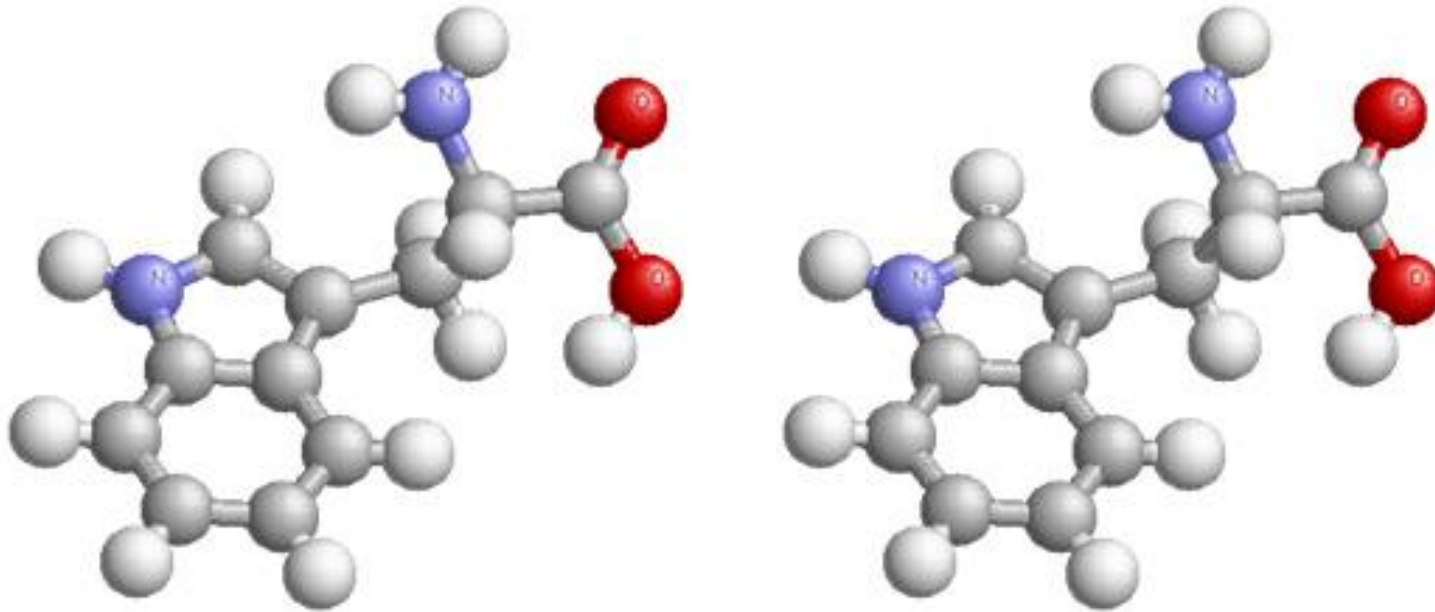
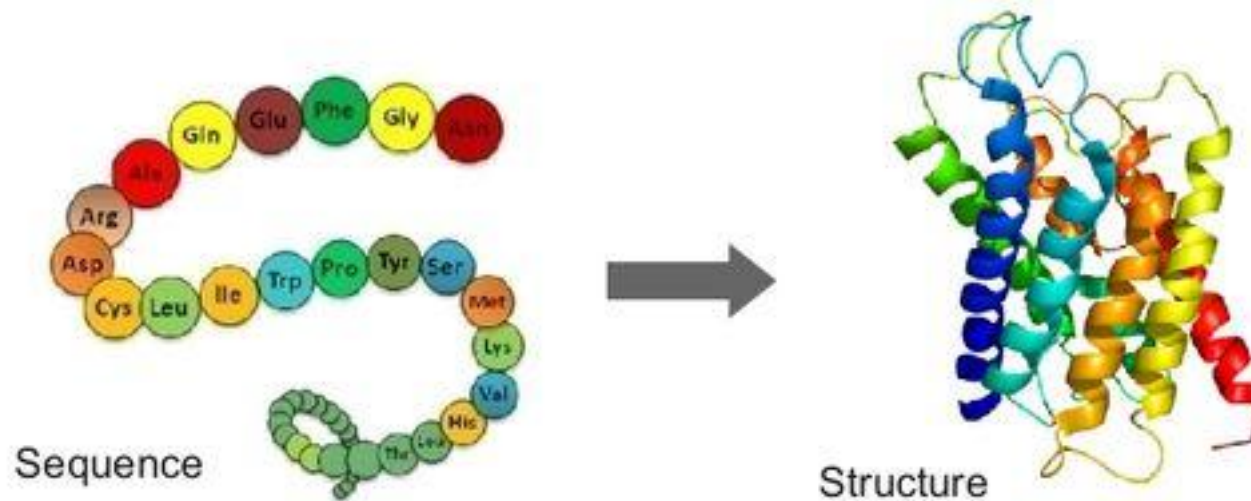


# PROTEIN STRUCTURE PREDICTION (Homology modelling)

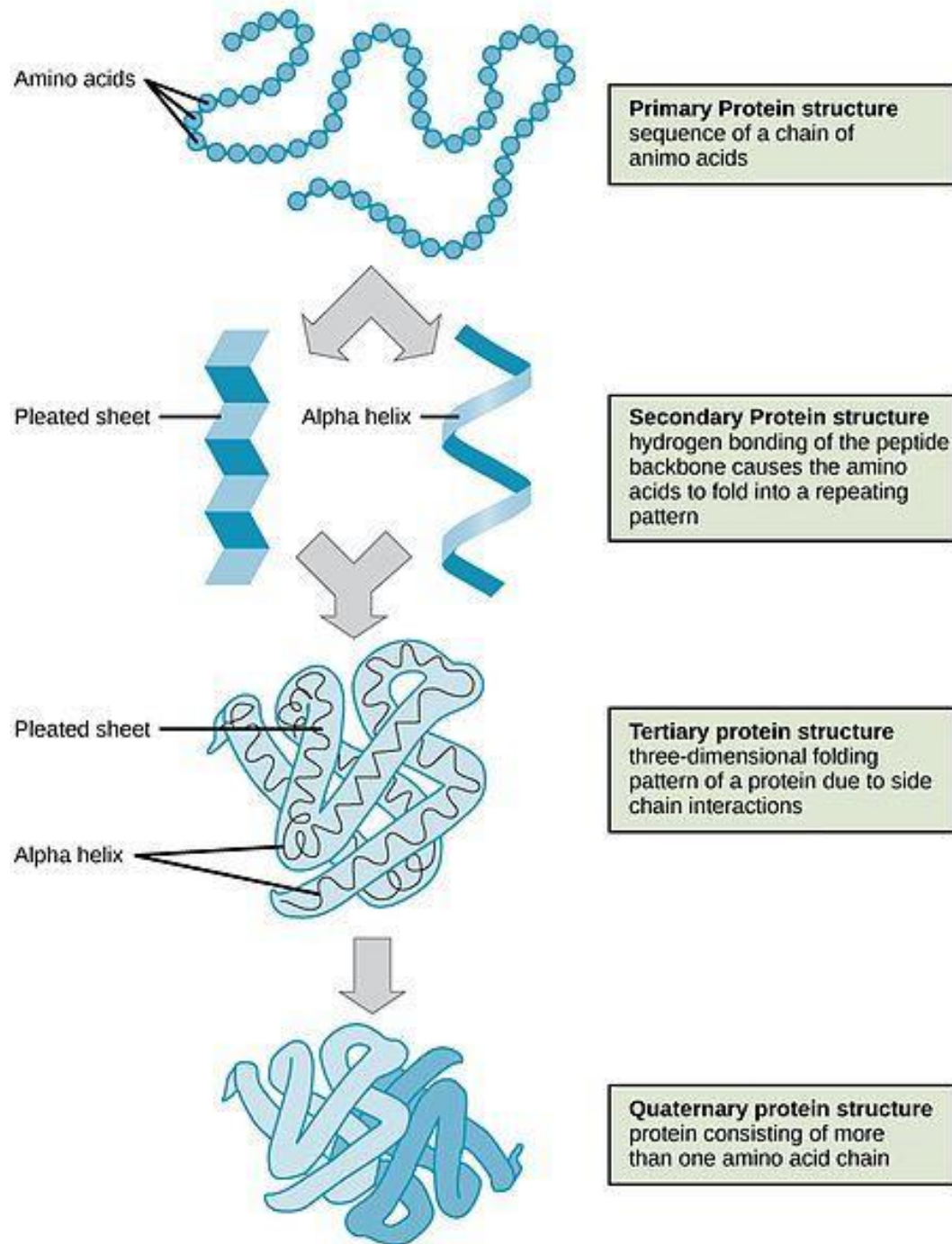


# BACKGROUND

- **Structural prediction-** powerful tool to understand the functions of biological macromolecules at the atomic level.
- **Protein structure prediction-** inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure.



# TYPES OF PROTEIN STRUCTURE



**Biological function of a protein is often dependent upon its tertiary structures.**

Two experimental methods for tertiary structure prediction:

1. X-ray crystallography
2. Nuclear magnetic resonance (NMR) Spectroscopy

Protein structures vary  
depending on the sequences.

NMR and X-Ray crystallography  
are slow in determining  
structures

**Gap between sequences and  
structures**

**Need for protein  
structure prediction arises**

# CLASSICAL METHODS FOR STRUCTURE PREDICTION

**1. X-Ray Crystallography**

**2. NMR Spectroscopy**



Faces difficulty in solving certain structures

Time consuming and limited in their approach

# COMPUTER-BASED STRUCTURE PREDICTION

Three computational approaches to protein three-dimensional structural modeling and prediction:

## Homology Modelling

Knowledge-based

builds an atomic model based on an experimentally determined structure that is closely related at the sequence level

## Threading

Knowledge-based

identifies proteins that are structurally similar, with or without detectable sequence similarities

## *ab initio* method

Simulation- based

predicts structures based on physicochemical principles governing protein folding without the use of structural templates

# HOMOLOGY MODELLING

# HOMOLOGY MODELLING

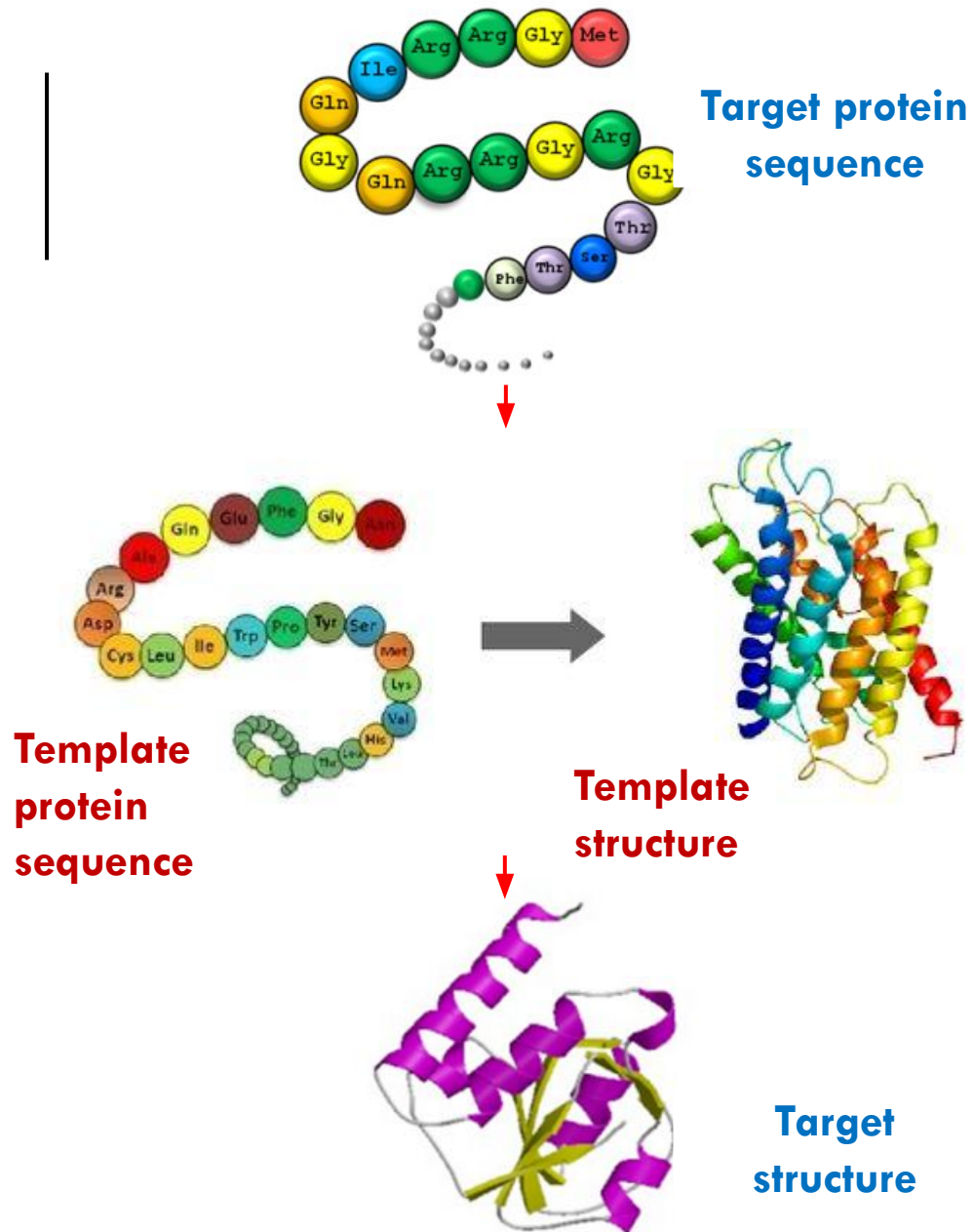
- **Homology modeling** predicts protein structures based on sequence homology with known structures.
- It is also known as **comparative modeling**.
- The principle behind it is *that if two proteins share a high enough sequence similarity, they are likely to have very similar three-dimensional structures*.
- If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a **high degree of confidence**.
- Homology modeling produces an **all-atom model**, based on alignment with template proteins.



# HOMOLOGY MODELLING

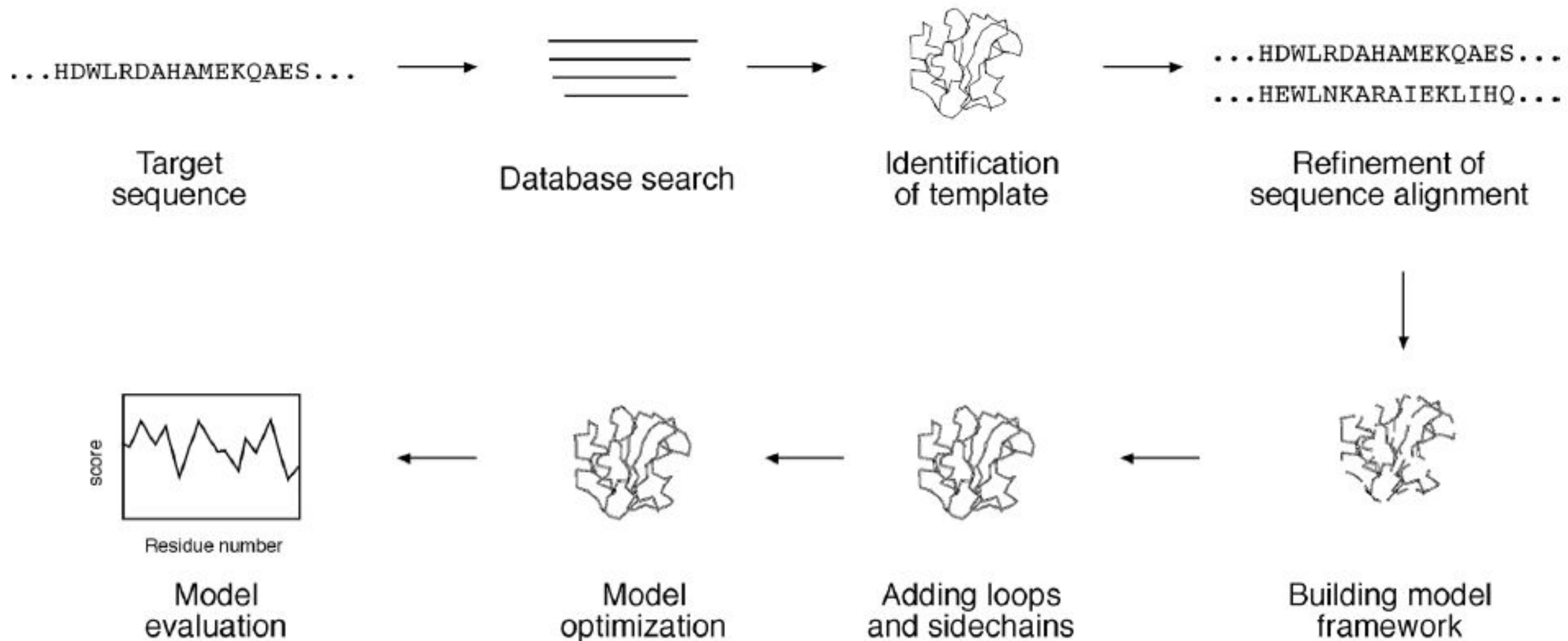
1. Search for homologous sequences in Protein Databases

2. Prediction of structure for target protein from template structure



# SEVEN STEPS IN HOMOMOLOGY MODELLING

1. Template selection
2. Alignment Correction
3. Backbone generation
4. Loop modelling
5. Side-chain modelling
6. Model Optimization
7. Model Validation



**Figure 15.1:** Flowchart showing steps involved in homology modeling.

# 1. TEMPLATE SELECTION

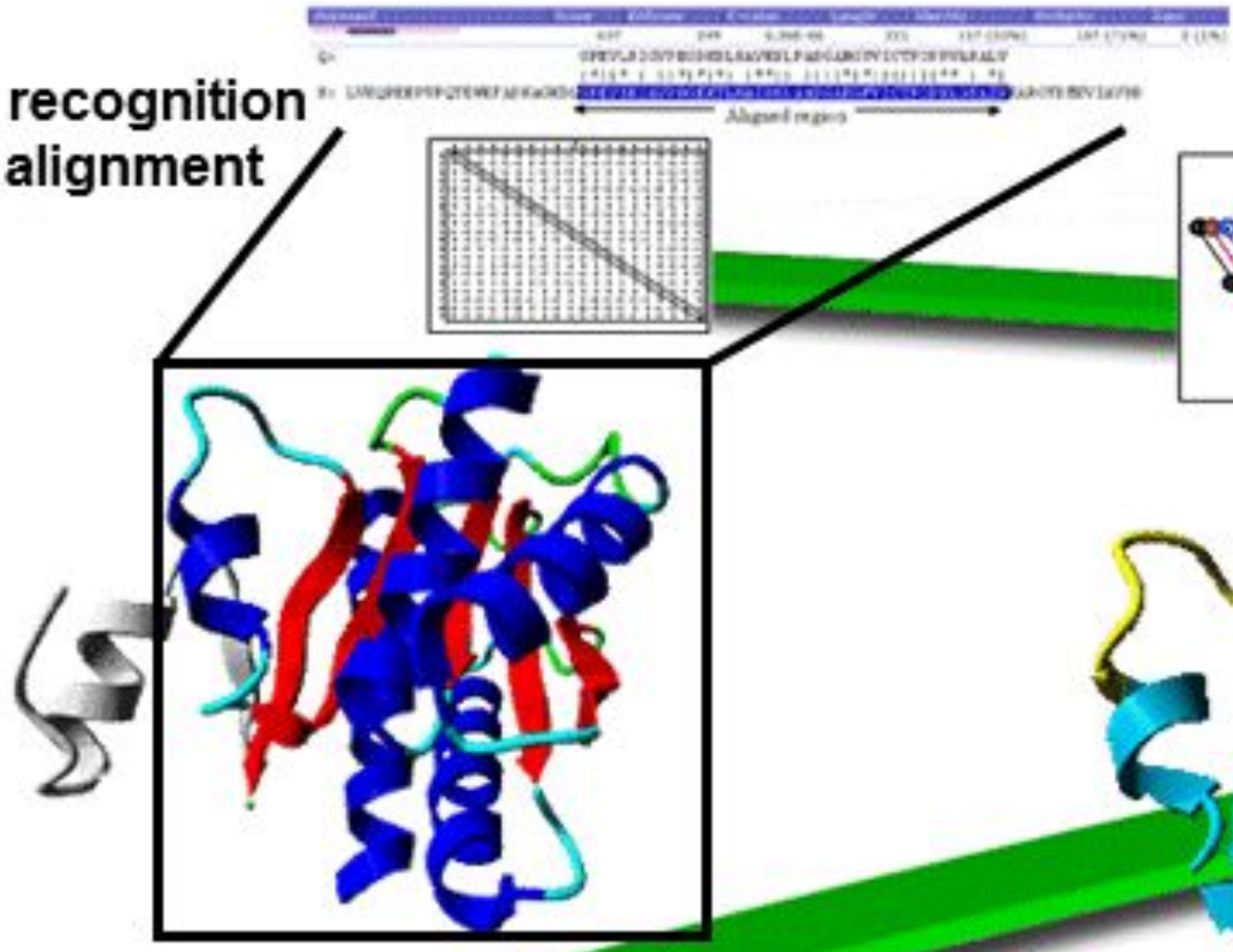
- **BLAST search for the template** in protein databases (e.g. PDB)
- **PSI-BLAST** helps to choose between the sequences which are similarly closer to the similarity threshold.
- In case of greater number of sequences above the similarity threshold, one with **lowest E-value** is considered to be the template.
- Minimum eligibility- 30% sequence identity, (can also be below - conditional)



## **Tools/ Software:**

1. BLAST
2. PSI-BLAST
3. SSEARCH or ScanPS (for sensitive searches)

# 1: Template recognition and initial alignment



## 2. ALIGNMENT CORRECTION

- when a good aligned sequence isn't found, homologous templates are chosen to create a **hybrid model to reach the consensus structure** -by MSA.
- **MSA uses progressive alignment approach** to align two most closely related sequences.
- MSA provides information based on:
  - **conserved residues**
  - **varying residues**

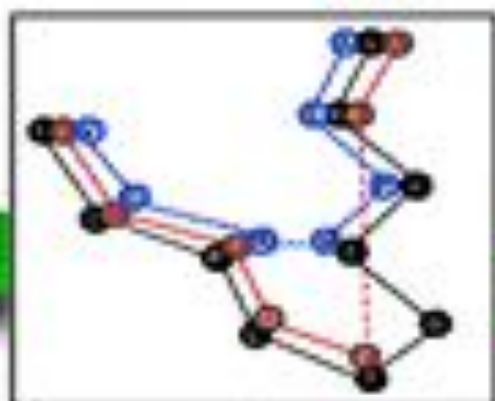
- These aligned sequences may not be error free, requires **manual evaluation** of conserved key residues.
- **Errors made in the alignment step cannot be corrected in the following modeling steps.**

### **Tools/ Software:**

1. T-Coffee
2. ClustalW



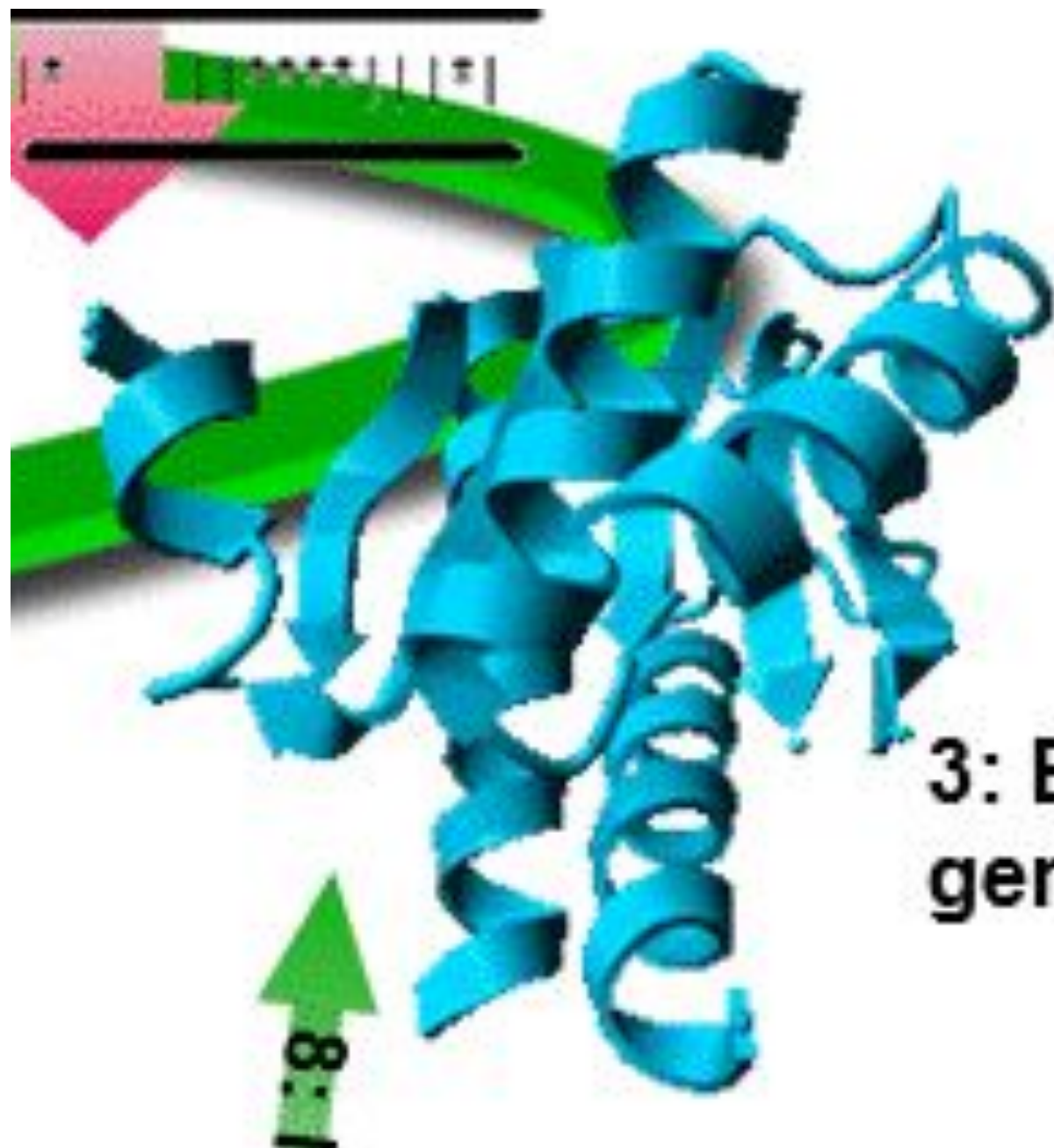
## 2: Alignment correction



### 3. BACKBONE GENERATION

- Protein backbone consists of **N,C and O**, linked through **peptide bond**.
- Backbone prediction is done **by aligning the residues at the atomic coordinates of the residues present at the template**.
- It can also be done with **multiple template modeling**, i.e. using more than one homologous template.

- It is **simplest to use only one template structure**; with the **best quality and highest resolution** if multiple options are available
- This structure tends to carry the fewest errors.



**3: Backbone  
generation**

## 4. LOOP MODELING

- **Gaps created** while aligning regions of model and template
- **Filled by means of indels** with some conformation changes (loop, coils, etc.) to the backbone.
- Rarely happens to secondary structure.
- Steps:
  1. **Knowledge-based**, PDB is searched for known loops with the same end points of the residue loop or simple copy of loop formation.
  2. **Energy-based**, to determine the best loop, energy function is minimized using molecular dynamics.
- **DEMERIT:** interchanging side chains and loops, may lead to change in orientation and spatial arrangement.

## Tools/ Software:

1. FREAD
2. PETRA
3. CODA

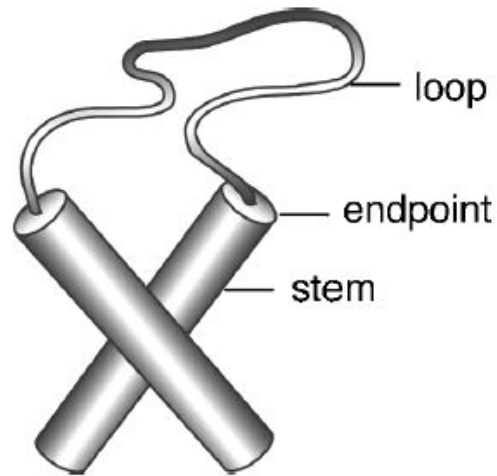
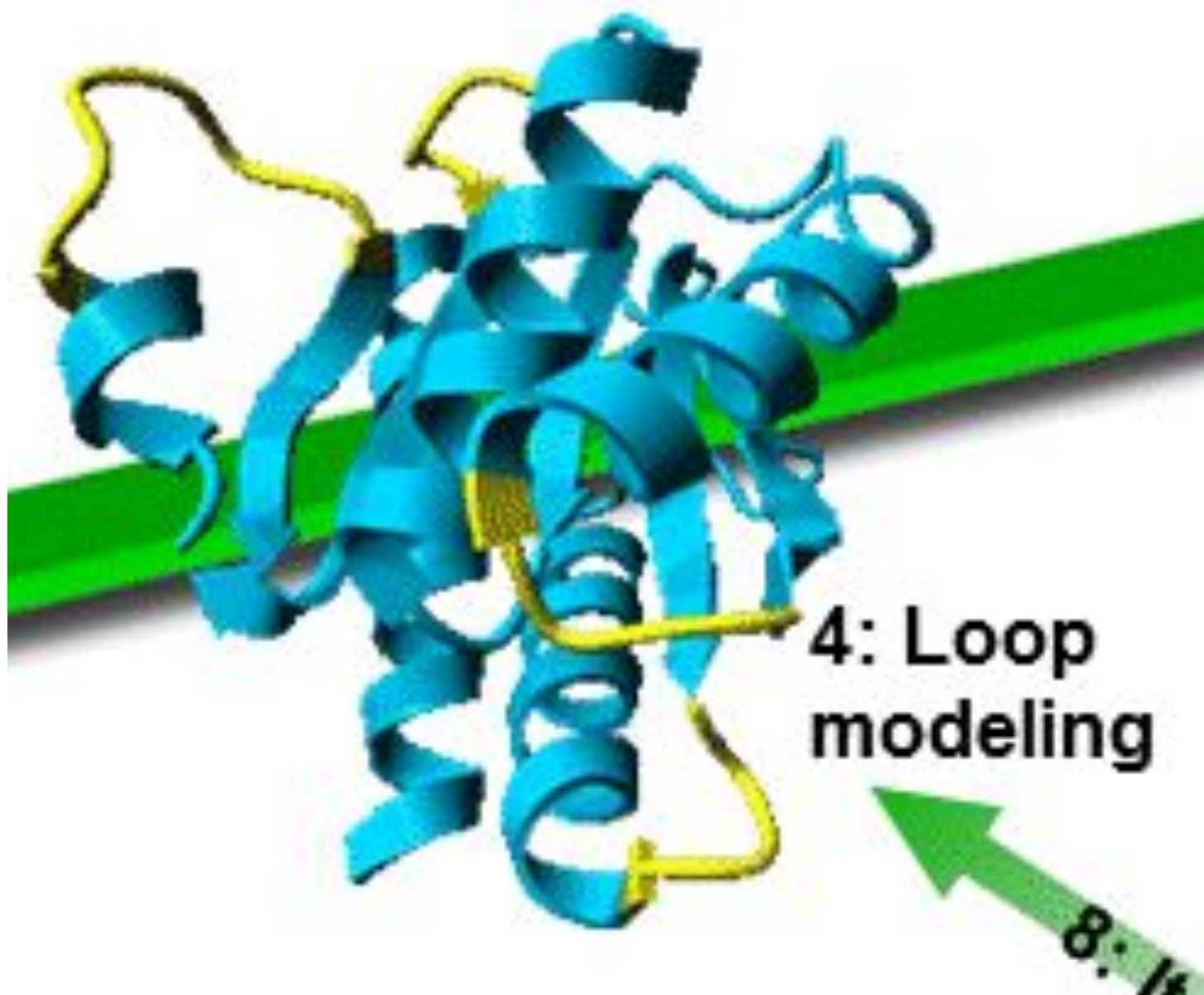


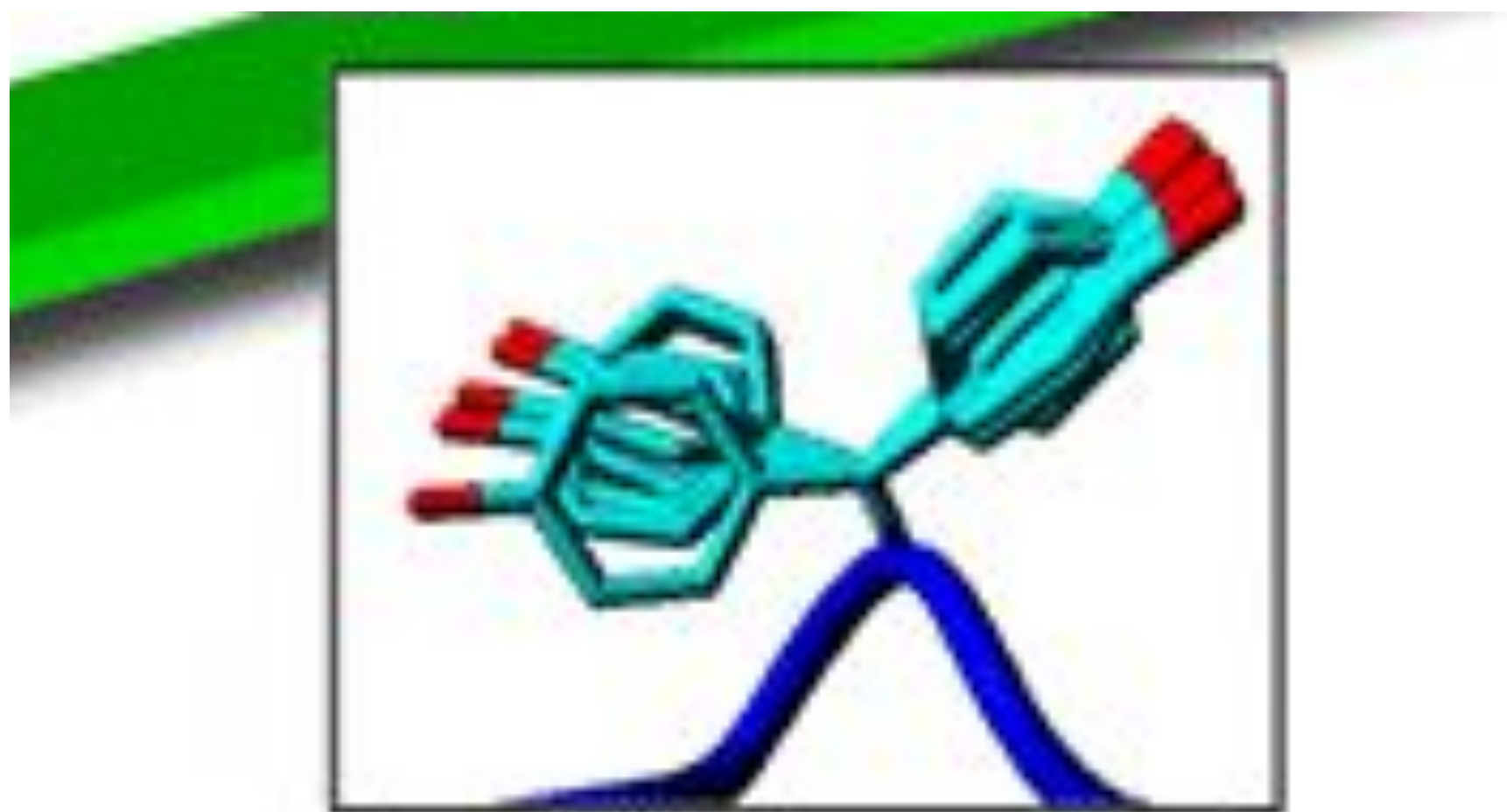
Figure 15.2: Schematic of loop modeling by fitting a loop structure onto the endpoints of existing stem structures represented by cylinders.



## 5. SIDE CHAIN MODELING

- **Side chains (roamers or rotamers) are modeled according to the atomic co-ordinates of the template available.**
- **The structure of side chains are copied at exact C1 angles as present in template structure.**
- **Feasible with structures that have high sequence similarities, computationally exhaustive combinatorial explosion.**
- **Prediction of single roamer, acts as anchors for others.**
- **Core residues remains unhindered and conserved – 90% similar C1 angles compared to experimental data.**





## 5: Sidechain modeling

## 6. MODEL OPTIMIZATION

In these loop modeling and side chain modeling steps, **potential energy calculations are applied** to improve the model.

However, this does not guarantee that the entire raw homology model is free of structural irregularities such as unfavorable bond angles, bond lengths, or close atomic contacts.

These kinds of structural irregularities can be corrected by:

- (i) **energy minimization procedure.**
- (ii) **molecular dynamic simulation**

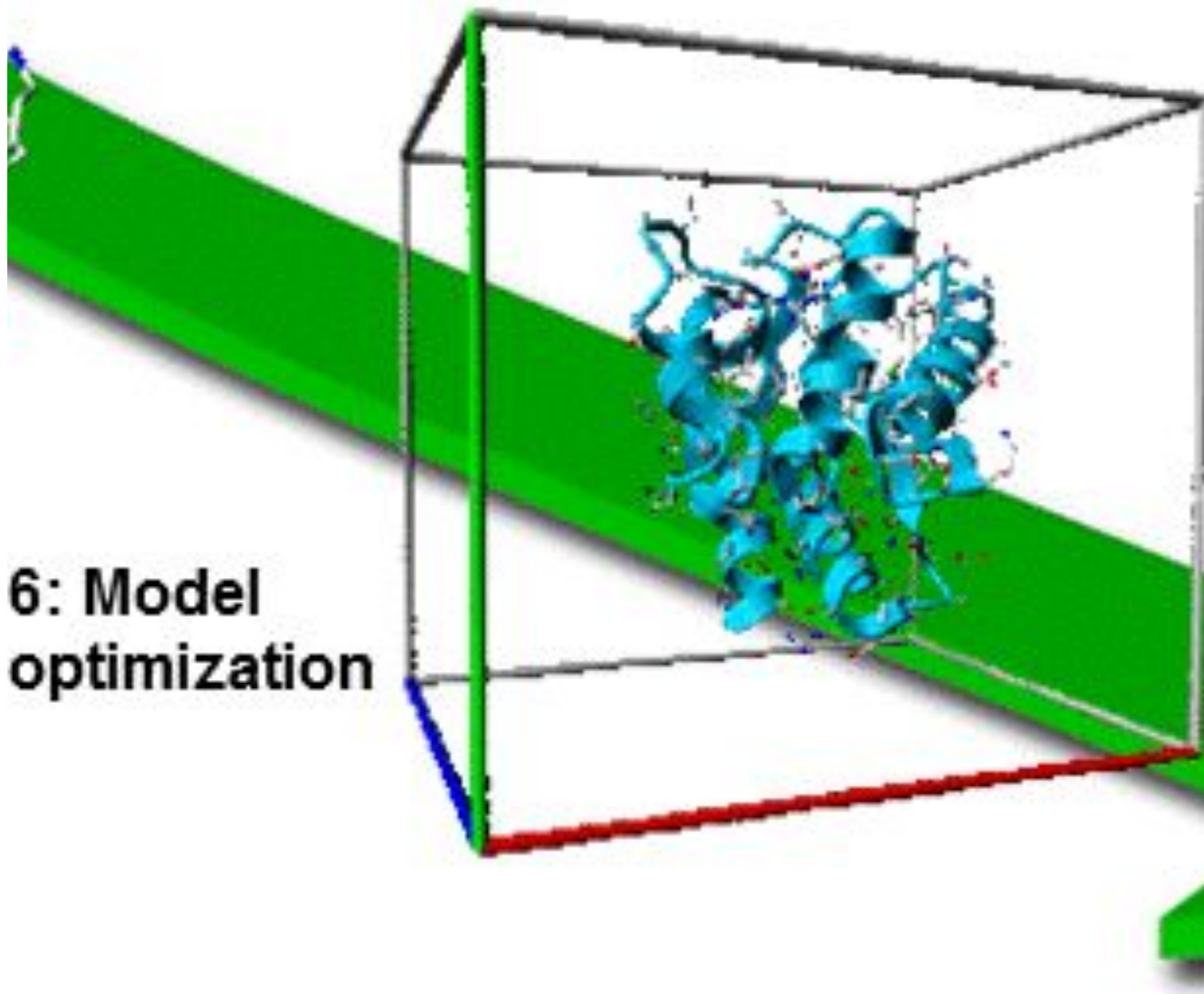
## ***(i) Energy Minimization Procedure***

- moves the atoms in such a way that **the overall conformation has the lowest energy potential**
- **Goal:** to relieve steric collisions and strains without significantly altering the overall structure.
- However, **energy minimization has to be used with caution** because **excessive energy minimization often moves residues away from their correct positions.**
- Therefore, ***only limited energy minimization is recommended*** (a few hundred iterations) to remove major errors, such as short bond distances and close atomic clashes.

## *(ii) Molecular Simulation*

- Energy minimization only moves atoms toward a local minimum **without searching for all possible conformations, often resulting in a suboptimal structure.**
- To search for a global minimum **requires moving atoms uphill as well as downhill in a rough energy landscape.**
- This requires thermodynamic calculations of the atoms.
- In this process, a protein molecule is “**heated**” or “**cooled**” to simulate the uphill and downhill molecular motions.
- Thus, it helps overcome energy hurdles that are inaccessible to energy minimization.

**6: Model  
optimization**



## 7. MODEL EVALUATION

- The final homology model has to be evaluated **to make sure that the structural features of the model are consistent** with the **physicochemical rules**.
- This involves checking anomalies in
  1.  $\varphi$ – $\psi$  angles
  2. bond lengths
  3. bond angles

▪ Another way of checking the quality of a protein model is to implicitly take these **stereochemical properties** into account.

- detects errors by **compiling statistical profiles** of spatial features and interaction energy from experimentally determined structures.

- It reveals *which regions of a sequence appear to be folded normally and which regions do not.*

▪ If structural irregularities are found, the region is considered to have errors and has to be further refined.

▪ **SOFTWARES:** Procheck, WHAT IF

The models we obtain may contain **errors**. These are errors depend on two values:

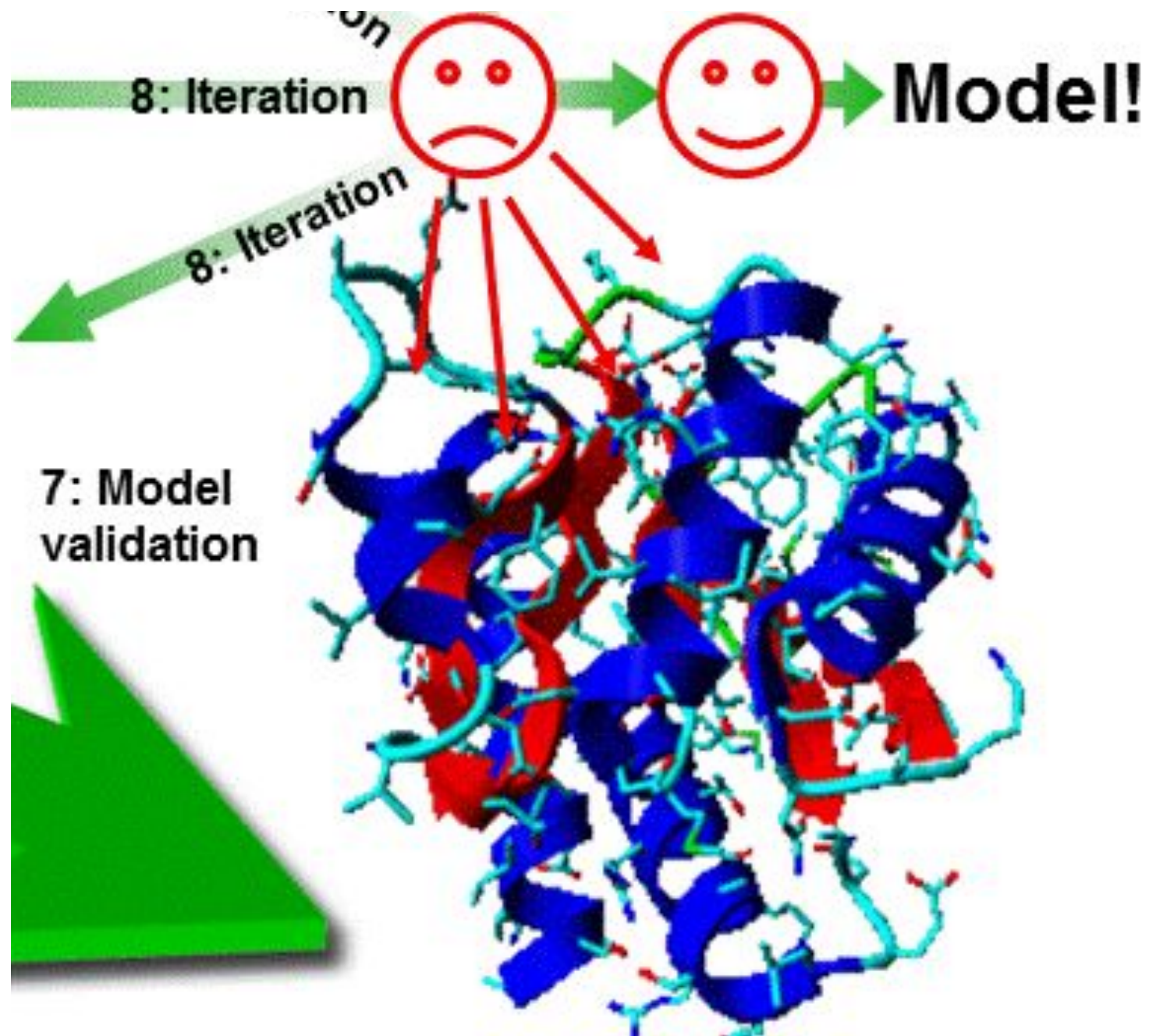
**(i) Percentage identity between template and target**

- **>90%** : accuracy is compared to crystallography
- **50-90%**: R.M.S.D. error can be as large as 1.5%..
- **<25%**: alignment becomes difficult for homology modelling, leads to large errors

**(ii) Number of errors in template**

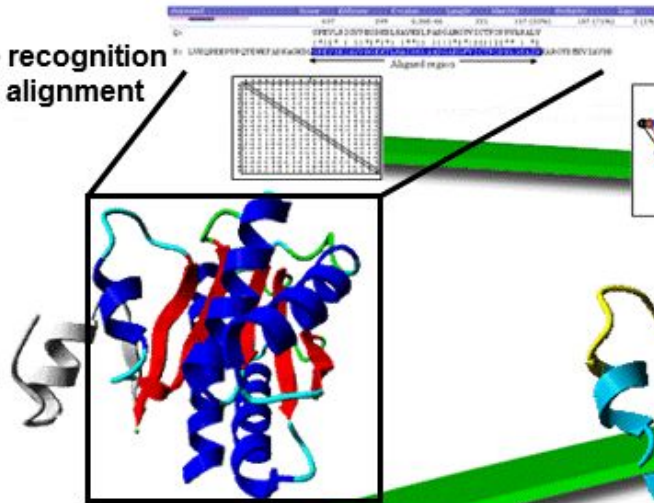
- Errors can be estimated by calculating model's energy based on force field.
- This method checks if bond lengths and angles are in normal range.



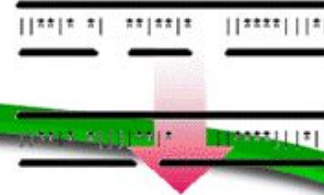
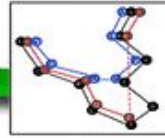


# SUMMARY OF STEPS

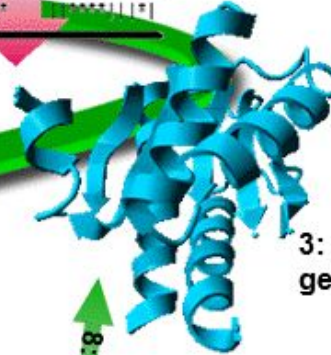
1: Template recognition  
and initial alignment



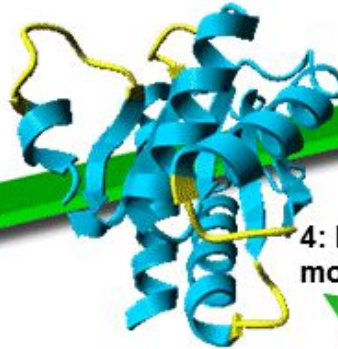
2: Alignment correction



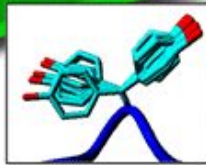
3: Backbone  
generation



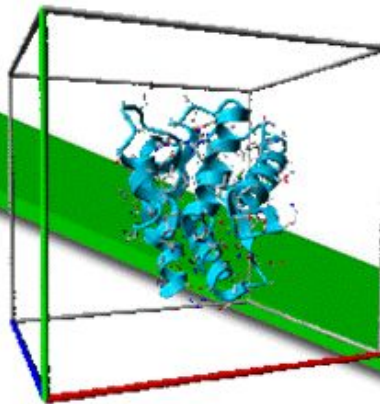
4: Loop  
modeling



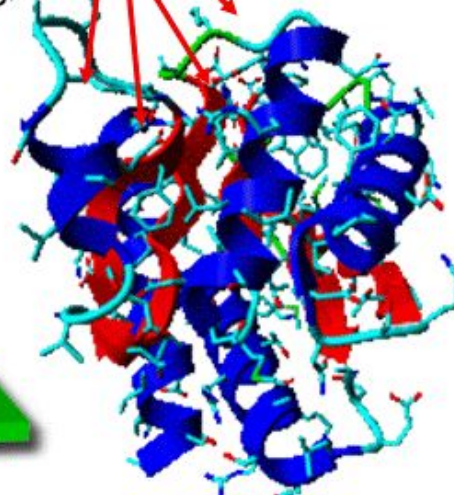
5: Sidechain  
modeling



6: Model  
optimization



7: Model  
validation



8: Iteration

8: Iteration

8: Iteration

Model!



# TOOLS OR SOFTWARE



# AVAILABLE TOOLS FOR HOMOLOGY MODELLING

Server name	URL
SwissModel	<a href="http://swissmodel.expasy.org/">http://swissmodel.expasy.org/</a>
MODELLER	<a href="https://salilab.org/modeller/">https://salilab.org/modeller/</a>
3D-Jigsaw	<a href="http://www.bmm.icnet.uk/servers/3djigsaw/">http://www.bmm.icnet.uk/servers/3djigsaw/</a>
CPHModels	<a href="http://www.cbs.dtu.dk/services/CPHmodels/">http://www.cbs.dtu.dk/services/CPHmodels/</a>
EsyPred3D	<a href="http://www.fundp.ac.be/urbm/bioinfo/esypred/">http://www.fundp.ac.be/urbm/bioinfo/esypred/</a>
Robetta	<a href="http://robeta.bakerlab.org/">http://robeta.bakerlab.org/</a>
Geno3D	<a href="https://geno3d-prabi.ibcp.fr/">https://geno3d-prabi.ibcp.fr/</a>
Phyre2	<a href="http://www.sbg.bio.ic.ac.uk/phyre2/">www.sbg.bio.ic.ac.uk/phyre2/</a>
RaptorX	<a href="http://raptorx.uchicago.edu/">raptorx.uchicago.edu/</a>
WHAT If	<a href="http://swift.cmbi.kun.nl/">http://swift.cmbi.kun.nl/</a>
HOMER	<a href="http://protein.cribi.unipd.it/homer/help.html">http://protein.cribi.unipd.it/homer/help.html</a>
RAMP	<a href="http://software.compbio.washington.edu/ramp/">http://software.compbio.washington.edu/ramp/</a>

# 1. SWISS- MODEL



**BIOZENTRUM**  
University of Basel  
The Center for Molecular Life Sciences

SWISS-MODEL

[Modelling](#) [Repository](#) [Tools](#) [Documentation](#) [Log in](#) [Create Account](#)

## Welcome to SWISS-MODEL

SWISS-MODEL is a fully automated protein structure homology-modelling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make protein modelling accessible to all life science researchers worldwide.

[Start Modelling](#)

- <http://swissmodel.expasy.org>
- protein structure homology modeling server

Three different modes:

### **1. automated mode:**

- availability of highly similar template structure.( $>50\%$ )
- either simply amino acid sequence or UniProt accession number is required.

### **2. Alignment mode:**

- Requires MSA in file formats like FASTA, ClustalW, MSF, PFAM, etc.
- Employs comparative modeling on MSA results.

### **3. Project mode:**

- Employed when unavailability of correct alignment found between template and target sequence.
- Allows manual manipulation of SWISS-MODEL project files, through DEEP VIEW program.
- User has full control on modelling parameters.
- Also used for improving results gained from other two modes.

# 2. MODELLER

## Modeller

Program for Comparative Protein  
Structure Modelling by Satisfaction  
of Spatial Restraints



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

About MODELLER

MODELLER News

Download & Installation

Release Notes  
Data file downloads

Registration

Non-academic use

Discussion Forum

Subscribe  
Browse archives  
Search archives

## About MODELLER

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (3,4), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is [available for download](#) for most Unix/Linux systems, Windows, and Mac.

Several graphical interfaces to MODELLER are [commercially available](#). There are also many other [resources and people using Modeller](#) in graphical or web interfaces or other frameworks.



- “Modeling by satisfaction of restraints.”
- Minimizes the restraints obtained from alignments or experiments like NMR.
- It calculates optimized structure with all non-hydrogen atoms.
- Also performs comparison of protein structure or sequences, clustering of proteins, searching of sequence databases.

# 3. GENO 3-D




## PRABI-GERLAND RHONE-ALPES BIOINFORMATIC POLE GERLAND SITE

Institute of Biology and Protein Chemistry

 Home


 Services

 Teaching

 Publications

 Links

 Jobs

 Contact

GENO3D Release 2 : AUTOMATIC MODELING OF PROTEINS THREE-DIMENSIONAL STRUCTURE

[Abstract] [[GENO3D help](#)] [Original server]

Database :

Sequence 1 name (optional) :

Paste protein 1 sequence below : [help](#)

- <http://geno3d-pbil.ibcp.fr>
- automatic web server for protein molecular modeling.
- Generates 3D model, with wholesome available information of the molecule.
- Information includes Ramachandran plot, deviation between models, restraints deduced from template.

## 4. PHYRE2

# Phyre<sup>2</sup>

Protein **H**omology/analog**Y** Recognition **E**ngine V 2.0

**Subscribe to Phyre at Google Groups**

Email:

[Visit Phyre at Google Groups](#)

 [Follow @Phyre2server](#)



[Cambridge 2019 Workshop](#) | [Older Workshops](#) | [Phyre2 paper](#)

- <http://www.sbg.bio.ic.ac.uk/>
- Protein homology/analogY recognition engine.
- Remote homology generation technique to generate reliable protein models, when other techniques can't.
- It has an user friendly interface, available with various platforms.
- Uses hidden MARKOV models using HHSearch, to improve alignment.
- Ab-initio simulations can be carried out through POING2 interface.

# 5. RAPTORX



[My Jobs](#) | [Docs](#) | [Download](#) | [Inquiry & Bug Report](#) | [About](#) | [Xu Group](#) | [Forum](#)

## RaptorX: Protein Structure and Function Prediction Powered by Deep Learning

RaptorX is developed by Xu group, excelling at secondary, tertiary and contact prediction for protein sequences without close homologs in the Protein Data Bank (PDB). RaptorX predicts protein secondary and tertiary structures, contact and distance map, solvent accessibility, disordered regions, functional annotation and binding sites. RaptorX also assigns confidence scores to predicted structures. See details below and [HERE](#).

**Currently the following servers are running. [Instructions](#).**

**RaptorX Structure Prediction**: the main server predicting protein secondary and tertiary structure, binding site and GO annotation. This server is ranked very top in CASPs and the fully-automated, live benchmark CAMEO. See [here](#) for a ranking list of the publicly-released structure prediction servers. Please use the RaptorX-Property server if you only want to predict secondary structure, solvent accessibility and disordered regions, and RaptorX-Contact for only ab initio folding and contact prediction.

[[Submit](#)][Find jobs by: [ID or sequence](#),[Email](#)][[Example](#)][Refs: [1](#),[2](#),[3](#),[4](#)]

**RaptorX Property Prediction**: protein structure property prediction without using templates,

### Login to your account

Once you submit one job with your email, an account is automatically created for you. Fill in your email address below to see the status of all your jobs.

Email:

### Server Status

821 jobs pending  
146 jobs done in the last 24 hours  
5420 jobs done in the last 30 days

#server users: 57986

#processed jobs: 478679

- <http://raptorx.uchicago.edu>
- protein structure prediction
- uses remote homologous templates in absence of close homologues in PDB.
- Filters various predicted models on the basis of P-scores, probability scores.
- Apart from secondary and tertiary structure prediction, it also determines disordered regions and solvent accessibility.

**THREADING**



# THREADING AND FOLD RECOGNITION

- Term – “**threading**” coined by David Jones in 1991.
- Approach to fold recognition **using a detailed 3D representation of protein structure.**
- Predicts the structural fold **by fitting the sequence into structural databases or by selecting the best fitting fold.**
- Work on two back-end algorithms:
  - Pair wise energy based- threading
  - Profile based- fold recognition
- Works on algorithms based on sequence databases if sequence similarity is found for folds, otherwise deals with trying to fit available folds present as of now.

# THREADING SERVERS:

- RaptorX
- LOOPP
- THREADER
- PHYRE and PHYRE2
- RAPTOR



# ***AB INITIO*** **METHODS**

# **AB-INITIO PROTEIN STRUCTURE PREDICTION**

▪ *ab initio* means “from the beginning”

- **Uses amino acid sequence alone to reach to the target protein structure.**
- **Technique uses prior knowledge of protein modeling.**
- **It is heuristic approach of minimizing the energy background to get to most validated model from all possible conformations.**
- **From the amino acid sequence, a native conformation is extracted based on possibly most stable known interactions.**
- **Unlike homology modeling and threading it focuses on stability of the target model, based on automated hit and trial basis.**

## Tools used:

### Rosetta

- Relies on “mini threading method”.
- Sequence is breakdown to small sub sequences of length 3-9 residues.
- Secondary structures are predicted for these using hidden markov model based program, HMMSTR.
- These secondary structures are assembled to random 3D conformations, and the one with overall lowest global free energy is selected on the basis of stability.
- Although not so reliable but it is a an approach to solve the structures to which homologous structures or sequences are not available.

# **CRITICAL ASSESSMENT OF TECHNIQUES FOR PROTEIN STRUCTURE PREDICTION (CASP)**

# CASP

- A **community-wide, worldwide experiment** for protein structure prediction taking place every two years since 1994.
- With so many protein structure prediction programs available, **there is a need to know the reliability of the prediction methods.**
- For that purpose, **a common benchmark is needed to measure the accuracies** of the prediction methods.
- It allow programmers to ***design for same molecule with different structure prediction methods, surveying the closest or most reliable prediction method.***

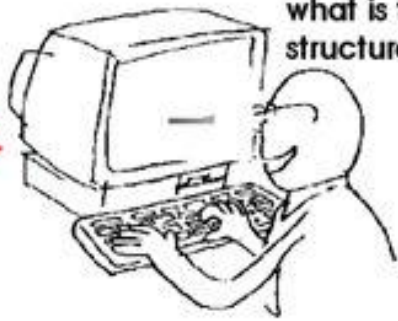
- CASP contestants are given protein sequences whose structures have been solved by x-ray crystallography and NMR, but **not yet published**.
- Each contestant **predicts the structures** and **submits the results to the CASP organizers** before the structures are made publicly available.
- Results are compared to solved structures with programs like VAST, SARF, DALI, etc.



- This approach provides **a valuable insight into protein structure prediction** and also acts as driving force of development of the field.
- The predictions can be made at various levels of detail (secondary or tertiary structures) and in various categories (homology modeling, threading, ab initio).

NDPLSPDYSPKAMQKGLTPBAAGYLDFFERRPLDMKQYLNLTIRGQGVYKNGQREPYCRFQDGL  
FFPQEMRYORHPPEARENWNGWYVTSPRAKWHWELNHTSFANTQFTFPDEAKQHFSDLFQGIN  
AQQGQRYSELLANGLLEGLLFRKAR  
QKADPNCAQDWQPGQKCYTETAMTALAETNCKMLGQRLASQSEENSPQTLNAGVYVWQD  
SACLGADAWTWSDQTFMMRFQWCSTKPSQVLAACQMGMAAAGQWQDLFCFASHEQVCAMTP  
ECUKWASLQVTSFNCCTOMSNVCQPTERISKWIEDGTFYVQKRCVATVQYQDELQKQVNEFCNPE  
DQGLLRSEFQGVYVNLKGGVLSVMPFRRSNNGEHLWMTSSITFQNPYQDEEYUCTKT  
SCFALISDTAMGVYCVRFRVXKZHRHYETCTTWTFTVADNGAERQDQAGLTFQSPSQRGDFLR  
WPLFPQMMSGTASLDF  
ASALERVQPTTVYQEQDPAADPAQWVQTOPNGYTYVLMAPQYQNPFEVYQDAENPVKITSPOV  
HDPHYQTMNVEVLPQEVSTVYTFKRPSEYRQNGYCOLSHQAMPSTVYVRE  
MAAPTLLEAHDVDRBALWSKRSDNAQSGERLAERKWFAVQAADEMLNQDSTETAFADVQHTQV  
SASTLRDQVYQVQAFAPDWAALYDOR  
LQNFSGTCTKAGQGVLTSTCERKNOQVNTSSDLSNVEVYDQSLKQSPNPFETCRNTGLAQQS  
ELANECVTRAGQGVYTKMLDQKADOTLKYE  
MRQSEVPGLTDLSPVQITDSSGLRWTFLESTISYRITVNAQEGPREDVQSEVGYTITGLSPQ  
DIDHSVTLNQGSEAPTTLTQGT  
ETDCRWDQCSQEFQGGGLVHNSHNDKSEFVCHWDQCRRLRPFKADYALVYVHMRKTOE  
KPKCTFBQCKKQYKALENLKTLHRSHTQENPYMCQWEGCSKAFSNAQKAKQMTKQKRPVY  
CALPSCVRYTDPSSLRKXNVTYKQ  
YKPPYLYALITMAALQSFQKLTLSQICPFIHNRFPYFEEKFPAWQNSRMLSLADCFVQPREQMP  
QKQKVVTLQPGSEDMFQAGSLRRRRR  
APQFHWQDPLNFDQAMQSDQTAHAKRLETQENYVNDQPELFEABEYADMCQCHQVYASD  
KQFQLANQWTVQGMETVQLFSTLYQATQGMQPMQGLTLEMLRTNMYVHLTYDQPKDAS  
WLTDEQKADTFPQP  
SATDAPKAKYACESARNPREKLAECLEGRCAEQYDMYRQNVYTRDQECQWRSRYFKPFI  
NSTTHPADLRKRCRNPQGSFQPVCTYTSPLRRESCSVVPCQGRVTVYVPR  
QSPFQHSKRYNKAULTKAGQKMLKRPQDQGLVSRKNEPNSVQSFQDQKPKCEVQGEQGT  
VMLQHSFQSLVQLHVEKMLYKMLKLYPWEENS  
LQDAENYQWQSRQEVNHLRQTAQDTFLVRDASTKMKQDITLTKKQDNKLKFRHQQKQYFS  
QPLFFNEVYSLNHYRNELAGYVFKLQVLLFVSKY  
TQAMALDAQRLSQACTTSQKQYVMSQVQSPKQGLFYQAEQATATGRQSEVRYQKLVY  
ELMQGRTTQDQACKEAVSRVYVNRQKMLKQKQVQFALMKQETQATCQDQNFAYWQKQRE  
LETFOALK

## Unknown Sequences



what is the  
structure?

Theoretic groups

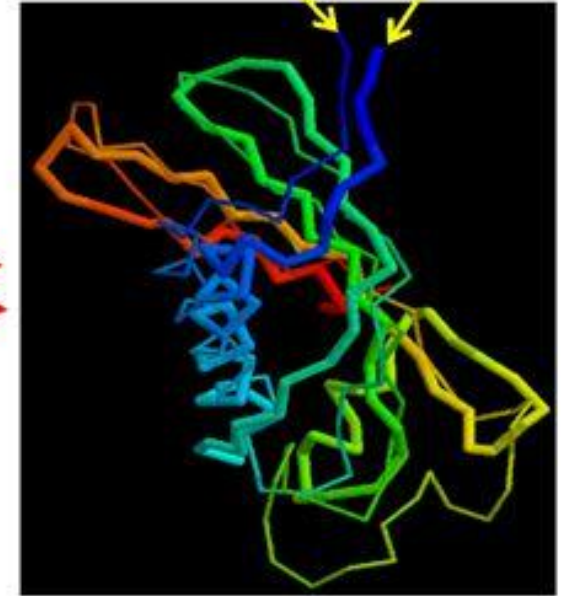


Experimental groups

3month  
later ...

model

native



Meeting ...  
Comparing ...  
Ranking ...

# APPLICATIONS OF HOMOLOGY MODELLING

(1) studying the **effect of mutation**

(2) identifying **active and binding sites** on protein (useful for ligand design )

(3) **searching for ligands** of a given binding site  
(database mining)

(4) **designing novel ligands** of a given binding site

(5) modeling **substrate specificity**

- (6) predicting **antigenic epitopes**
- (7) **protein-protein docking simulations**
- (8) **molecular replacement** in X-ray structure refinement
- (9) **rationalizing known experimental observations**
- (10) **planning new computational experiments** with the provided models.