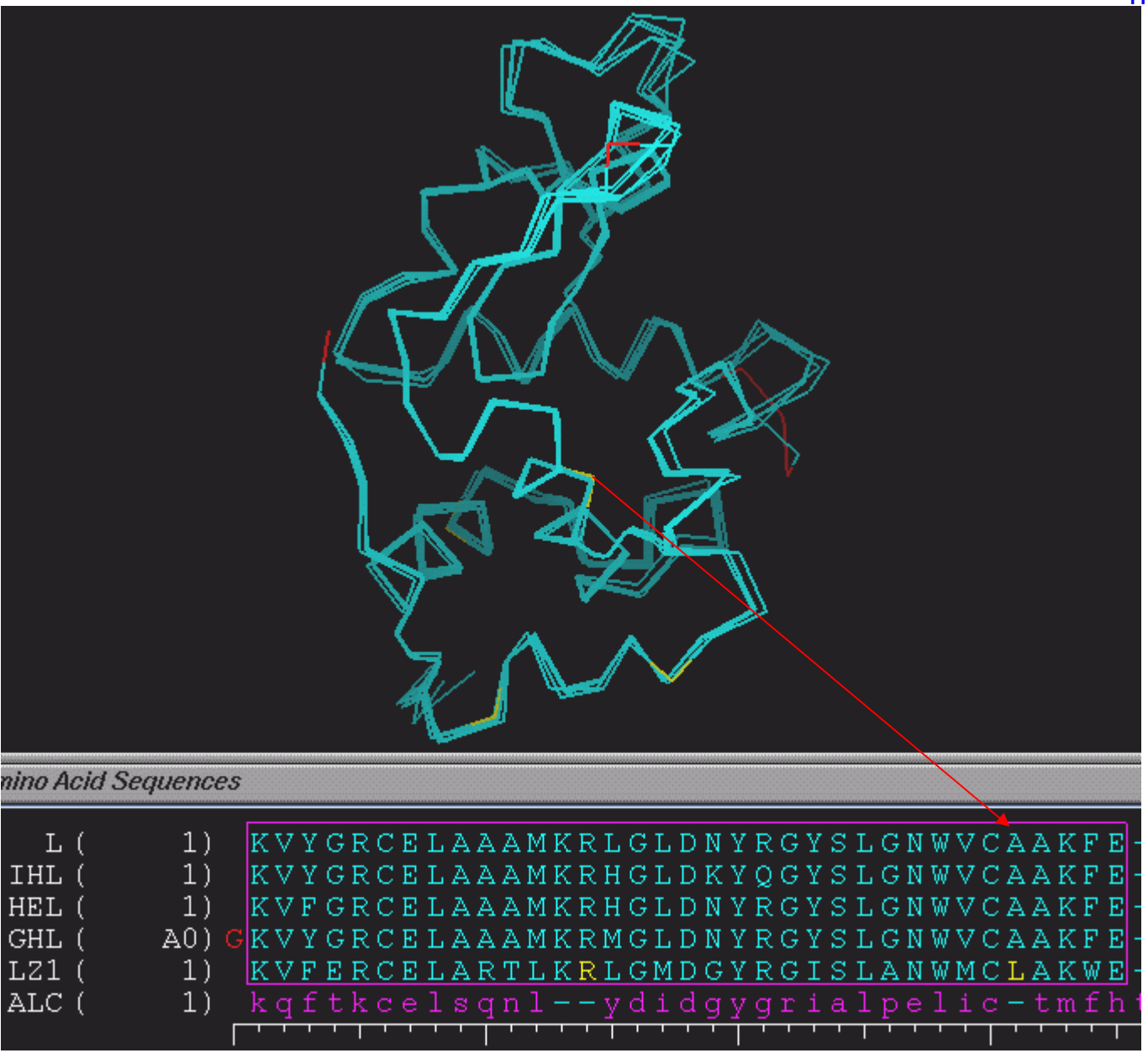
```
  L (     1)  KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFE-SNFNTH (41
IHL (     1)  KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFE-SNFNTQ (41
HEL (     1)  KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFE-SNFNTQ (41
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFE-SNFNTG (A41
LZ1 (     1)  KVFERCELARTLKRLGMDGYRGISLANWMCLAKWE-SGYNTR (41
ALC (     1)  kqftkcelsqnl--ydidgygrialpelic-tmfhtsgydtq (39
```

Made using InsightII, Accelrys

Gap insertion

Suspect the alignment

- Look at the alignment and if the gaps introduced are not in the surface exposed then go examine the parameters of the alignment (gap-penalty etc.)
- If the deletions occur at the end-terminus, surface exposed, not in any recognized secondary structure, then they may be valid deletions
- Finally, copy the coordinates from each conserved group of <u>one of the most similar</u> <u>sequence template</u> (based on phylogeny) to the model sequence.
  - Other alternative is “Distance Geometry” approach



Query length	matrix	Gap Costs (Exis, Exten)
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
>85	BLOSUM-62	(11,1)

1) Before alignment 2) wrong alignment parameters 3) correct alignment parameters (higher gap penalty)

1

```
  L (     1)  KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTH (41
IHL (     1)  KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFESNFNTQ (41
HEL (     1)  KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQ (41
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTG (A41
LZ1 (     1)  KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTR (41
ALC (     1) kqftkcelsqnlydidgygrialpelictmfhtsgydtqaiv (42
```

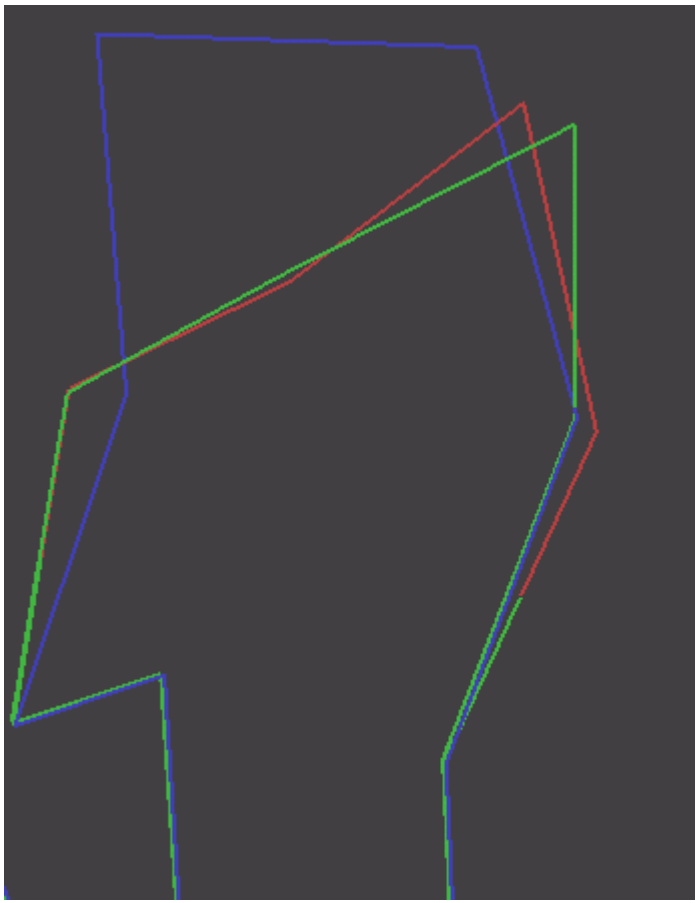
2

```
  L (     1)  KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFE-SNFNTH (41
IHL (     1)  KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFE-SNFNTQ (41
HEL (     1)  KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFE-SNFNTQ (41
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFE-SNFNTG (A41
LZ1 (     1)  KVFERCELARTLKRLGMDGYRGISLANWMCLAKWE-SGYNTR (41
ALC (     1)  kqftkcelsqnl--ydidgygrialpelic-tmfhtsgydtq (39
```

3

```
  L (     1)   KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTH (41
IHL (     1)   KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFESNFNTQ (41
HEL (     1)   KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQ (41
GHL (    A0)  GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTG (A41
LZ1 (     1)   KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTR (41
ALC (     1)   KQFTKCELSQnl--yDIDGYGRIALPELICTMFHTSGYDTQ (39
```

Loop Modeling

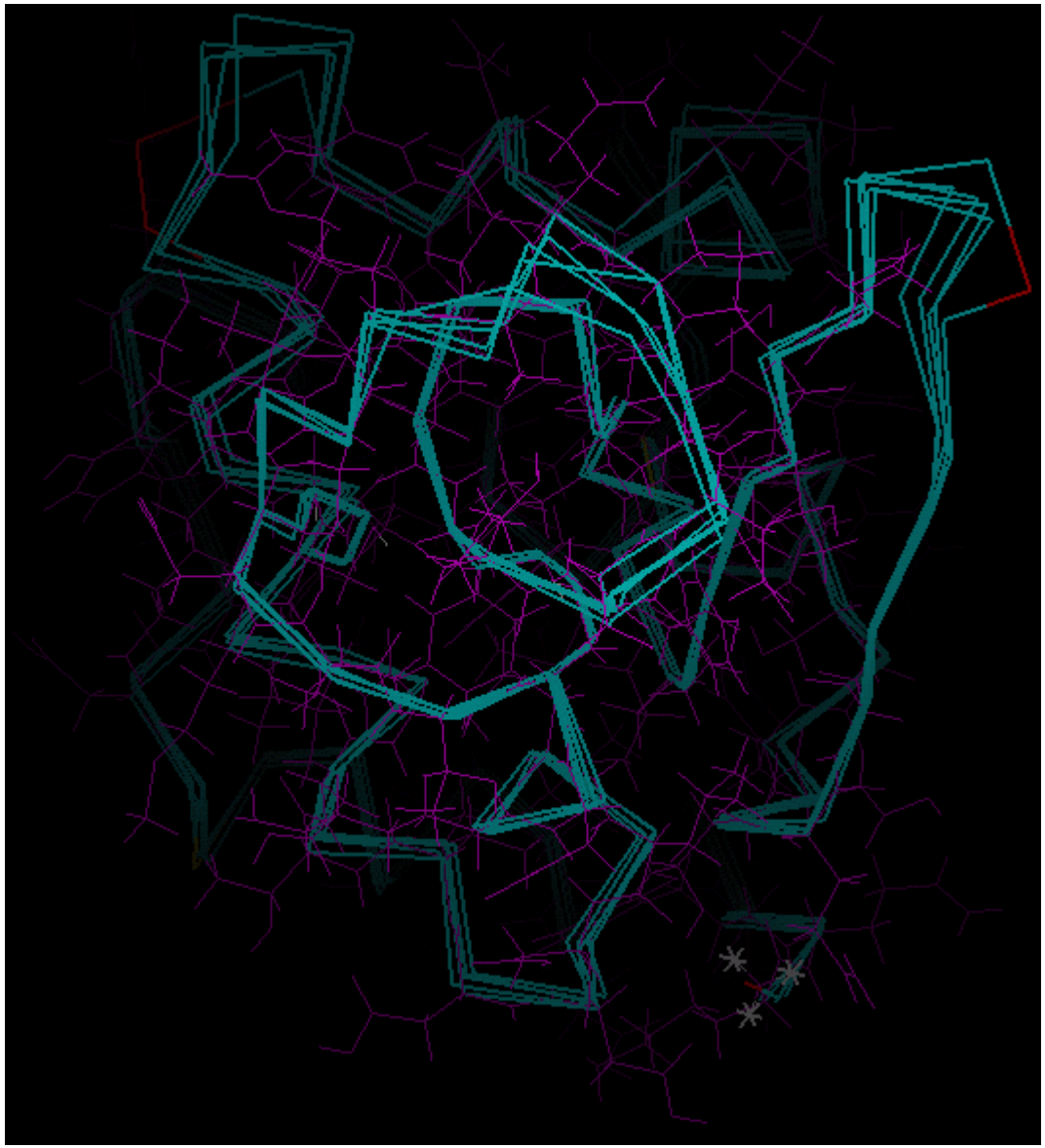


What do we know now?

Loop regions do not have random strs.

It depends on

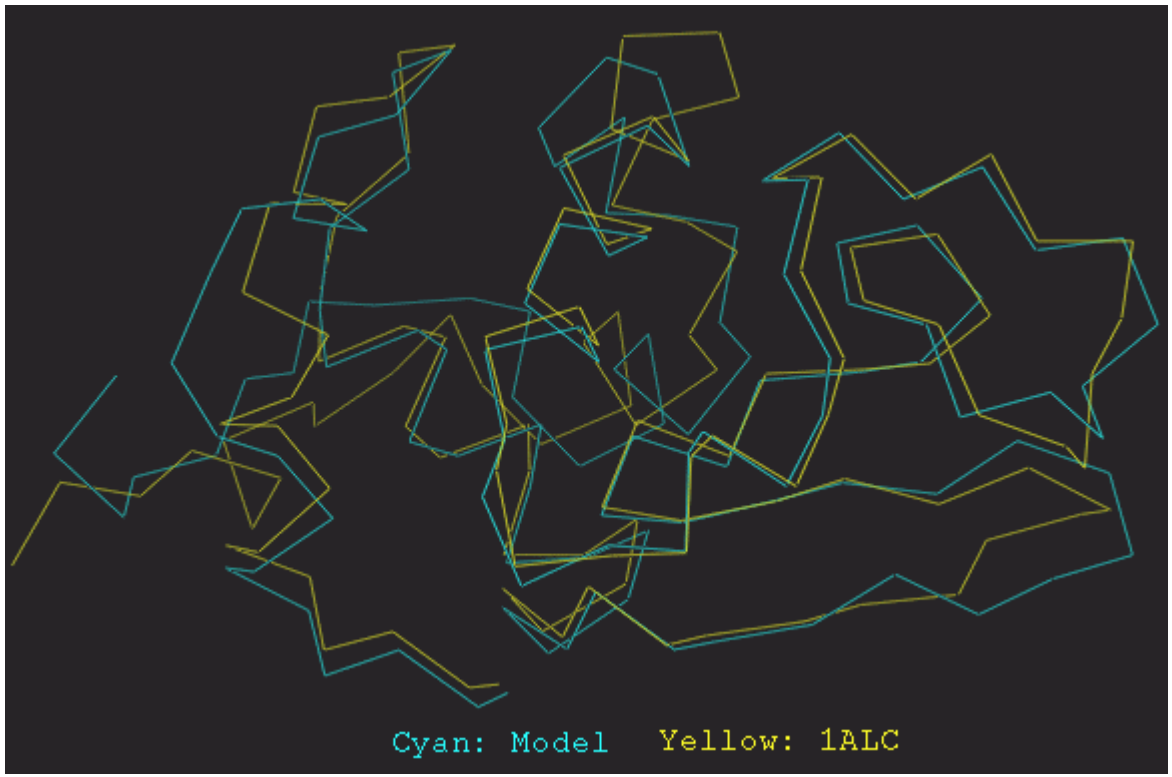
- aas in the loop
- Type of Sec Str elts it connects
  - α-α, β-β, α – β
- Loop database can be built (PDB)



Side Chains will be added if the template has identical residues

Rotamers will be generated which doesn't clash with the backbone

Final Model



Homology Model Evaluation

- Most automated Homology Modeling software provides a model, even with an inappropriate template
- How to judge the quality of the model?
  - Absence of R-factors-No way to evaluate the model
  - Correct models usually have atomic positions within the experimental uncertainty limit

Final Step: Energy Minimization

- Why? The final model now has backbone+side-chains+loops generated from the template(s)
  - Has atom clashes and non-optimal conformations
- Choose a program to perform Energy Minimization to repair the model structure (bad contacts)
  - Suitable Force-Fields-GROMOS, CVFF, GROMOS, CHARMm etc.
    - Swiss-Model uses GROMOS
- How many steps of Minimization ?
  - Vacuum (non-solvent)-minimum number of steps

Identifying Incorrect Models

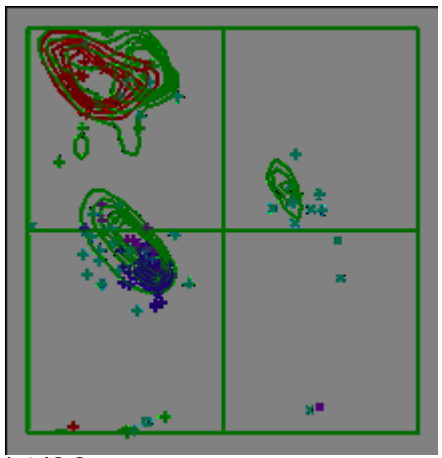
- Hydrophobic residues exposed
- Buried polar or ionic residues without the charges satisfied (H-bonds, salt-bridge etc)
- Clashes
- Unusual bond-lengths, bond-angles
- Sequence alignment is not-optimal
- Very large RMSD among the templates

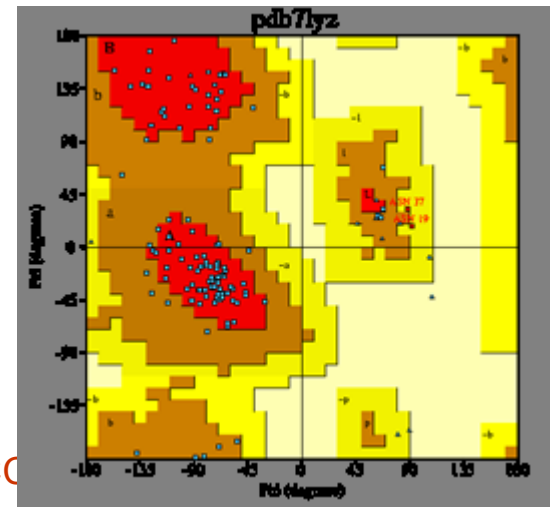
Quality of Models

- Procheck: Stereo-chemical quality of the protein and residue by residue analysis in figures

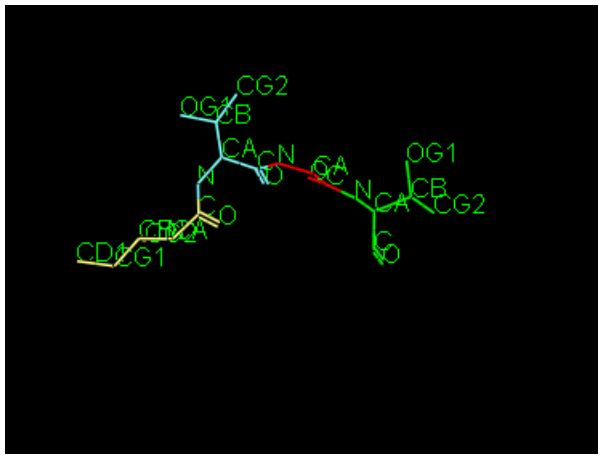
  http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

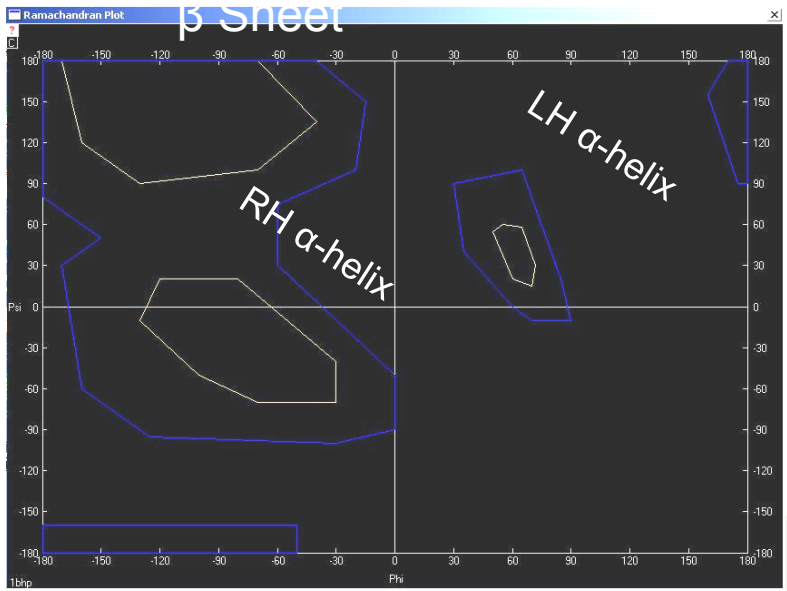
- PDBREPORT: http://www.cmbi.kun.nl/gv/pdbreport

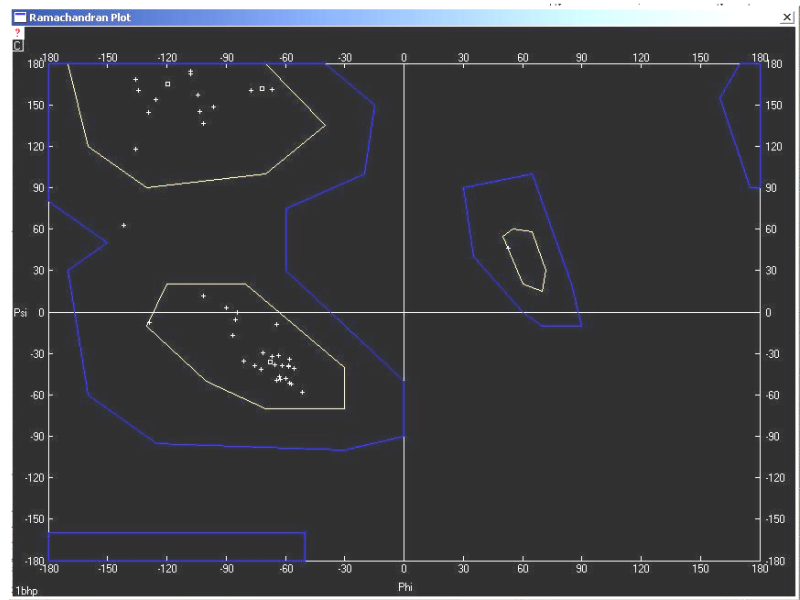




Ramachandran Plot

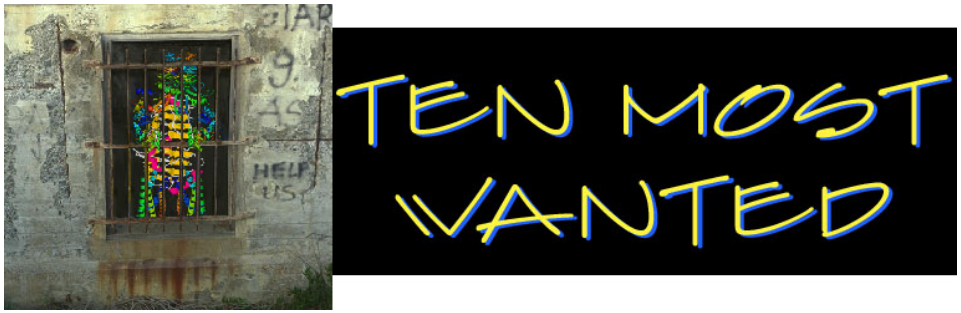






Quality of Models

- Stereochemical Accuracy
  - Choice of template(s) (low resolution etc.), bond-lengths, bond-angles, torsion angles, chirality of C-alpha atoms, planarity of amide bonds
- Packing Quality
  - How secondary structural elements packed together (α-α,β-β, α-β etc.)
    - Inspecting buried residues
- Folding Reliability
  - Comparing overall fold of template and model
    - 3D-profiles-reduce 3D structure to 1D string. This 1D string can then be compared with query sequence after some modification



CASP: Test of the Models

CASP6 [2004]
Proteins: Structure, Function, and Bioinformatics Volume 61, Issue S7 , Pages 225 - 236

- Critical Assessment of Techniques for Protein Structure

  http://predictioncenter.org/

  - Showcase for the latest methods in the structure prediction area
  - Once in two years
  - Competition open in three areas
    - Homology Modeling, Threading and ab-initio
  - CASP (1998, 2000, 2002 & 2004) showed the reliability of Homology Modeling when suitable templates
    are available (>30%, above Twilight Zone)

Database of Homology Models

- Project, 3D-Crunch (1984)
  - Project submitted all sequences of Swiss-Prot and trEMBL to SWISS MODEL server
- The resulting homology models (64,000) are stored and available to public from SWISS-MODEL Repository
  - Database contains: Final models, Entire modeling projects including aligned coordinates of templates

Database of Homology Models

- ModBase Sali and co-workers
  - Software: Modeller
  - Models were built based on spatial restraints
    - Restraints: distances between alpha carbons, distances within main-chain etc
  - E-minimization techniques are employed to obtain these restraints

Meta-Servers

The META server : Volker Eyrich and Burkhard Rost

BioInfoBank MetaServer

The Bio
models

Submit

Ads by Goooooogle

Find Secondary Structures
RAPTOR: casp5 winner, gives quality structure prediction. Free Demo.
www.bioinformaticssolut

SeqWright: DNA Sequencing
Genotyping, Mutagenesis, Synthesis. Shotgun Lib, SNPs, Contigs. GLP.
www.seqwright.com

Protein Analysis Software
User-friendly and Integrated Tools Fully Functional Demo Available!
www.clcbio.com

Sequence analysis tool
Easy-to-use free software for sequence alignment & tree building
www.geneious.com

Advertise on this site

Google

Web BioInfo.PL

Search

Type the required information into the fields (and select one or more services)

Struc
0 jobs

Amino

Submit domains separately
Remove coiled coil regions
Check LiveBench for evaluation of the reliability of the servers
Results are stored only for 1 month
Jobs queued for more than 7 days for servers with queue>30 are skipped
Use is limited to 30 jobs per week per domain
Contact us in case of problems with interpretation of results
Some servers return only models, no alignments (target sequence is shown)
Results published on this server are public and can not be used for patenting

profsec 33
3D-Jury

Description of field (click on description for help)

Your email address (watch for typos)
One-line name of protein (optional)
Paste, or type your sequence

- amino acids in one-letter code (any number of spaces allowed)
- other possible formats

For retrieving protein sequences from databases we recommend the Sequence Retrieval System SRS6

The Structure Prediction Meta Server provides access to various fold recognition, function prediction and local structure prediction methods. The Server takes the amino acid sequence of the query protein, the reference name for the prediction job, and the E-mail address as input. The E-mail address is used only for notification about errors during the execution of the job. The query sequence and the reference name are placed in the process queue. The Meta Server accepts only sequences, which have not been submitted before. In case of duplicate sequences the second user will be notified with a link to the previous submission. Sequences longer than 800 amino acids are not accepted by some services. The internal SQL database offers the possibility to find any previous jobs processed by the Meta Server using "regular expressions" addressing field like E-mail, Job Name and the host name, from which the job was initiated. Each server has its own process queuing system managed by the Meta Server. All results of fold recognition servers are translated into uniform formats. The information extracted from the raw output of the servers includes the PDB codes of the hits, the alignments and the similarity (reliability) scores specific for every server. Mapping of the hits to the SCOP and FSSP classifications are made either using known PDB representatives or alignment of the template sequence with the databases of proteins in both classifications. The secondary structure assignments for all hits are taken from the mapped FSSP (red for helices and blue for strands). Underscored amino acids indicate the first residue after an insertion in the template sequence. The Meta server provides translation of the alignments in standard formats like FASTA, PDB or CASP. The Meta Server is coupled to consensus servers. They provide jury predictions based on the results collected from other services. Not all fold recognition servers are used by the jury system. The data stored on the meta server is available through (http://meta.bioinfo.pl/). Jobs older than 2 months are not shown. The Meta Server is only a set of programs aimed to process and manage biological data, while the predictive power of the service comes from (mostly) remote prediction providers

Please cite the prediction servers and 3D-Jury: Ginalski K, Elofsson A, Fischer D, Rychlewski L. "3D-Jury: a simple approach to improve protein structure predictions." Bioinformatics. 2003 May 22;19(8):1015-8. [Library]

http://cubic.bioc.columbia.edu/meta/

http://bioinfo.pl/meta/

Meta-Servers

http://cubic.bioc.columbia.edu/meta/

GeneSilico Metaserver

This is a gateway to various methods for protein structure prediction. If you submit a single protein sequence or a multiple sequence alignment (hereafter called a "target"), you should obtain the following predictions:

Primary structure: domains identified by HmmPfam

Secondary structure: helical/extended/other conformation (H/E/-), transmembrane helices, and disordered regions.

Tertiary structure: alignments of the target sequence to sequences of proteins with known structures, identified by protein fold-recognition (FR) methods. The Pcons consensus server will evaluate to which extent the FR alignments agree with each other and if a particular fold can be singled out.

From FR alignments you can automatically generate crude three-dimensional models (without variable loops). You can also select a subset of alignments that belong to a particular fold and submit them to our FRankenstein server, which will splice and recombine the crude models to produce a cute little monster protein model that hopefully resembles the true structure.

Models produced by our server are scored by COLORADO3D. Use "color by temperature factor" option in RASMOL to visualize regions likely to be correct (blue) and those that may be wrong (red). Please remember that all theoretical models must be taken with a grain of salt - and even the most confident ones should be carefully verified.

http://cubic.bioc.columbia.edu/eva/index.html

EVA

Version
May 25, 2001

email

EVA mirrors - Secondary structure Comparative modelling Threading

OBJECTIVES:

EVA continously and automatically analyses protein structure pr
details)

RESULTS:

- PDB statistics
- secondary structure
- comparative modelling
- inter-residue distances and contacts
- threading
- FTP archives

CAFASP Critical Assessment of Fully Automated Structure Prediction

PDB-CAFASP

GOAL

The goal of PDB-CAFASP, like CAFASP3 and LiveBench, is to evaluate the performance of fully automatic 3D protein structure prediction servers. The difference in PDB-CAFASP is the set of targets on which it is applied. CAFASP uses CASP targets, LiveBench uses newly released pdb entries, while PDB-CAFASP uses as targets pre-release PDB entries. Pre-release PDB entries are entries soon to be released, whose structures are not yet published, but whose sequence is known. PDB-CAFASP thus increases the size of the test set on which the servers are evaluated.

PDB-CAFASP will take place at the end of 2002. All CAFASP participants are invited to participate in PDB-CAFASP3. Further details will be announced in this site.

PDB-CAFASP: Pre-release PDB entries

CAFASP3: Evaluation of fully automated Servers

Homology Modeling software in ABCC

Commercial Software: $$$$$

- Tripos: Composer, Match-maker, GeneFold (not a HM software)
- Accelrys: Homology, Modeller, GCG

Free Software:

- SWISS-MODEL, GeneMine etc.

Other software

- SwissModel
- LOOK (Free software)
- ProFit
- Modeller $$$
- Prime (Schrodinger) $$$
- Sybyl $$$
- ICM $$$
- MOE $$$

Some selected recent publications

- Sequence comparison and protein structure prediction Roland L Dunbrack Jr. Cur. Opin. Str. Biol, **2006**, 16:374
- A Comparative Study of Available Software for high-accuracy homology modeling: From sequence alignments to structural models, Akbar Nayeem et al, Protein Science, **2006** 15:808
- Protein Families and their evolution-A structural perspective, C.A. Orengo and J.M. Thornton, Annu. Rev. Biochem, **2005**, 74, 867-900
- Servers for protein structure prediction, D. Fischer, Curr Opin. Str. Biology, **2006**, 16:179
- In Quest of an empirical potential for protein structure prediction, J. Skolnick, Curr. Opin Str. Biology, **2006**, 16:166
- Comparative modeling for protein structure prediction, Krzysztof Ginalski, Cur. Opin Str. Biology, **2006**, 16:172-177
- Sequence Similarity, talk by Prof. W.R. Pearson, **2006**

Some selected recent publications

- Predicting protein function from sequence and structural data, J.D. Watson, R.A. Laskowski and J.M. Thornton, Cur. Opin Strl. Biology, **2005**, 15:275-284
- Application and Limitations of X-ray crystallographic Data in Structure-based Ligand and Drug Design, Andrew M. Davis, Simon J Teague and Gerard J. Kleywegt, Angew. Chem. Int. Ed **2003**, 42, 2718-2736
- Where did the BLOSUM62 alignment score matrix come from? Sean R. Eddy, Nature Biotechnology, 2004, 22(8) 1035-1036.
- From protein structure to biochemical function? R.A. Laskowski, J.D. Watson and J.M. Thornton, J Str and Funct. Genomics, **2003** 4:167
- Sequence and Structure differences between enzyme and nonenzyme homologs, A.E. Todd, C.A. Orango and J.M. Thornton, Structure, **2002**, 10:1435-1451
- Validation of protein crystal structures, G. J. Kleywegt, Acta Crystallographica Sec D, **2000**, D56, 249

Reference

- Visit the web-site

http://ncisgi.ncifcrf.gov/~ravichas/HomMod/