

HOMOLOGY MODELING

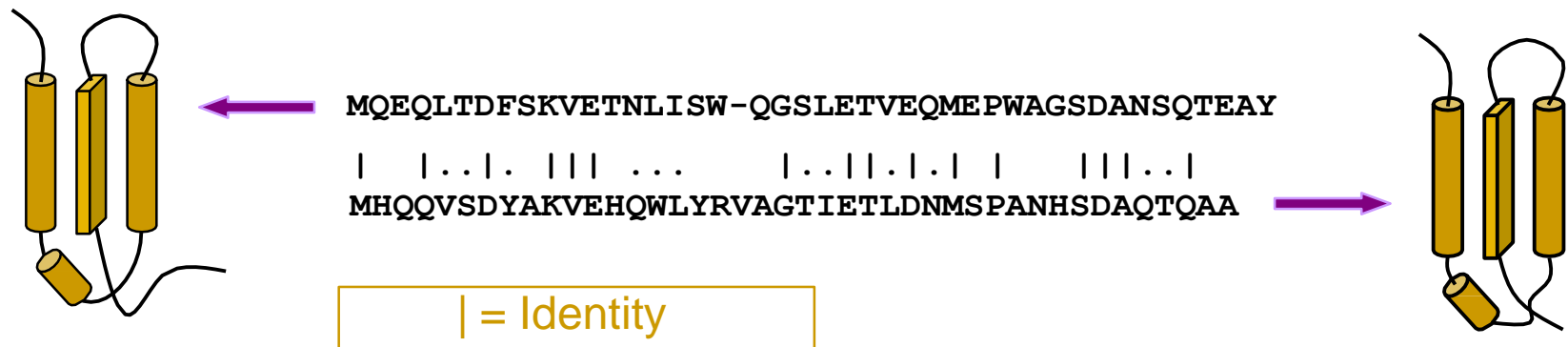
APARNA PATIL KOSE
LECTURER.

Homology Modeling

- Presentation
- Steps
- Testing methods

What is Homology Modeling?

- To predict the 3D-structure of a unknown protein (TARGET) based on the known experimental structure of a related homologous protein (TEMPLATE):
 - Search structure databases for *homologous* sequences
 - Transfer coordinates of known protein onto unknown



Why Modeling/Computational approach is Necessary?

- Current structure determination methods:
 - NMR technology – limited size (30 KD).
 - X-Ray Crystallography – proteins do not / difficult to crystallize.
 - Expensive, limited, difficult, time consuming.
 - Techniques involve elaborate technical procedures.
- Can't keep up with growth rate of sequence databases:
 - PDB: 40132 structures (14/11/2006)
 - pairsdb: 4000000+ sequences (11/2006)

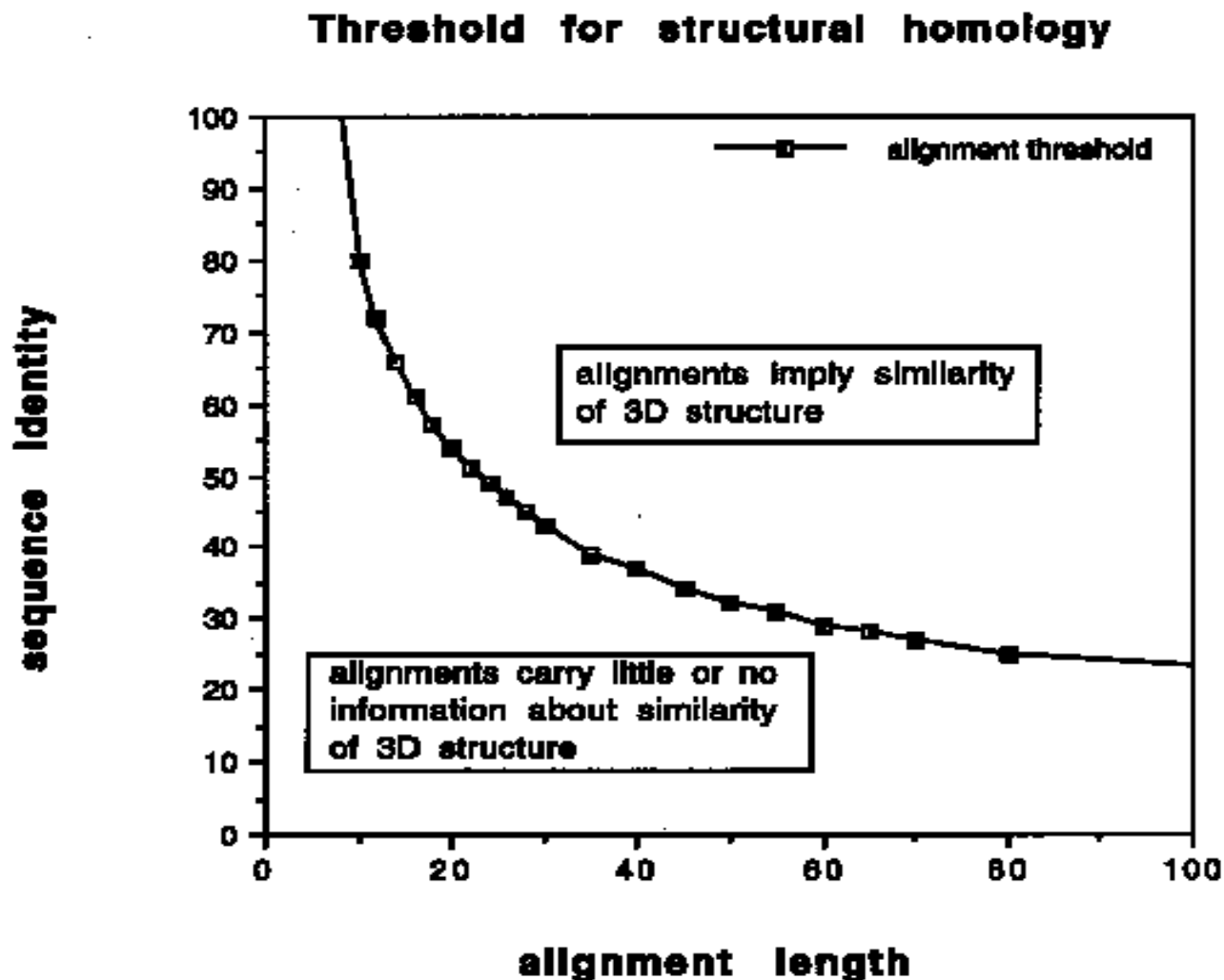
- ❑ The goal of research in the area of structural genomics is to provide the means to characterize and identify the large number of protein sequences that are being discovered.
- ❑ Knowledge of the three-dimensional structure
 - can be of great importance for the design of drugs
 - greatly enhances our understanding of how proteins function and how they interact with each other.
- ❑ **With a better computational method this can be done extremely fast.**

Modeling methods

To predict or model the three dimensional structure of proteins, the 3 different methods used are:

- Homology modeling
- Threading
- Ab initio methods

What can be Modelled ?



What can be Modelled ?

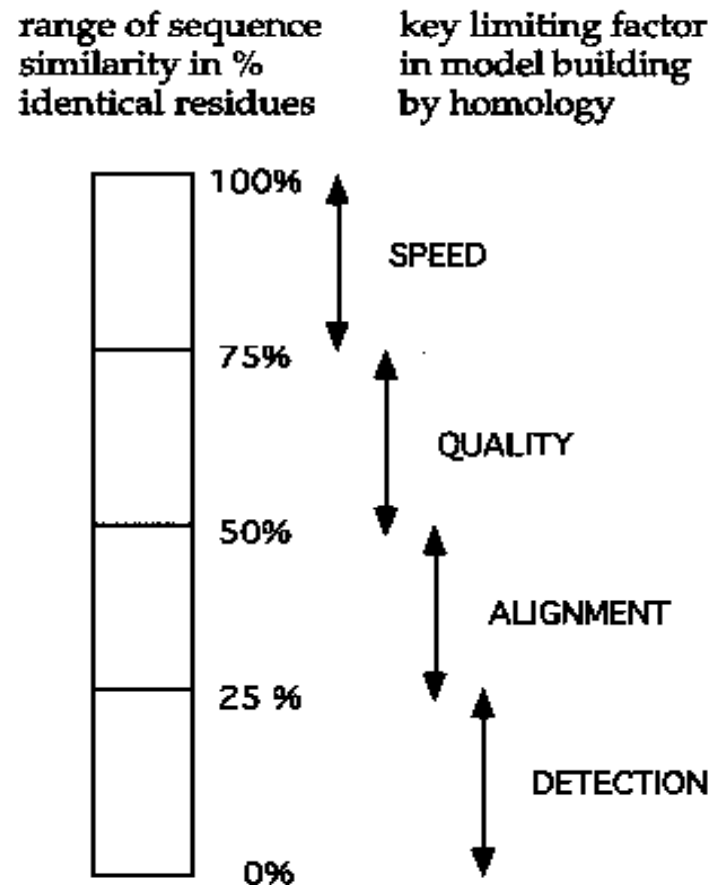
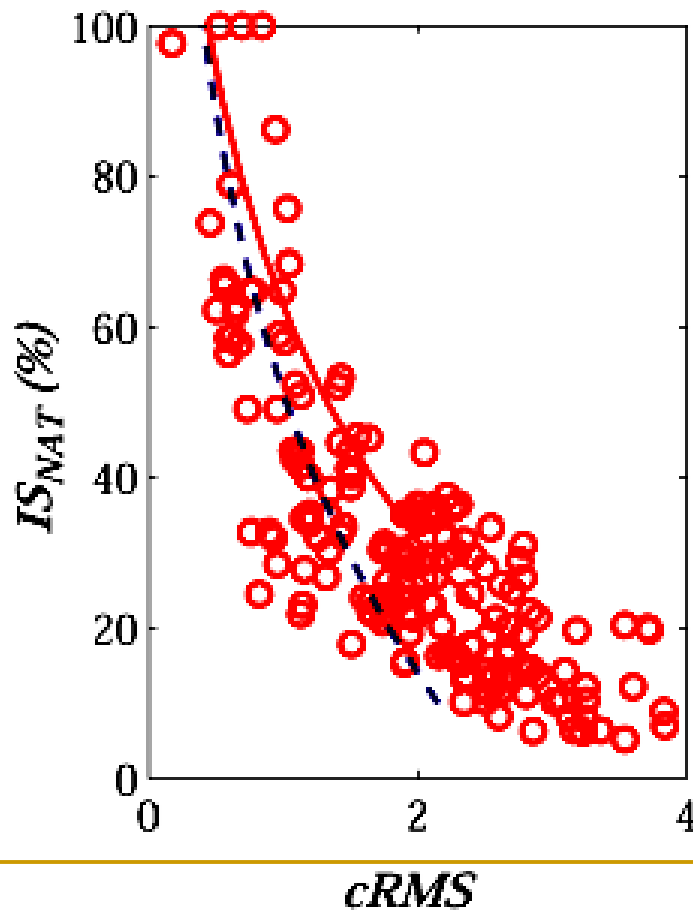


Figure 1. The main limiting steps for model building by homology as function of the percentage sequence identity between the structure and the model.

Homology Modeling: why it works

Native Sequences

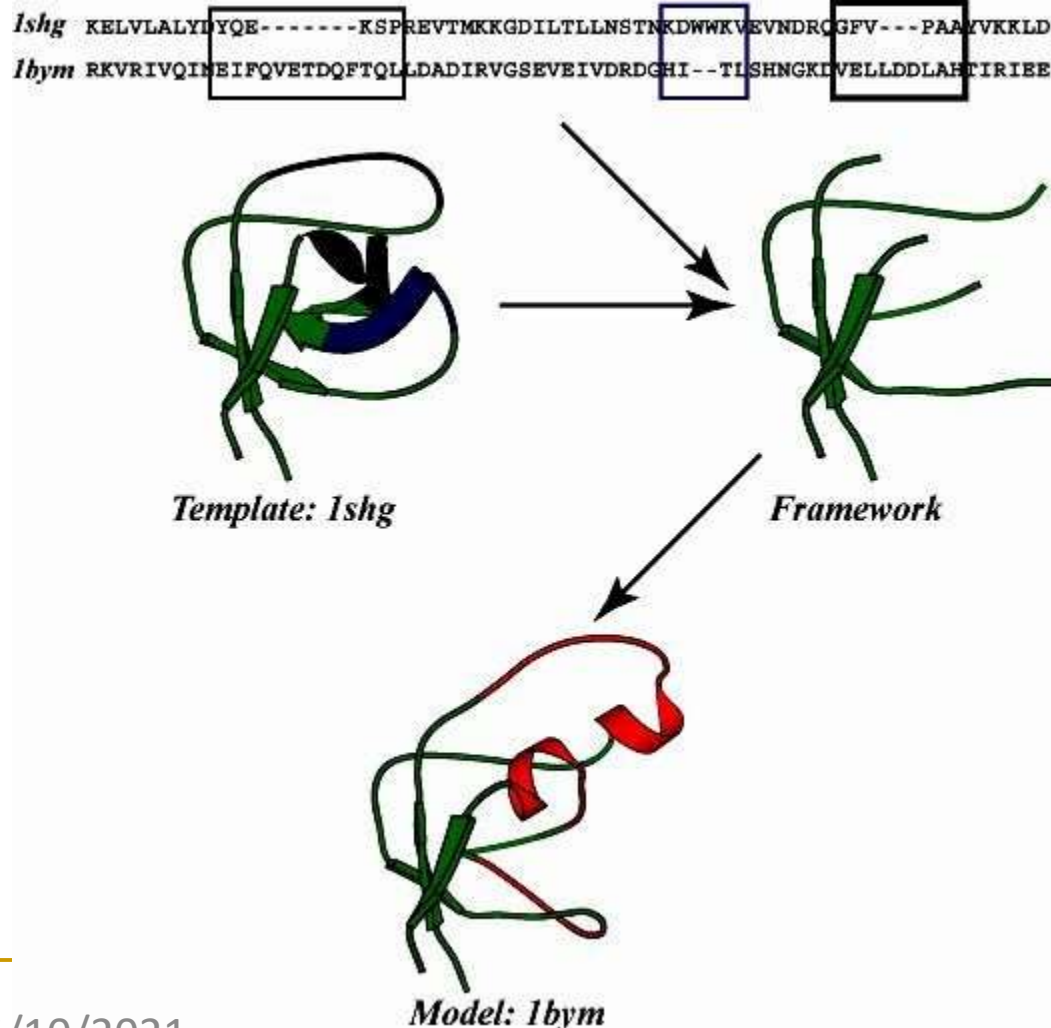


High sequence identity



High structure similarity

Homology Modeling: why it works

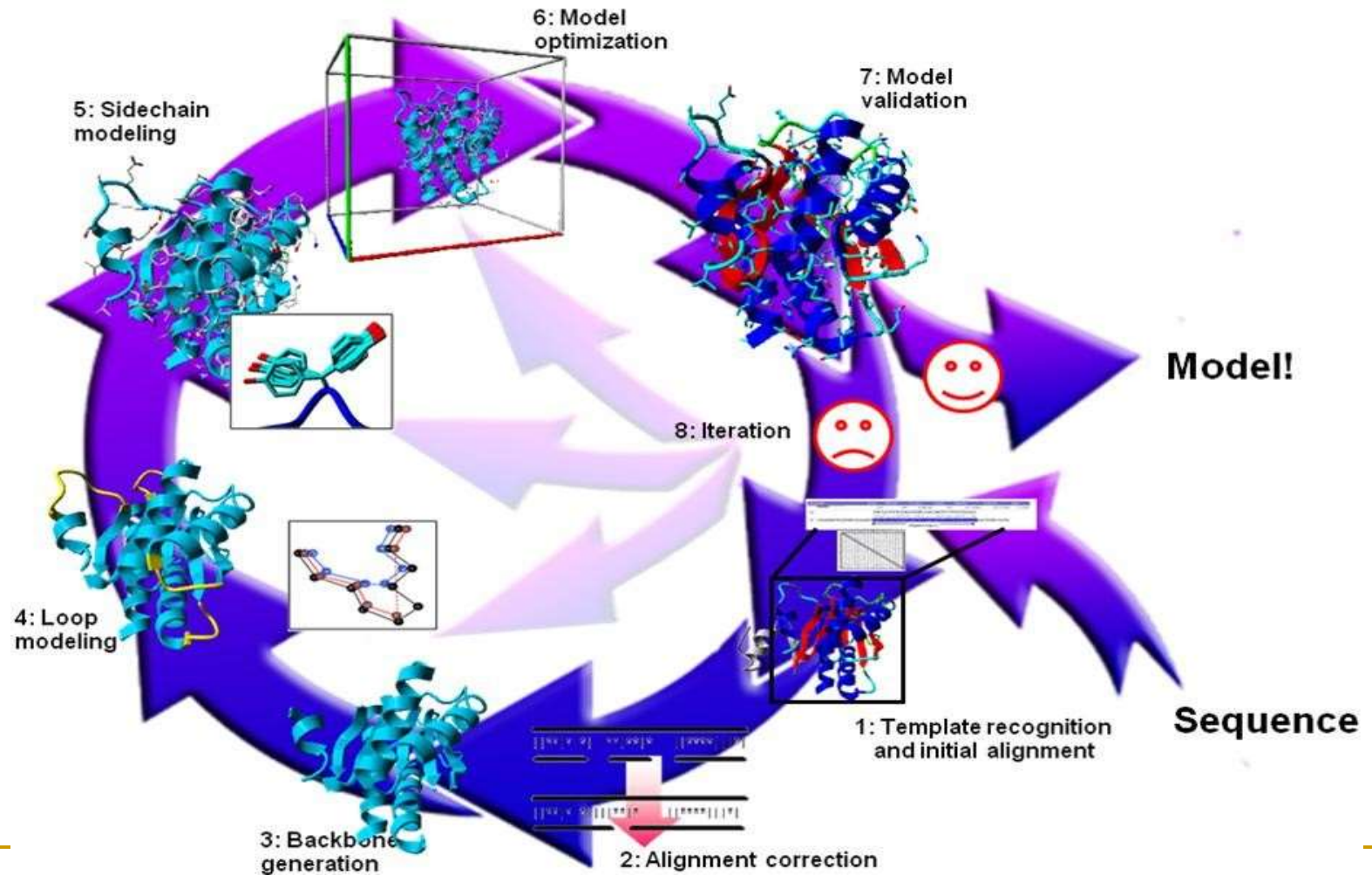


- o Find template
- o Align target sequence with template
- o Generate model:
 - add loops
 - add sidechains
- o Refine model

The '7' Steps of Modeling

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side chain modeling
6. Model optimization
7. Model Validation

Overview:

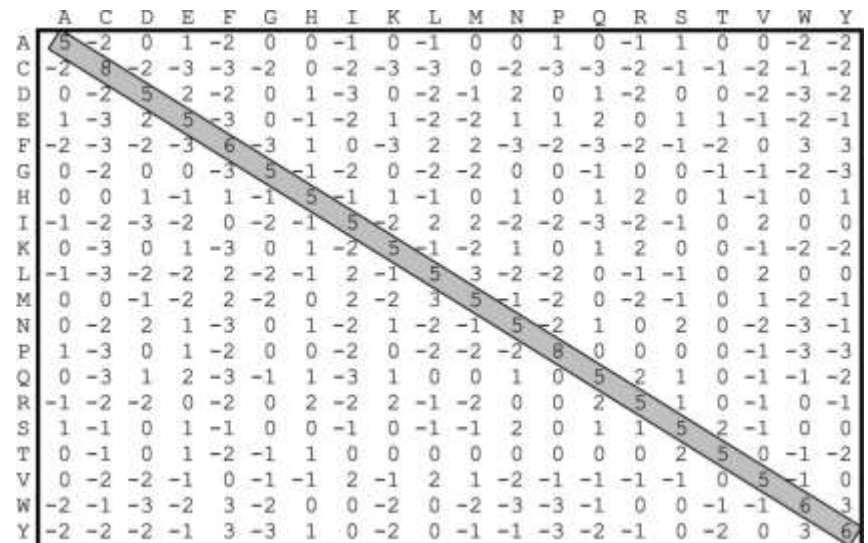


1. Template detection

- Sequences in safe homology modeling zone - share high % identity.
- BLAST- sequence alignment program.
- Involves comparing query seq. with all seqs. of known PDB structures.
- Uses 2 matrices.
- Template selected- Highest sequence similarity and other considerations.

A residue exchange matrix

- Defines the likelihood that any 2 of 20 amino acids ought to be aligned.
- Similar physicochemical properties- Better score.
- Conserved residues- Highest score.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	5	-2	0	1	-2	0	0	-1	0	-1	0	0	1	0	-1	1	0	0	-2	-2
C	-2	5	-2	-3	-3	-2	0	-2	-3	-3	0	-2	-3	-3	-2	-1	-1	-2	-1	-2
D	0	-2	5	2	-2	0	1	-3	0	-2	-1	2	0	1	-2	0	0	-2	-3	-2
E	1	-3	2	5	-3	0	-1	-2	1	-2	-2	1	1	2	0	1	1	-1	-2	-1
F	-2	-3	-2	-3	5	-3	1	0	-3	2	2	-3	-2	-3	-2	-1	-2	0	3	3
G	0	-2	0	0	-3	5	-1	-2	0	-2	-2	0	0	-1	0	0	-1	-1	-2	-3
H	0	0	1	-1	1	-1	5	-1	1	-1	0	1	0	1	2	0	1	-1	0	1
I	-1	-2	-3	-2	0	-2	-1	5	-2	2	2	-2	-2	-3	-2	-1	0	2	0	0
K	0	-3	0	1	-3	0	1	-2	5	-1	-2	1	0	1	2	0	0	-1	-2	-2
L	-1	-3	-2	-2	2	-2	-1	2	-1	5	3	-2	-2	0	-1	-1	0	2	0	0
M	0	0	-1	-2	2	-2	0	2	-2	3	5	-1	-2	0	-2	-1	0	1	-2	-1
N	0	-2	2	1	-3	0	1	-2	1	-2	-1	5	-2	1	0	2	0	-2	-3	-1
P	1	-3	0	1	-2	0	0	-2	0	-2	-2	-2	5	0	0	0	0	-1	-3	-3
Q	0	-3	1	2	-3	-1	1	-3	1	0	0	1	0	5	2	1	0	-1	-1	-2
R	-1	-2	-2	0	-2	0	2	-2	2	-1	-2	0	0	2	5	1	0	-1	0	-1
S	1	-1	0	1	-1	0	0	-1	0	-1	-1	2	0	1	1	5	2	-1	0	0
T	0	-1	0	1	-2	-1	1	0	0	0	0	0	0	0	0	2	5	0	-1	-2
V	0	-2	-2	-1	0	-1	-1	2	-1	2	1	-2	-1	-1	-1	-1	0	5	-1	0
W	-2	-1	-3	-2	3	-2	0	0	-2	0	-2	-3	-3	-1	0	0	-1	-1	5	3
Y	-2	-2	-2	-1	3	-3	1	0	-2	0	-1	-1	-3	-2	-1	0	-2	0	3	5

An alignment matrix

- Axes of matrix corresponds to 2 sequences to be aligned.
- Values from residue matrix.

		G	T	G	T	A	T	A	T	A	T	C	T	G
1)	G	-	-	-	-	-	-	-	-	-	-	3	-	-
	A	-	2	-	2	-	2	-	2	-	2	-	5	-
	C	3	-	5	0	-	-	-	-	-	-	-	-	8
	T	-	-	0	1	2	-	2	-	2	-	-	-	3
	A	-	2	-	2	-	4	-	4	-	4	-	2	-
	T	-	-	-	-	4	-	6	1	6	1	-	-	-
	A	-	2	-	2	-	6	1	8	3	8	3	2	-
	T	-	-	-	-	4	1	8	3	10	5	4	-	-
	A	-	2	-	2	-	6	3	10	5	12	7	6	1
	T	-	-	-	-	4	1	8	5	12	7	8	3	2
	G	-	-	-	-	-	-	3	4	7	8	10	5	-
	A	-	2	-	2	-	2	-	2	3	5	12	7	-
	G	-	-	-	-	-	-	-	-	-	6	7	-	-

2. Alignment

- Run BLAST on model & template sequence.
- Its difficult to align two sequences with low sequence identity.
- Do multiple sequence alignment.
- Keep only model and template
 - to characterize protein families
 - identify shared regions of homology,
 - to place insertions or deletions in strongly divergent areas.
 - Provides basis of the transforming of coordinates from the reference to the model.

How to align:

ASASASASASAS

YPYPYPYPYPYP

**(Poor alignment
possibilities....)**

How to align:

ASASASASASAS-
AYAYAYAYAYAY-
-YPYPYPYPYPYP

(MSA...)

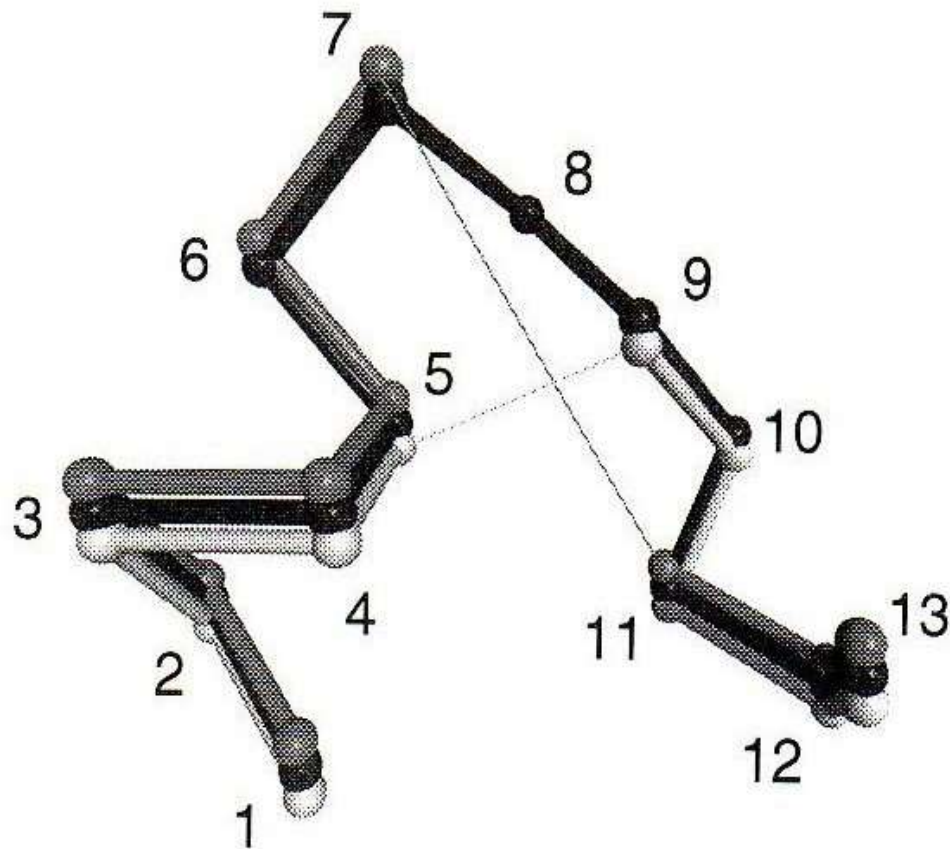
Alignment optimization

		1	2	3	4	5	6	7	8	9	10	11	12	13
Template		PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL
Model (bad)	1	PHE	ASN	VAL	CYS	ARG	ALA	PRO	---	---	---	GLU	ALA	ILE
Model (good)	2	PHE	ASN	VAL	CYS	ARG	---	---	---	ALA	PRO	GLU	ALA	ILE

Sequence alignment with 3-residue deletion.

- Alignment 1: Highest score (considering only sequence) but big gap in structure.
- Alignment 2: Better alignment, small gap- can be accommodated by small backbone shifts.

Alignment optimization (contd...)



3. Backbone generation

- Alignment ready- model building starts.
- Uses known structurally conserved regions to generate coordinates for the unknown
- Template-based fragment assembly.
 - i. Choose template with fewest errors.
 - ii. Find structurally conserved regions.
 - iii. Build model core.
- Multiple template modeling: combine good parts of both templates in a model.

4. Loop Modeling

- Alignment process involves gaps, either in model or template (insertions/deletions) sequence- implying conformational change of model (hole).
- Changes often do not occur in regular secondary structure elements.
- Loop conformations- hard to predict.
- The process involves assignment of coordinates in loop or variable region.
- Database search for segments from known protein structures fitting fixed end-points.

3 main approaches to model loops:

- Knowledge based
- In between or Hybrid based
- Energy based

(i) Knowledge based approach:

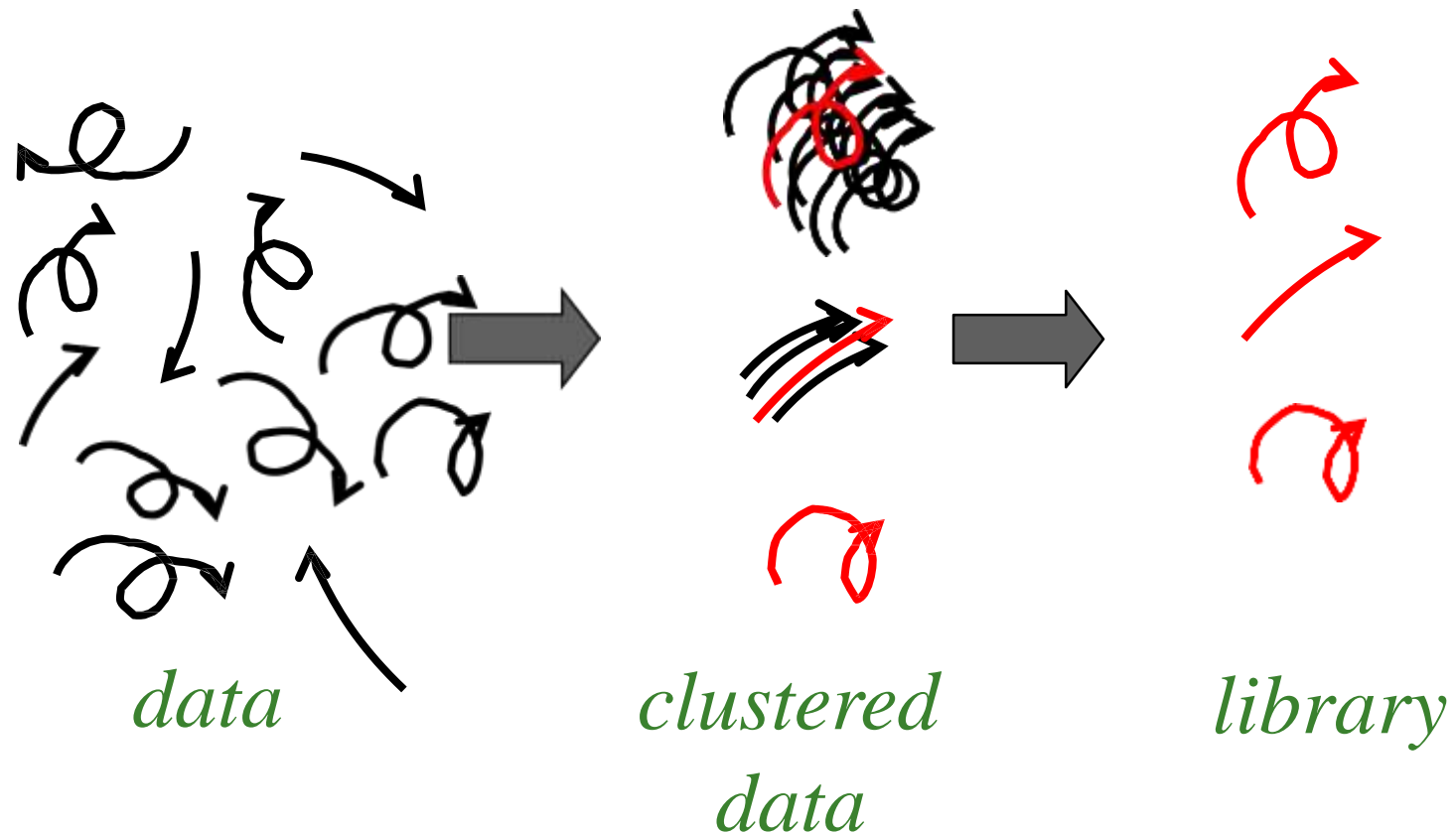
- Search through PDB for known loops.
- Scan database and search protein fragments with correct number of residues and correct end-to-end distances.
- Identified coordinates of loop transferred.
- Supported by: Modeller, 3D-Jigsaw, What if...

(ii) In Between or Hybrid based:

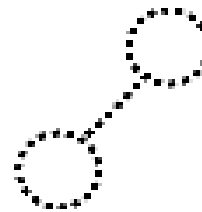
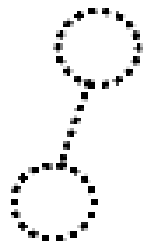
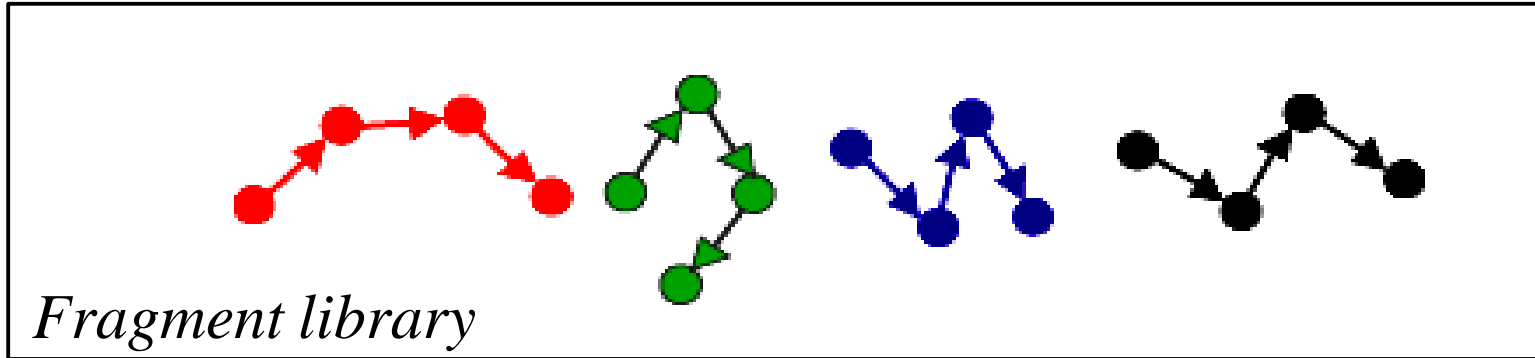
- A fragment-based approach.
- Loops- small fragments- separately compared to PDB.
- Rosetta method.

Long loops: A fragment-based approach

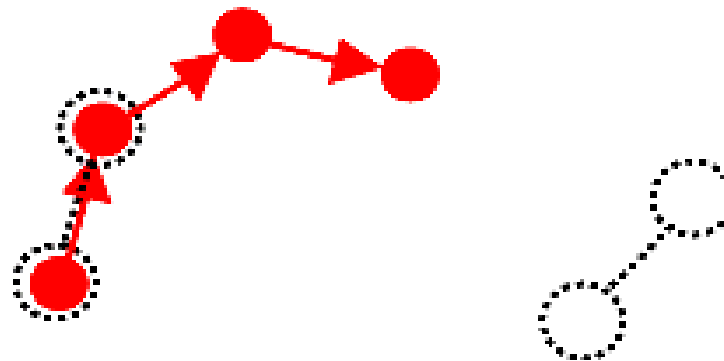
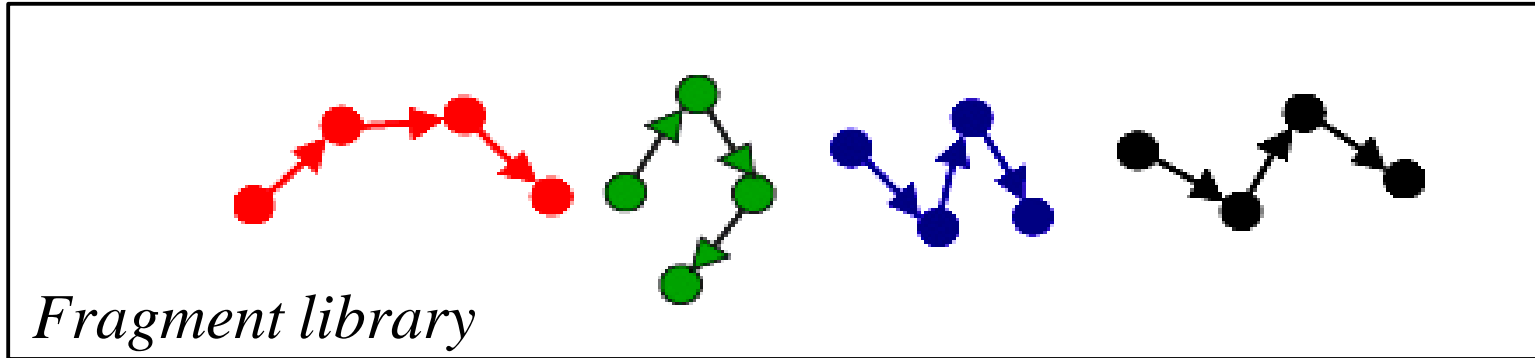
Clustering Protein Fragments to Extract a Small Set of Representatives



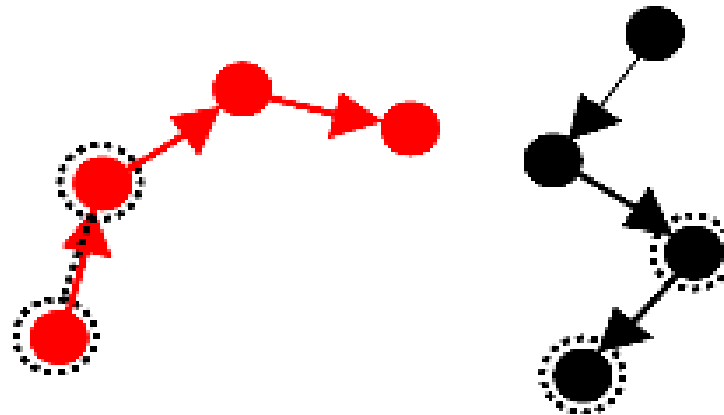
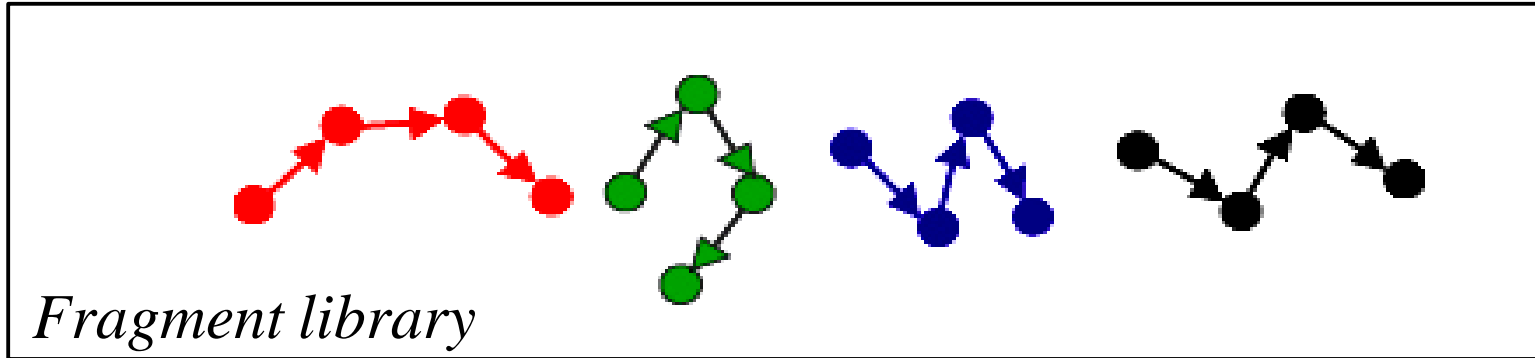
Generating Loops



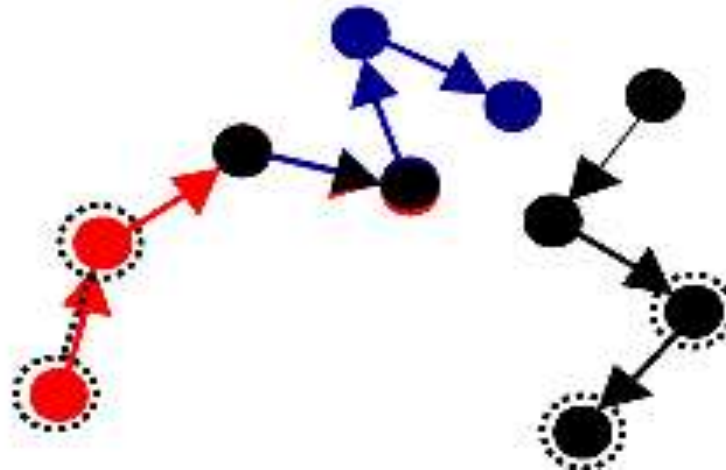
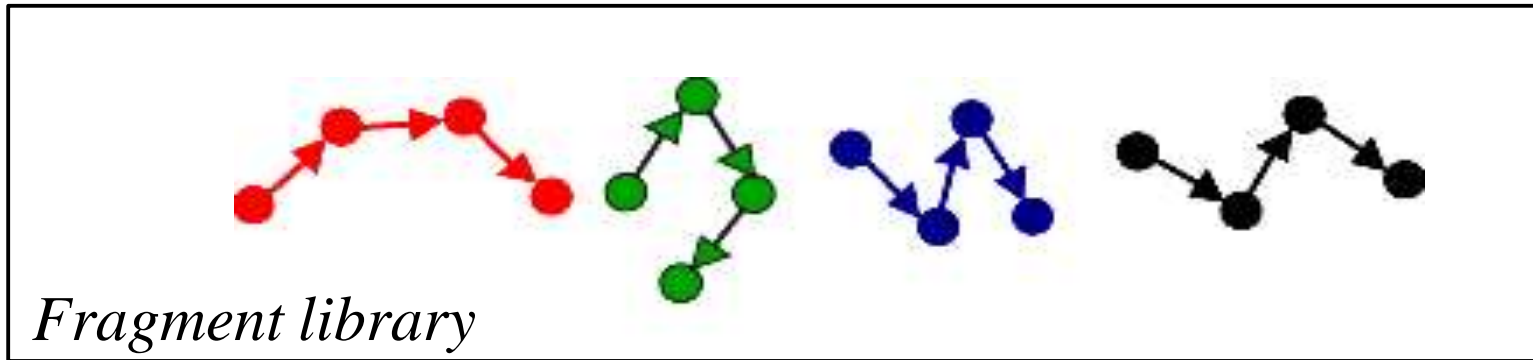
Generating Loops



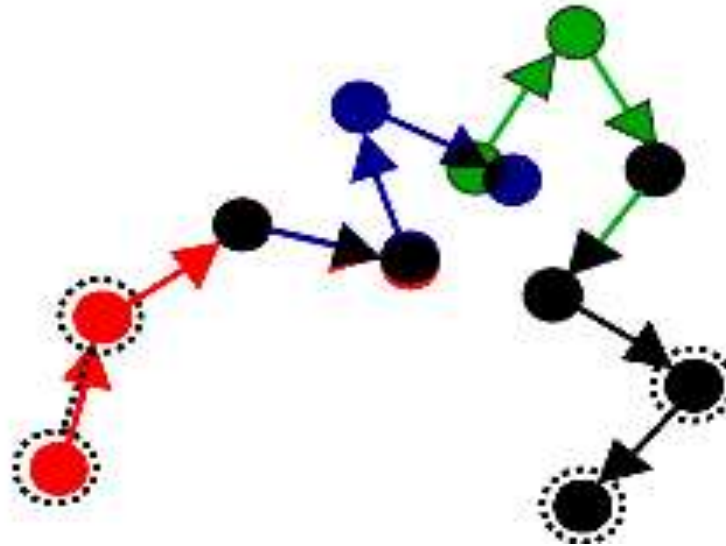
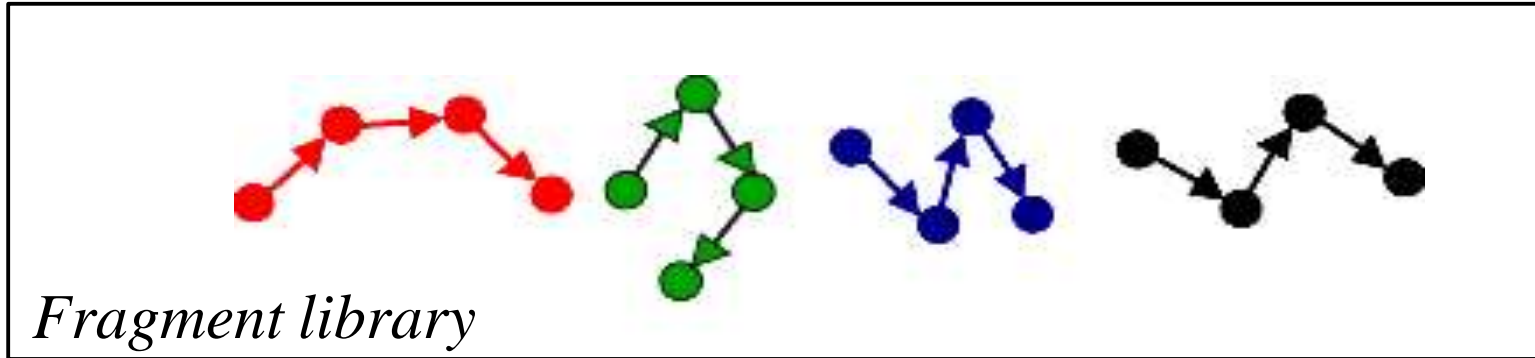
Generating Loops



Generating Loops

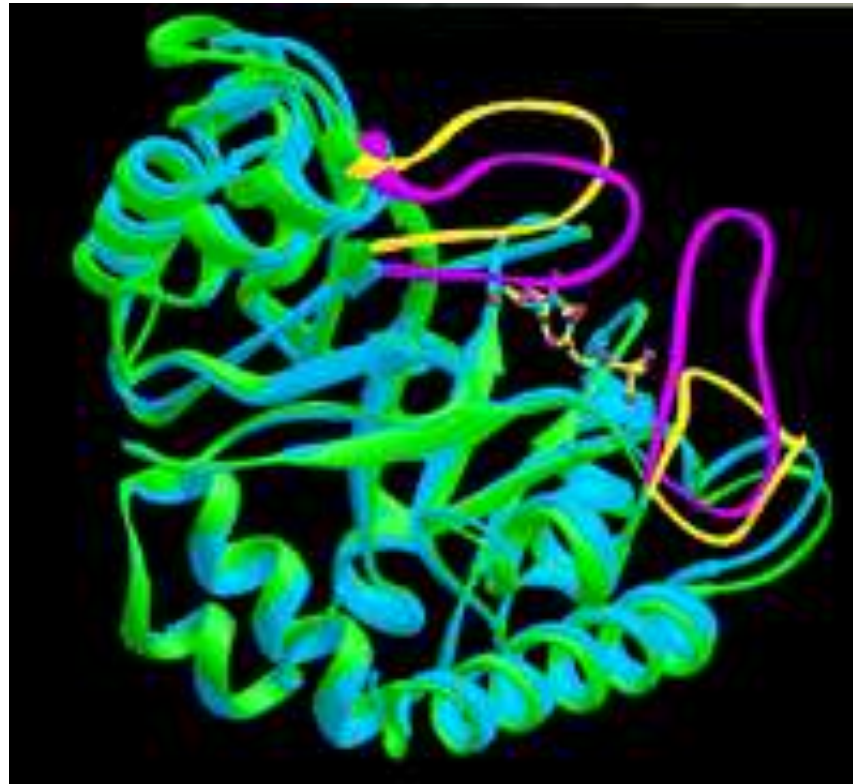


Generating Loops



(iii) Energy based approach:

- Energy function is used to judge the quality of a loop.
- Minimalization of structure
 - Monte Carlo
 - Molecular dynamics techniques.
- Best loop conformation.
- Short loops- superimposes well on true structure.



Loop Modeling Structure

5. Side chain modeling

- ❑ With bond lengths, bond angles and two rotatable backbone bonds per residue ϕ and ψ , it's very difficult to find the best conformation of a side chain.
- ❑ Hence Side chain conformational search in loop regions is a must.
- ❑ Side chain residues replaced during coordinate transformations should also be checked.

- Highly conserved homologous proteins ($> 40\%$)- torsion angle, same orientation- thus copy conserved residues.
- Gives higher accuracy than by copying just the backbone and repredicting the side chains.
- Isolated ($<35\%$)- rotamers differ.

Find the most probable side chain conformation, using

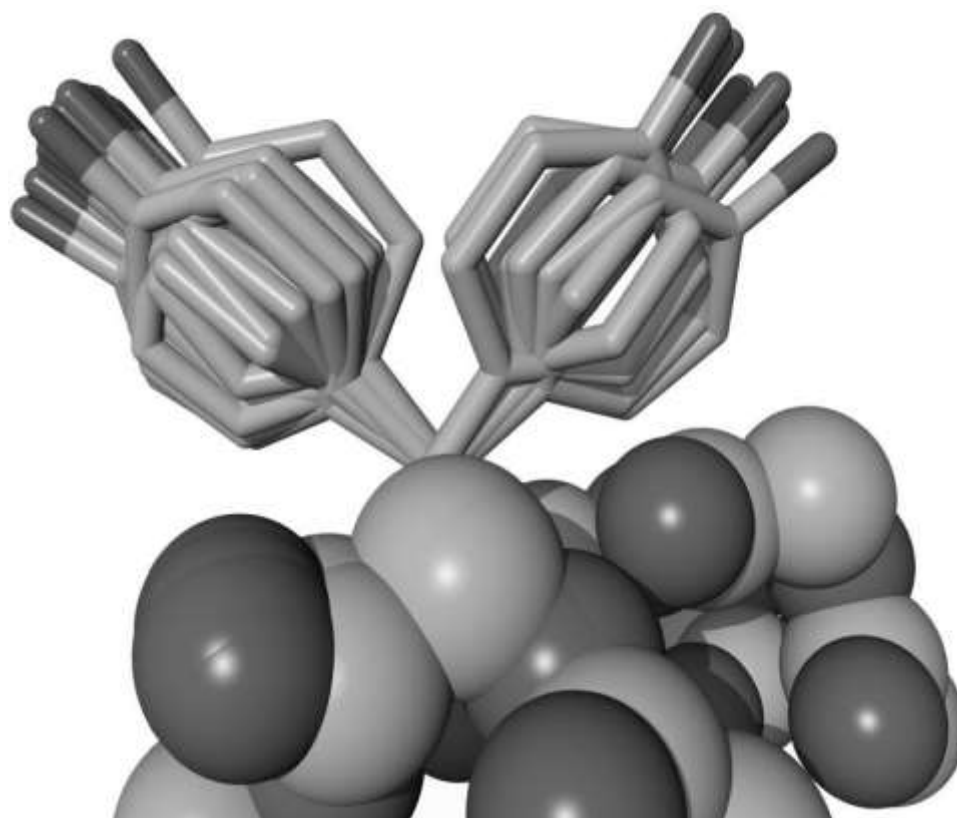
- homologues structure information
- back-bone dependent rotamer libraries
- energetic and packing criteria

- Keep template rigid
- Rotamer libraries provide an ensemble of likely conformations
- The propensity of rotamers depends on the backbone geometry
- Determine best rotamer
- Do not optimize rotamers
- If best rotamer doesn't fit, start thinking.
- If the model is bad, you had the wrong template, or the wrong alignment.
- Make sure your model exists...

Selection Of Good Rotamer

It is observed that:

- ❑ In homologous proteins, corresponding residues virtually retain the same rotameric state ([Ponder and Richards 1987](#), [Benedetti et al. 1983](#))
- ❑ Certain rotamers are almost always associated with certain secondary structure([McGregor et al. 1987](#)).



Backbone dependent rotamer library.

6. Model Optimization

- High accuracy side-chain rotamers-- correct backbone-- depends on the rotamers and their packing.
- Iterative approach used.
- Use 25 – 50 steps energy minimization, or use a force field that has been especially designed for the optimization of homology models.

Model optimization depends on energy function/ force field precision.

Better energetic models for energy minimization:

- Quantum force fields
- Self-parameterizing force fields

Quantum force fields

- Energies expressed as function of position of atomic nuclei only.
- Apply methods of quantum chemistry to entire proteins
- Accurate descriptions of the charge distribution

Self-parameterizing force fields

- force field depends on the parameters
- computationally expensive
- Initial parameter--change—model energy minimization—result basis—keep/dismiss new model.
- Perform molecular dynamics simulations (1 femtosecs)
- Prediction increases

7. Model Validation

- ❑ Every homology model contains errors. Two main reasons
 - % sequence identity between reference and model
 - The number of errors in templates
- ❑ Hence it is essential to check the correctness of overall fold/structure, errors of localized regions and stereochemical parameters: bond lengths, angles, geometries

Many structural artifacts can be introduced while the model protein is being built

- ❑ Substitution of large side chains for small ones
- ❑ Strained peptide bonds between segments taken from different reference proteins
- ❑ Non optimum conformation of loops

2 ways of estimating errors in structure:

- Calculating the model's energy based on a force field: Checks bond lengths, angles; fold corrections.
- Determination of normality indices: Quality of structure; distribution of polar, nonpolar residues; distance and direction of atomic contacts.

ITERATION

CONCLUSION:

- Homology modeling- Not an easy experimental work.
- Further uses threading, ab initio and molecular dynamics simulations to home into a true structure.

Homology Modeling: Practical guide

Approach 1: Manual

- Submit target sequence to BLAST; identify potential templates
- For each template:
 - Generate alignment between target and template
 - Build framework
 - build loop + sidechain
 - assess model (stereochemistry, ...)

Homology Modeling: Practical guide

Approach 2: Submit target sequence to automatic servers

- *Fully automatic:*

- **3D-Jigsaw** : <http://www.bmm.icnet.uk/servers/3djigsaw/>

- **EsyPred3D**: <http://www.fundp.ac.be/urbm/bioinfo/esypred/>

- **SwissModel**: <http://swissmodel.expasy.org//SWISS-MODEL.html>

- *Fold recognition:*

- **3D-PSSM**: <http://www.sbg.bio.ic.ac.uk/~3dpssm/>

- *Useful sites:*

- **Meta server**: <http://bioinfo.pl/Meta>

- **PredictProtein**: <http://cubic.bioc.columbia.edu/predictprotein/>

Model Evaluation

- ❑ WHAT IF <http://www.cmbi.kun.nl/gv/servers/WIWWWI/>
- ❑ SOV <http://predictioncenter.llnl.gov/local/sov/sov.html>
- ❑ PROVE <http://www.ucmb.ulb.ac.be/UCMB/PROVE/>
- ❑ ANOLEA <http://www.fundp.ac.be/pub/ANOLEA.html>
- ❑ ERRAT <http://www.doe-mbi.ucla.edu/Services/ERRATv2/>
- ❑ VERIFY3D
http://shannon.mbi.ucla.edu/DOE/Services/Verify_3D/
- ❑ BIOTECH <http://biotech.embl-ebi.ac.uk:8400/>
- ❑ ProsaII <http://www.came.sbg.ac.at>
- ❑ WHATCHECK <http://www.sander.embl-heidelberg.de/whatcheck/>

Challenges

- ❑ To model proteins with lower similarities(eg < 30% sequence identity)
- ❑ To increase accuracy of models and to make it fully automated
- ❑ Improvements may include simultaneous optimization techniques in side chain modeling and loop modeling
- ❑ Developing better optimizers and potential function, which can lead the model structure away from template towards the correct structure
- ❑ Although comparative modelling needs significant improvement, it is already a mature technique that can be used to address many practical problems

Automated Web-Based Homology Modeling

- ❑ SWISS Model : <http://www.expasy.org/swissmod/SWISS-MODEL.html>
- ❑ WHAT IF : <http://www.cmbi.kun.nl/swift/servers/>
- ❑ The CPHModels Server : <http://www.cbs.dtu.dk/services/CPHmodels/>
- ❑ 3D Jigsaw : <http://www.bmm.icnet.uk/~3djigsaw/>
- ❑ SDSC1 : <http://cl.sdsc.edu/hm.html>
- ❑ EsyPred3D : <http://www.fundp.ac.be/urbm/bioinfo/esypred/>

Comparative Modeling Server & Program

- ❑ COMPOSER <http://www.tripos.com/sciTech/inSilicoDisc/bioInformatics/matchmaker.html>
- ❑ MODELER <http://salilab.org/modeler>
- ❑ InsightII <http://www.msi.com/>

References