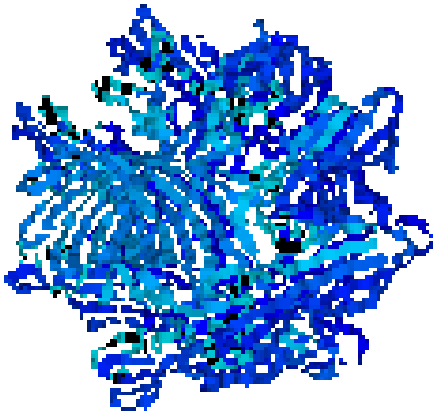


Animation using pdb2mgif

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
10/20/2004

Advanced Biomedical Computing Center

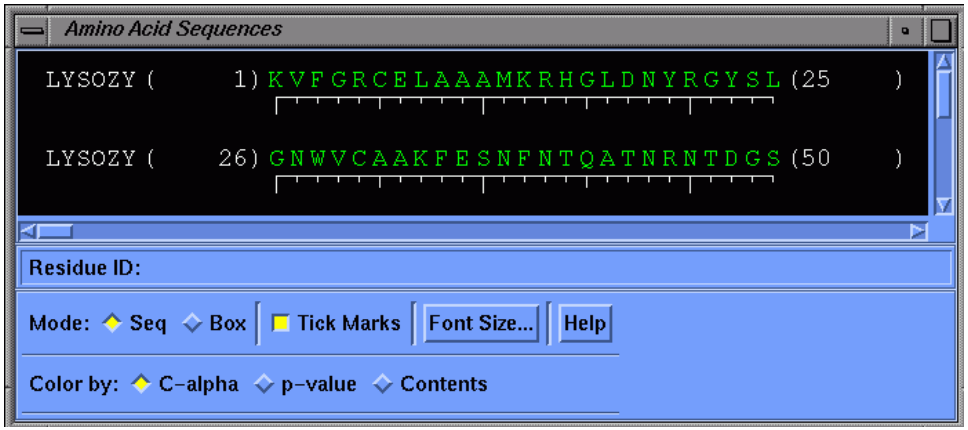
- Supercomputing facility located at NCI-Frederick (Bldg 430)
  - NCI, NIH, NIA ……
- What do we do?
  Consultation, Training, Research
- Biomedical Research groups at ABCC
  - Quantum-Mechanics, Molecular Modeling, Bioinformatics, Structural Biology…
- Web: http://www-fbsc.ncifcrf.gov

Overview

- Basics of Homology Modeling
- Hands-on exercise
  - Homology Modeling using Sybyl
  - 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 the Sybyl (Composer) and InsightII (homology) modules

Overview of Homology Modeling

Sequence
from experiment



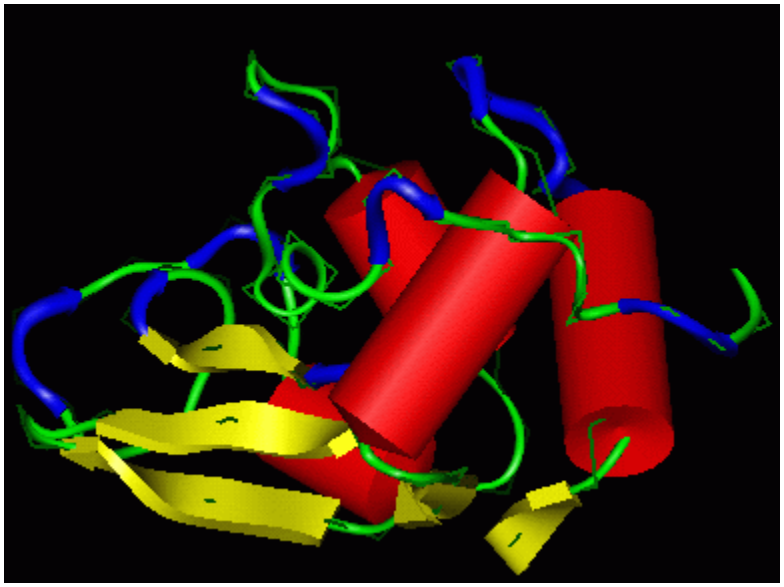
Experiments

X-ray, NMR, e-Diffraction

Physicochemical
Simulations

Comparative Modeling
Knowledge-Based

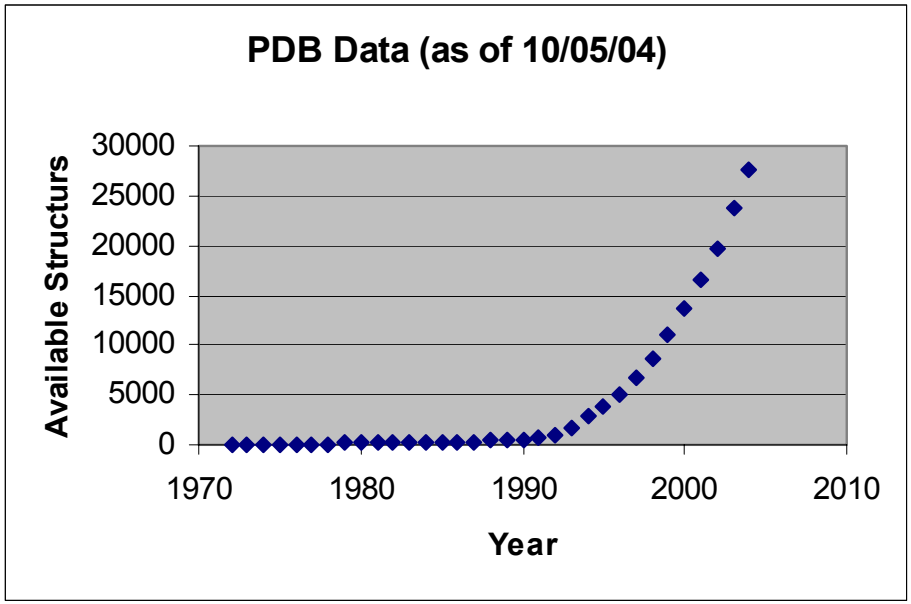
Modeling



3D-Structure

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
    - Swiss-Prot Release 44.7 as of 10/11/04 162,781 entries
      - (Release 42.4 had 138,347 entries 11-14-03)
    - PDB as of 10/12/04 has 27,663 structures
      - PDB: 23,188 (11-11-03);



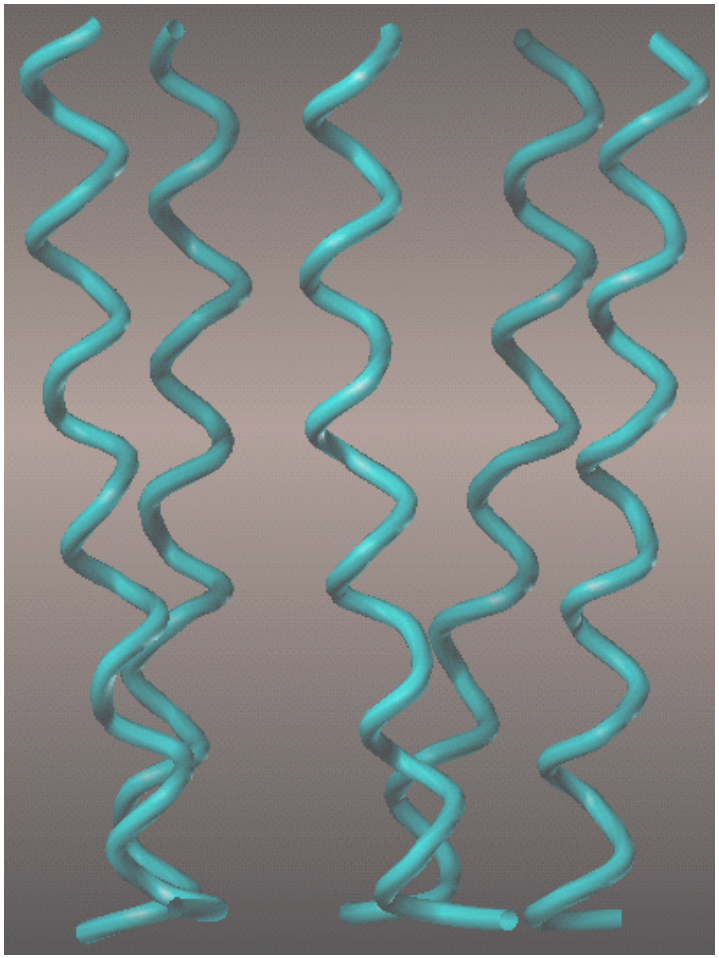
Why Homology Modeling?

- 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
- Structural information enhances our understanding of protein-protein or protein-DNA interactions

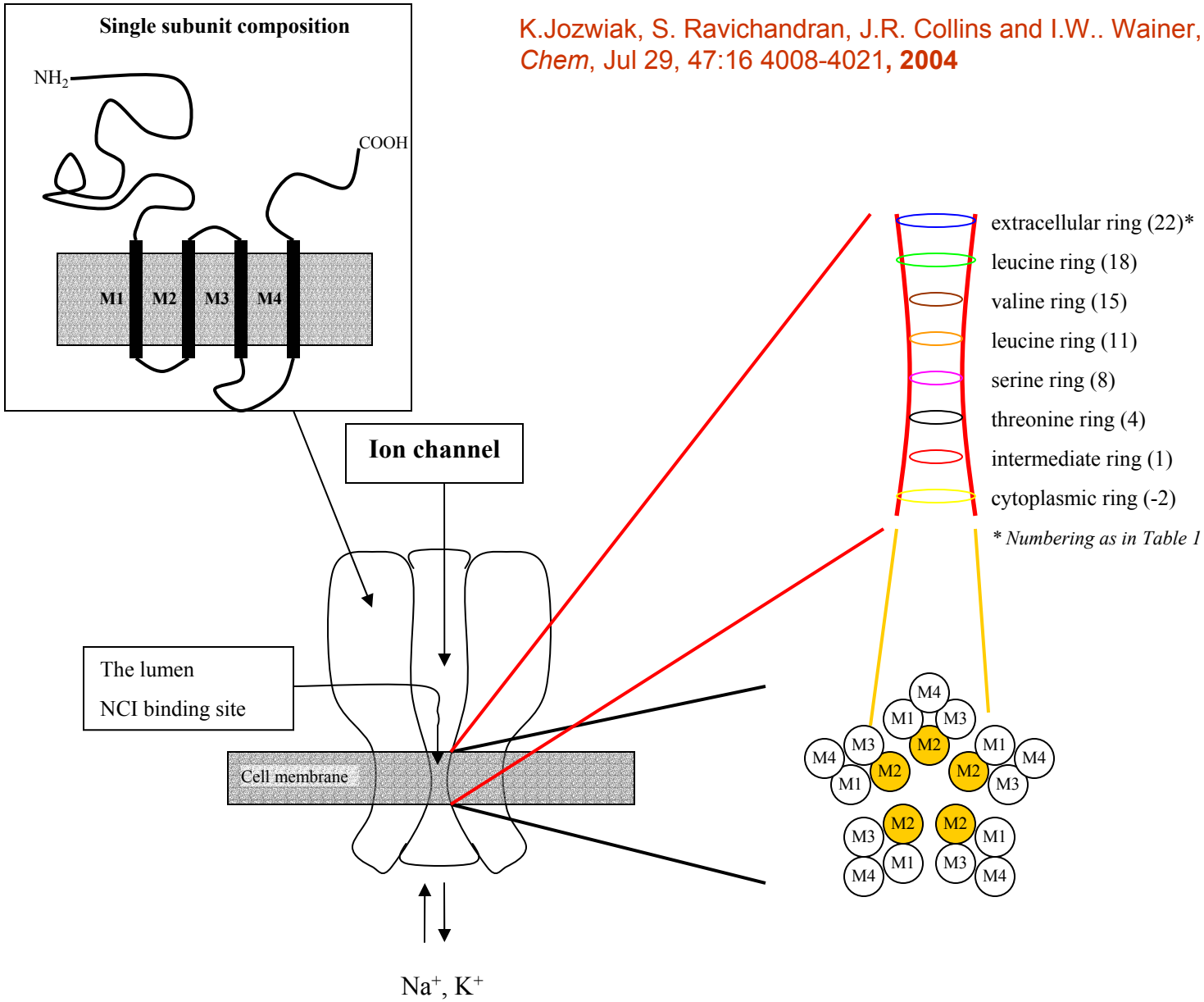
K.Jozwiak, S. Ravichandran, J.R. Collins and I.W.. Wainer, *J. Med. Chem*, Jul 29, 47:16 4008-4021**, 2004**

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:
  - "Having a common evolutionary origin"
    - Evolved 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 Eukoryates, rats and humans

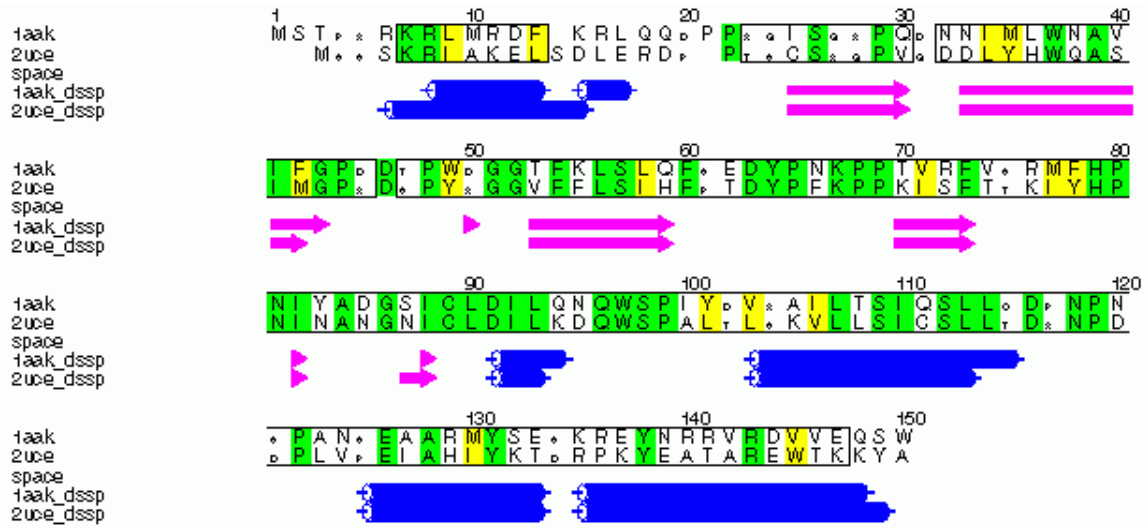
How to identify homologous proteins? Can we exploit sequence information?

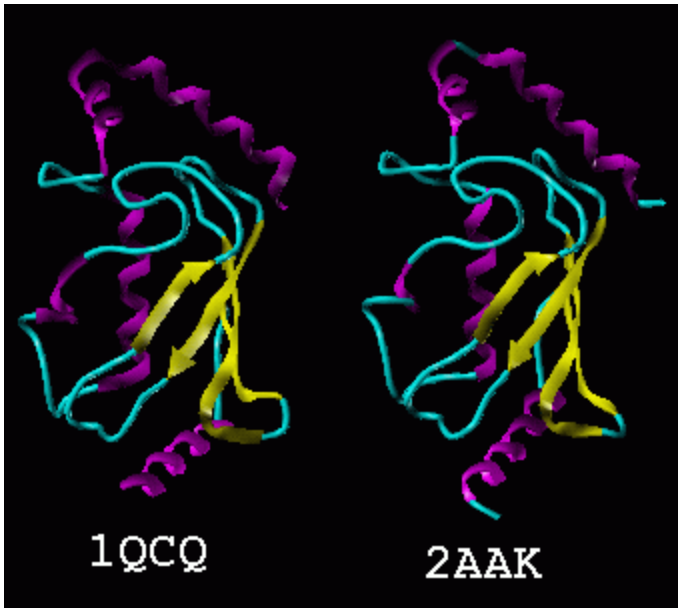
Comparing Homologous enzymes

Family: Ubiquitin Conjugating enzyme

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

Sequence Identity 43%





Using PredictProtein

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

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

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 RMSD 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

NMR & X-ray

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

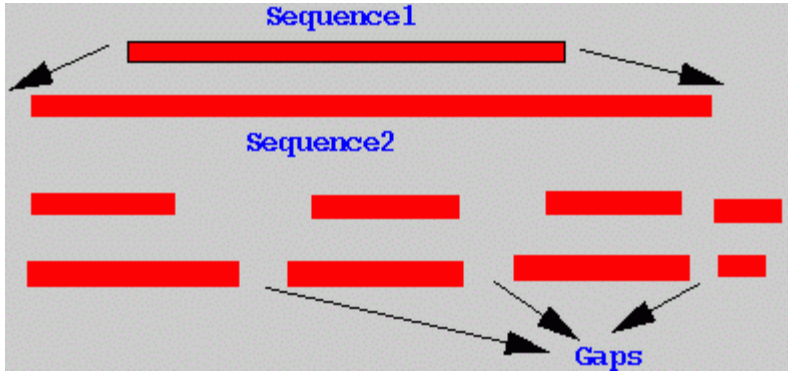
- X-ray
  - Static
    - Only one model
  - Crystal
  - Limitations
    - Not limited by size
- Examples
  - 7LYZ

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



Global Alignment

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

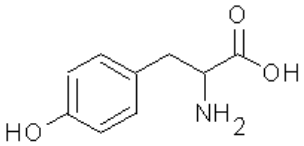
- Blosum (BLOcks SUbstitution Matrix) Heinkoff & Heinkoff (1992)
  - Derived from BLOCKS database
  - Distant relationships explained better than PAM
  - Blosum62 obtained from sequence BLOCKS clustered at >= 62% identity
  - BlosumX are observed from actual alignments not extrapolated
  - Sequences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

PAM250 Matrix (identities at 20% level)

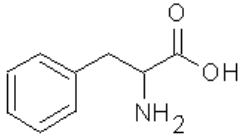
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	.
Ala	2	-2	0	0	-2	0	0	1	-1	-1	
Arg	-2	6	0	-1	-4	1	-1	-3	2	-2	
Asn	0	0	2	2	-4	1	1	0	2	-2	
Asp	0	-1	2	4	-5	2	3	1	1	-2	
Cys	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	
Gln	0	1	1	2	-5	4	2	-1	3	-2	
Glu	0	-1	1	3	-5	2	4	0	1	-2	
Gly	1	-3	0	1	-3	-1	0	5	-2	-3	
His	-1	2	2	1	-3	3	1	-2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

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)

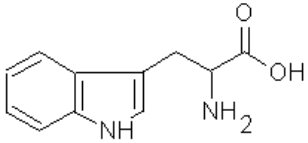
Cystein: Well-known for S-S linkage Important for structure



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	0	1

Searching for Templates

- Do a Blast/Fasta or use programs within GCG (Align, gap, bestfit, etc.) for sequence alignment. 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
- 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 Sequence Alignments, secondary structure elements
  - What happens when we have only one template?
- The next step is to align the Structurally aligned templates with the unknown sequence
  - No gaps are allowed within the SCR regions
    - Special sequence alignment algorithm used which discourages gaps within SCR.

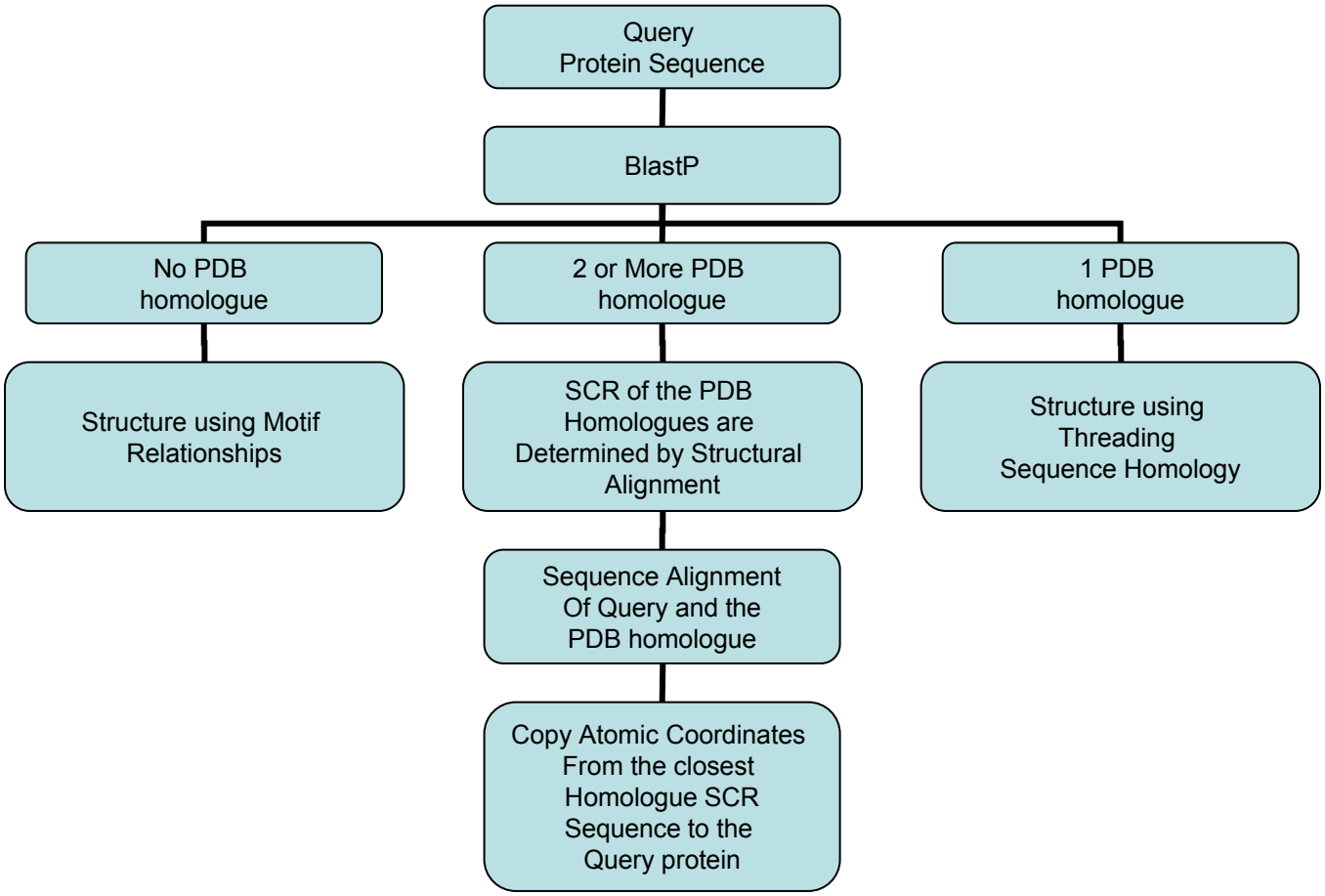
Structurally Variable Region (SVR) Modeling (3 methods)

- 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
- Mostly MM calculations carried out as the last step

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



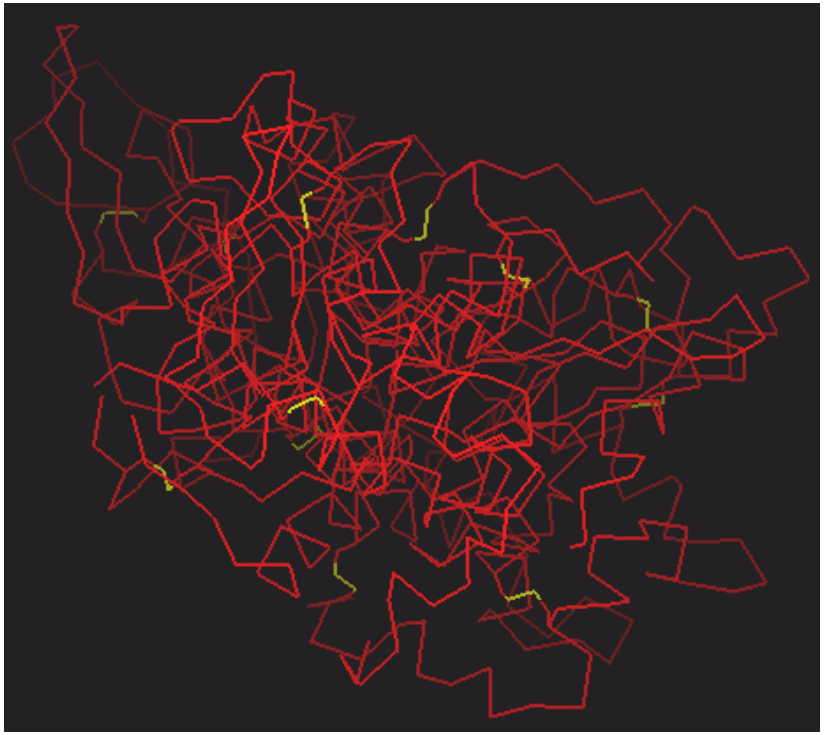
Bioinformatics Basics:
Rashidi & Buehler

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



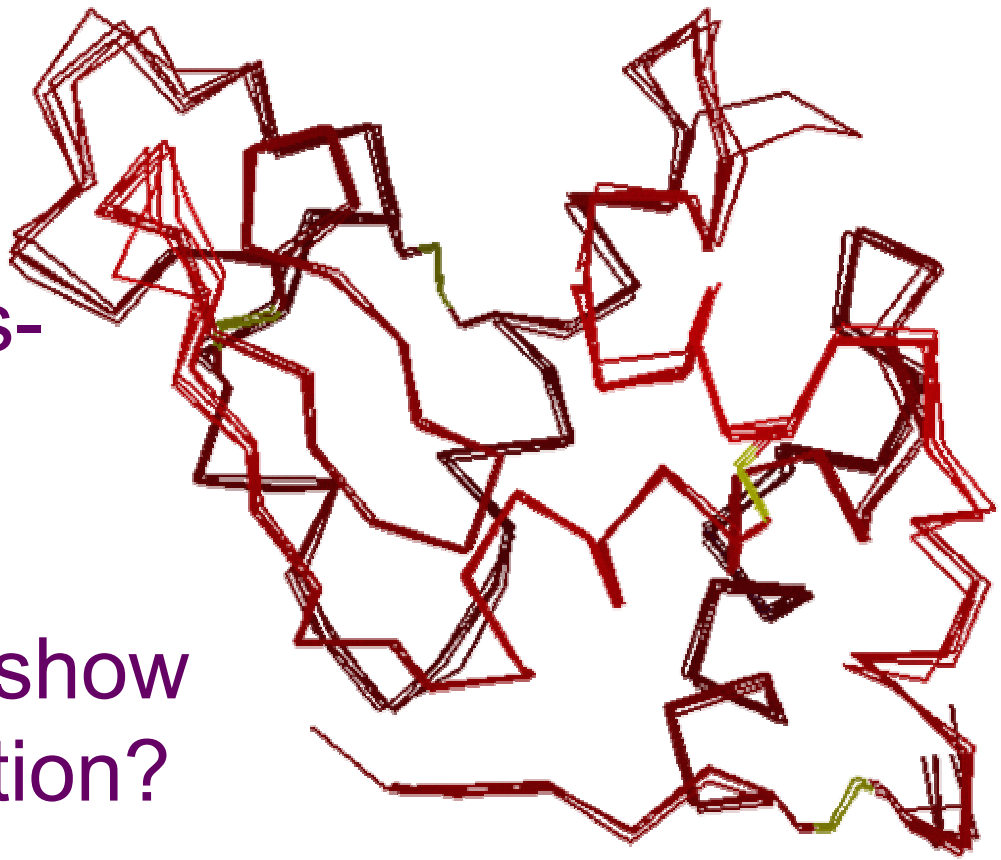
```
  L (     1) KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHA (42
IHL (     1) KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFESNFNTQA (42
HEL (     1) KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQA (42
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTG (A41
LZ1 (     1) KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRA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

Studying the corrected template alignment

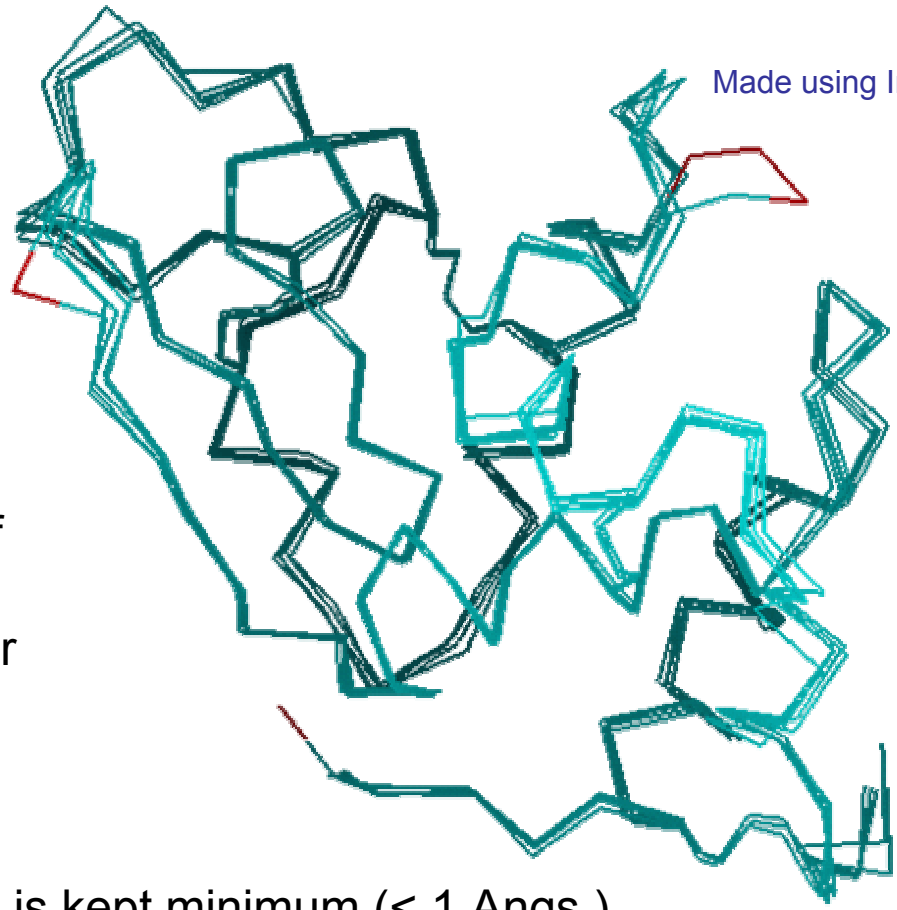
- Look at cys

  How about the Structural Cons-veration?

- Which regions show structural variation?



Structurally corrected MSA



Made using InsightII, Accelrys

	A	B
A	0	0.5
B	0.5	0

Do you see the location of the variable region (core or 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
```

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

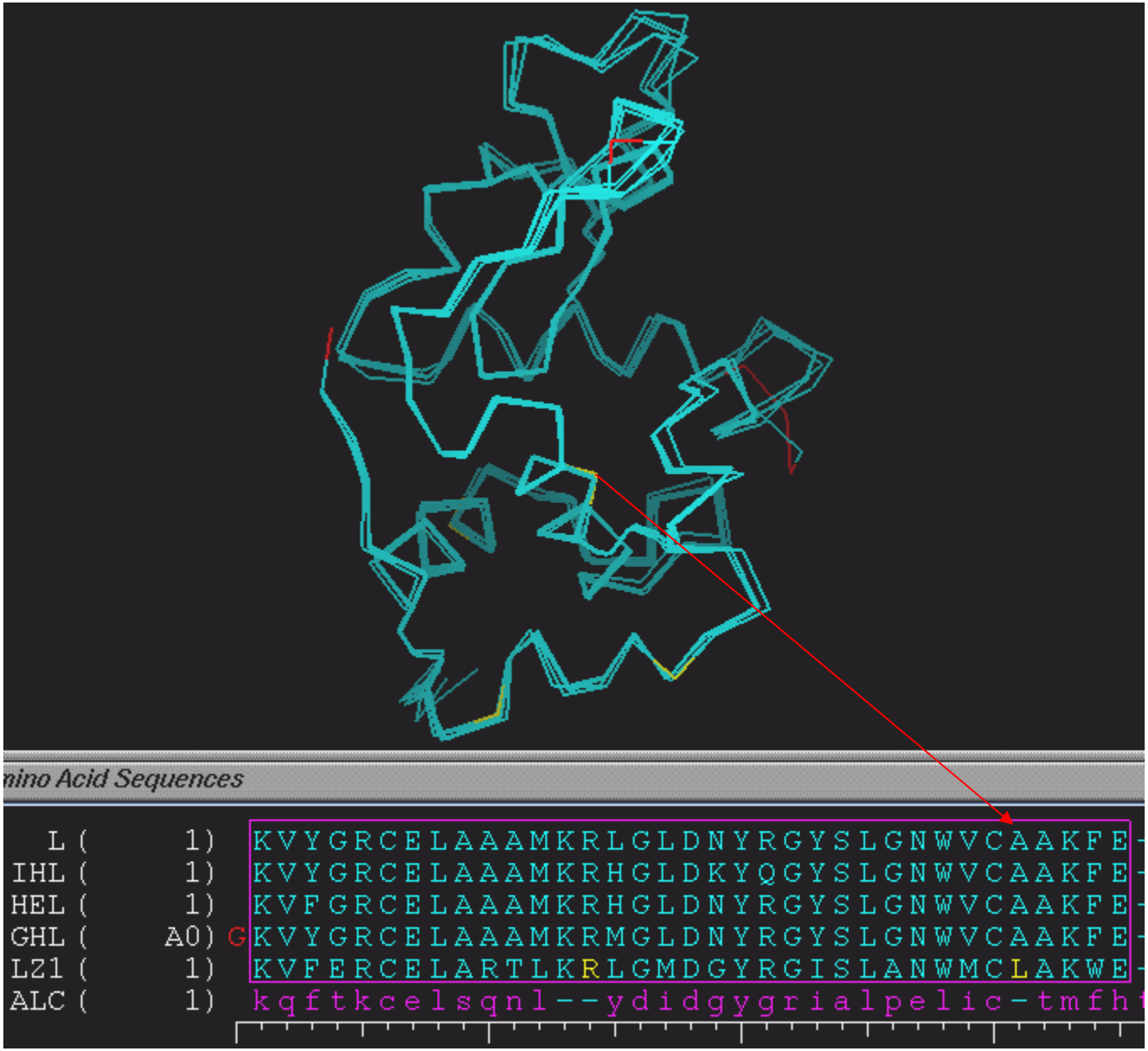
```
  L (     1)  KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFE-SNFNTH (41
IHL (     1)  KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFE-SNFNTQ (41
HEL (     1)  KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFE-SNFNTQ (41
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFE-SNFNTG (A41
LZ1 (     1)  KVFERCELARTLKRLGMDGYRGISLANWMCLAKWE-SGYNTR (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 sequence template</u> to the model sequence.
  - Other alternative is “Distance Geometry” approach


mino Acid Sequences
L (1) KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFE
IHL (1) KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFE
HEL (1) KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFE
GHL (A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFE
LZ1 (1) KVFERCELARTLKRLGMDGYRGISLANWMCLAKWE
ALC (1) kqftkcelsqnl--ydidgygrialpelic-tmfh

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
```

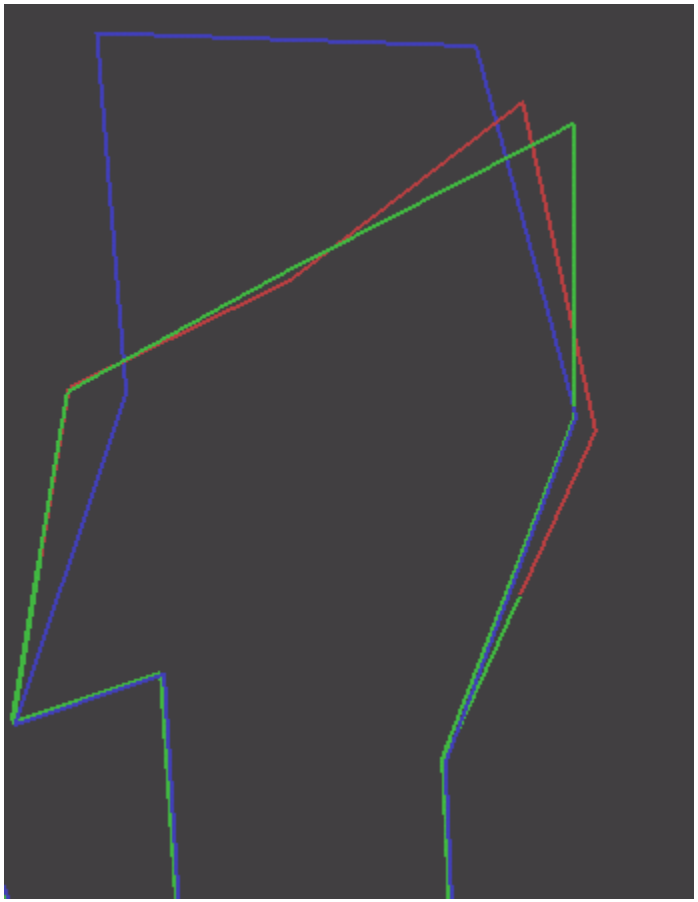
2

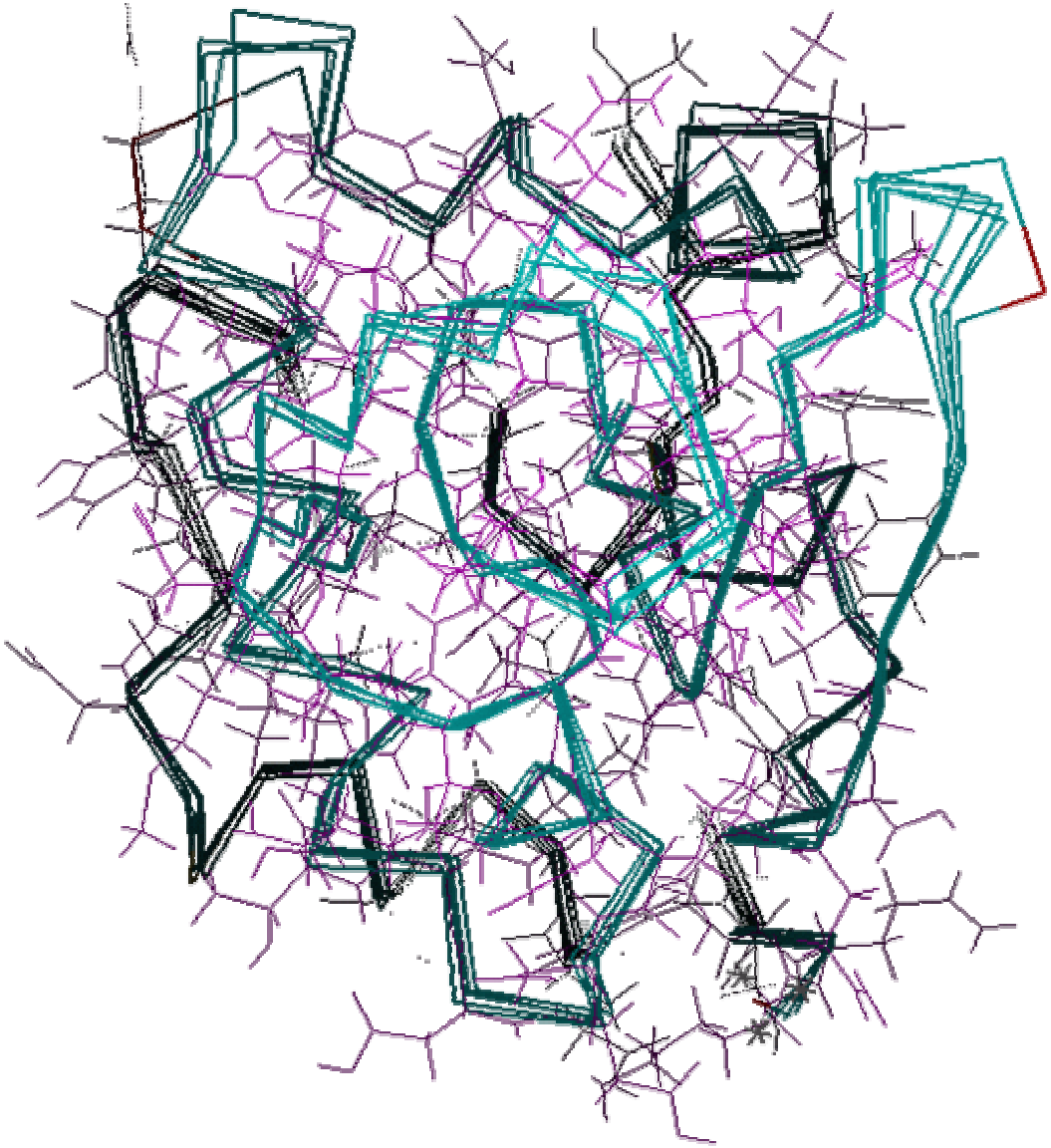
```
  L (     1)  KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFE-SNFNTH (41
IHL (     1)  KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFE-SNFNTQ (41
HEL (     1)  KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFE-SNFNTQ (41
GHL (    A0) GKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFE-SNFNTG (A41
LZ1 (     1)  KVFERCELARTLKRLGMDGYRGISLANWMCLAKWE-SGYNTR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

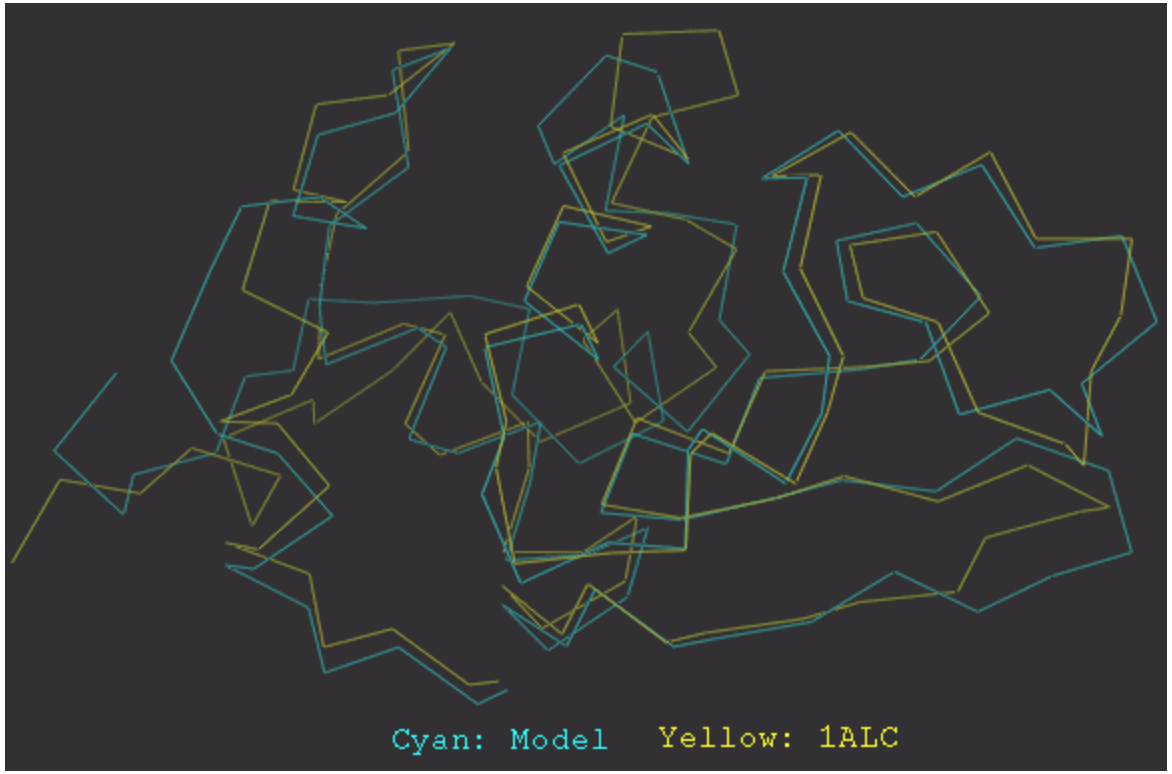




Side Chains will be added if the template has identical residues

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
  - One option is to look at Luzzati plot
  - 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)
  - Swiss-Model uses GROMOS
- How many steps of Minimization ?
  - Vacuum (non-solvent)

Identifying Incorrect Models

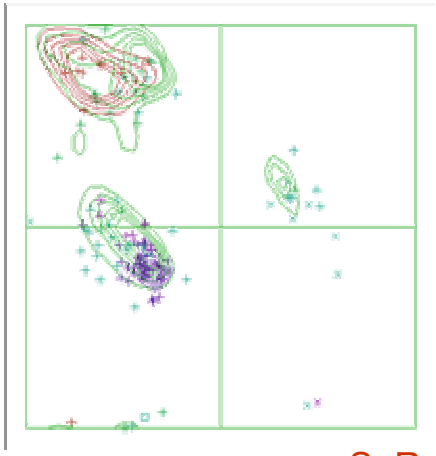
- Hydrophobic residues exposed
- Buried polar or ionic residues without the charges satisfied (H-bonds, salt-bridge etc)
- Clashes
- Unusual bond-lengths, bond-angles
- Sequence alignment is not-optimal
- Very large RMSD among the templates

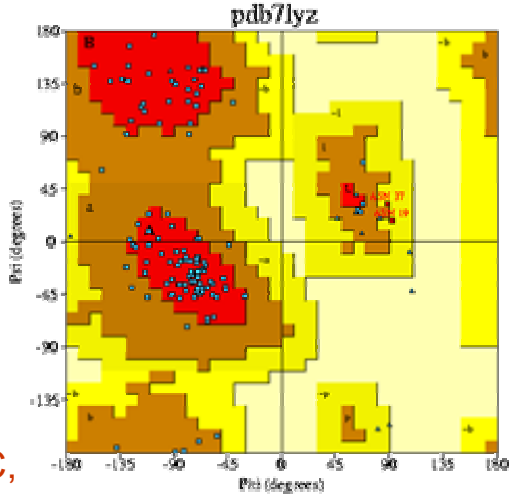
Quality of Models

- 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





CASP: Test of the Models

- Critical Assessment of Techniques for Protein Structure

  http://predictioncenter.llnl.gov/

  - 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 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 used 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

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

Reference

- Visit the web-site

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