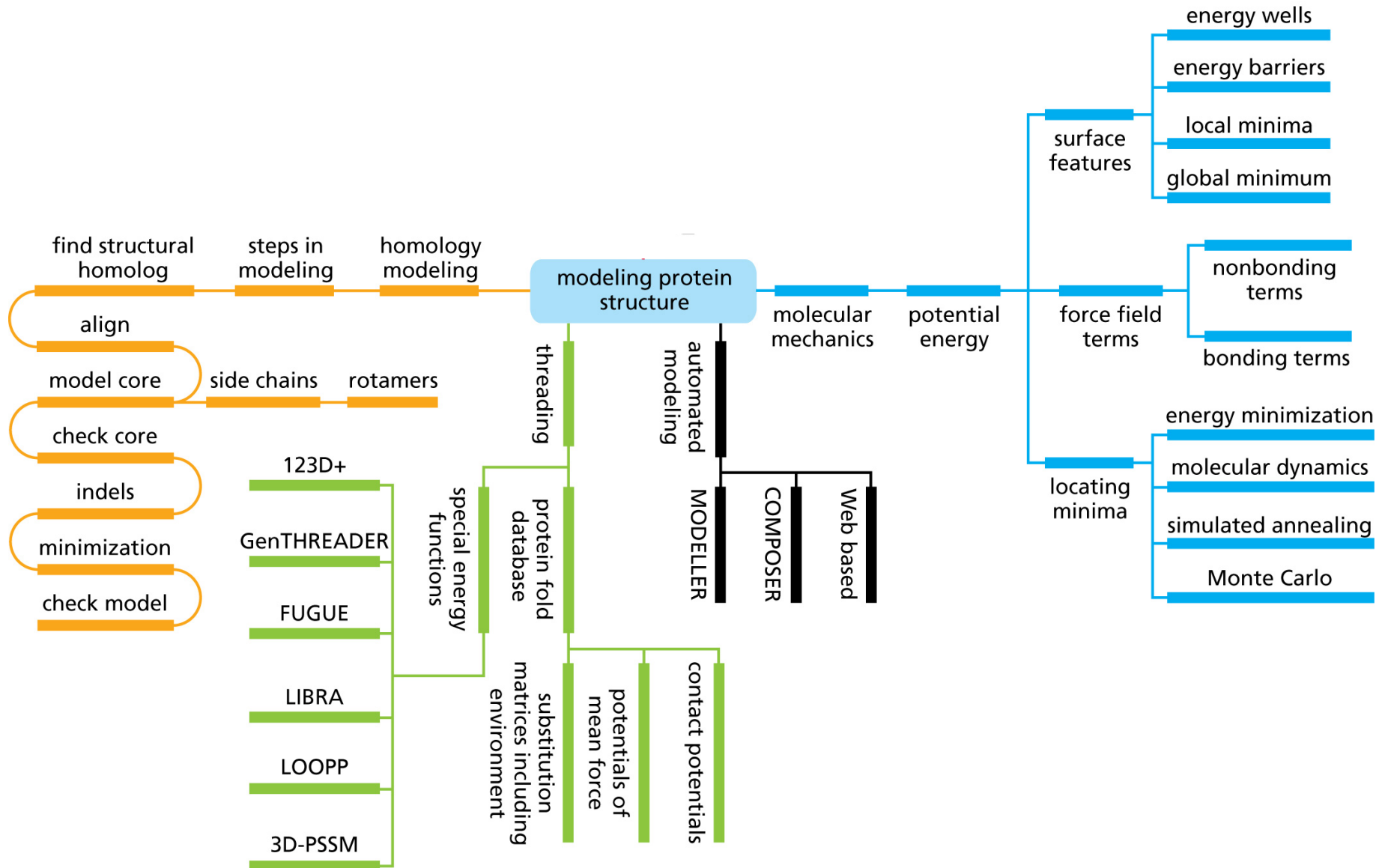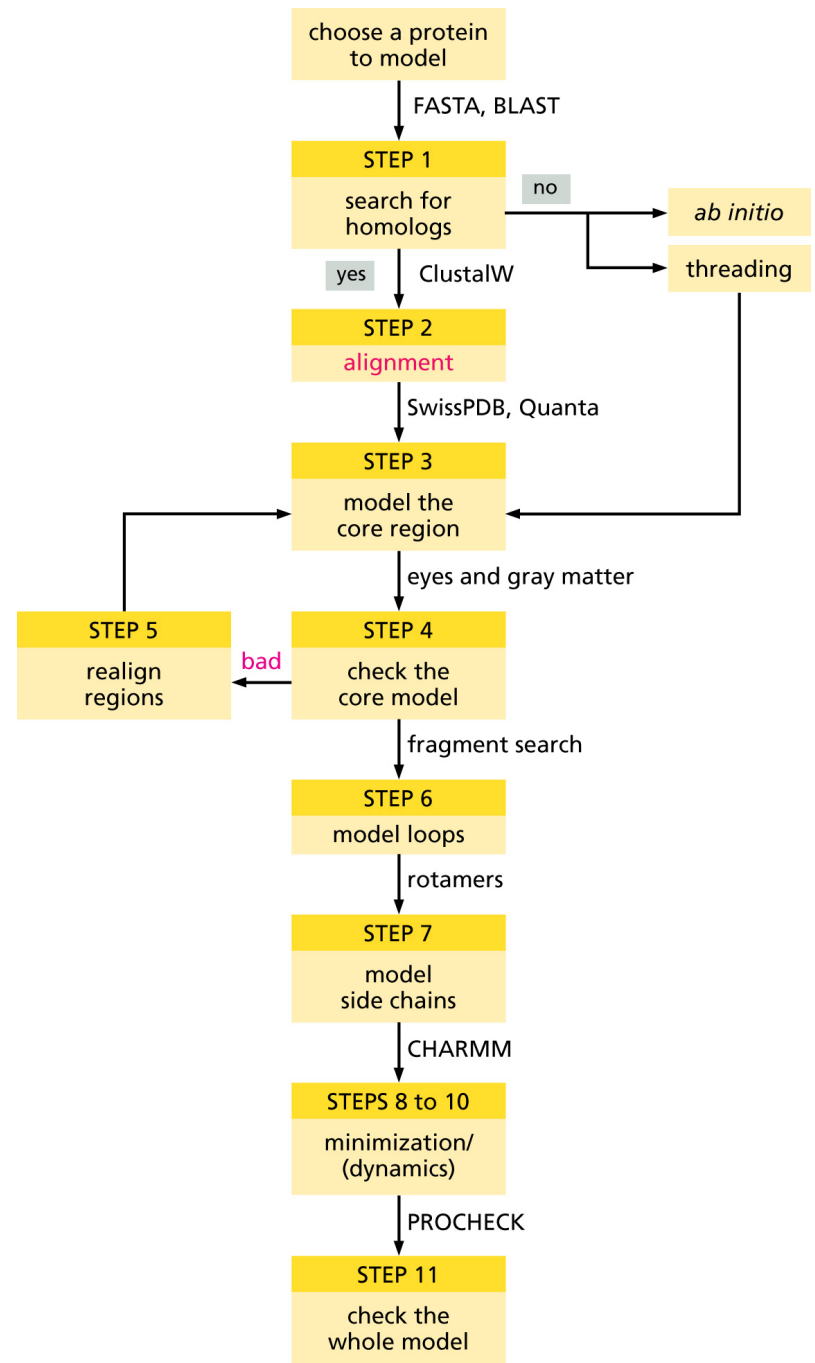# Modeling protein structures

Homology modeling

It is also called comparative modeling or knowledge-based modeling

The first homology models were built as early as the 1960s using wire and plastic models of bonds and atoms.

The first published homology model structure in 1969 was obtained for a small globular protein, the alpha-lactalbumin (**TARGET**). It was modeled on the structure of lysozyme as the **TEMPLATE**.

choose a protein to model

FASTA, BLAST

STEP 1
search for homologs

no → *ab initio*
→ threading

yes ClustalW

STEP 2
alignment

SwissPDB, Quanta

STEP 3
model the core region

eyes and gray matter

STEP 5
realign regions

bad ← STEP 4
check the core model

fragment search

STEP 6
model loops

rotamers

STEP 7
model side chains

CHARMM

STEPS 8 to 10
minimization/ (dynamics)

PROCHECK

STEP 11
check the whole model

Homology modeling: searching for homologs

▪ Studies of the relation between the sequence similarity and the 3D structure have indicated that the cut-off point for successful modeling is 25% sequence identity.

▪ Significant sequence alignment depends on the length of the sequence
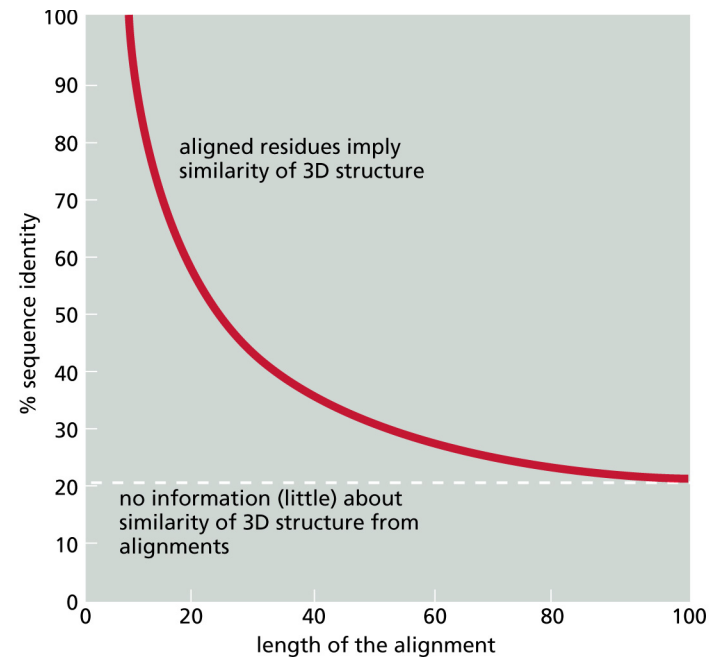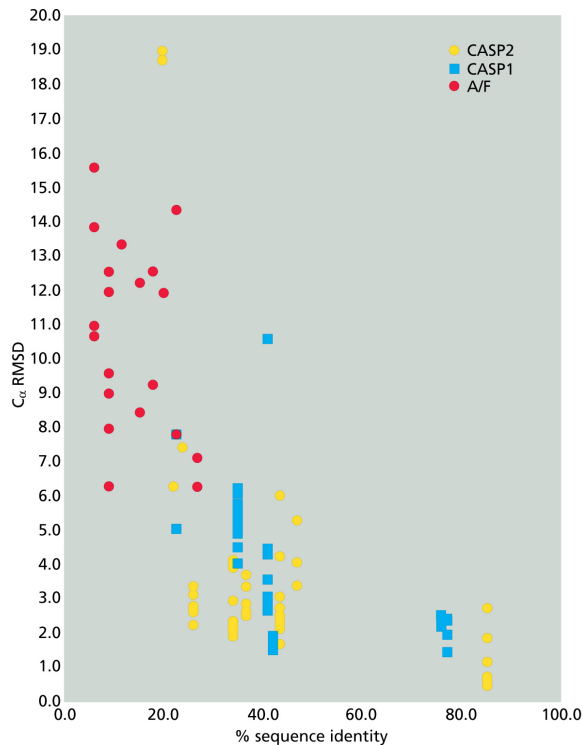
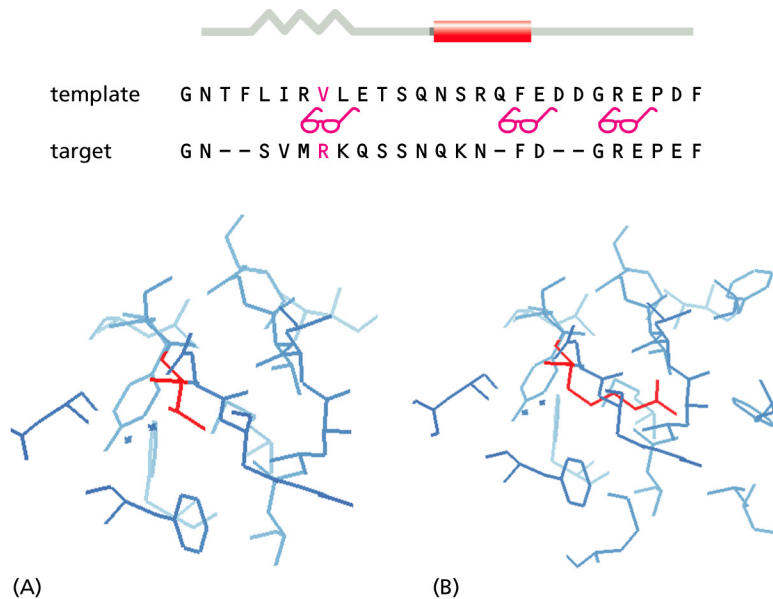In manual modeling, there are 3 options:

1. Use of the most homolog as template.

2. Use of a template that is the average of all the possible templates.

3. Use of different fragments from each structure to make up a template.

Homology modeling

ASSUMPTIONS

1. The polypeptide backbone of regions conserved between template and target have identical spatial coordinates.

- is the template similar or homolog?

- Divergent evolution may have contributed to different structures

- The spatial coordiantes will be similar but never identical.

2. Insertions and deletions in the sequence alignments will fall mainly in loop regions and considered as random coils.
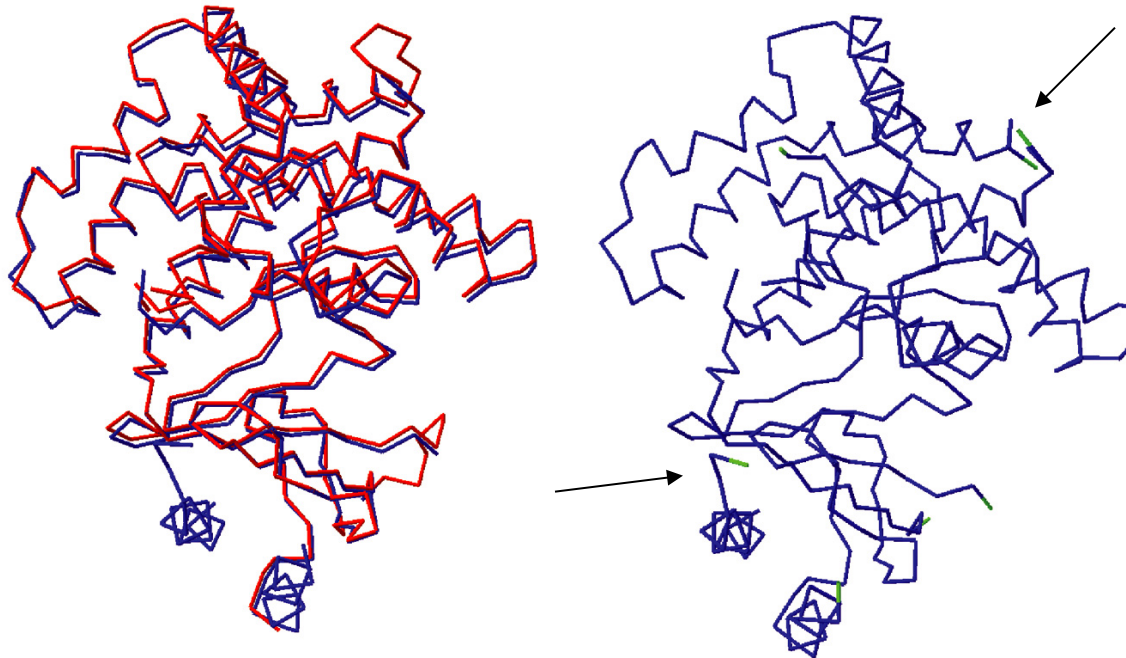
The importance of the alignment

template   G N T F L I R **V** L E T S Q N S R Q F E D D G R E P D F

target   G N - - S V M **R** K Q S S N Q K N - F D - - G R E P E F

The effect the alignment can have on a model structure.

An arginine is uncorrectly aligned to a valine. A) the valine (red) points into the hydrophobic core of the protein. B) The arginine (red) of the model will point into the hydrophobic core which is energetically unfavorable. Moreover the large side chain of arginine clashes with many other side chains.

(A)   (B)

Homology modeling

1. Structurally conserved regions are modeled first by transferring the x, y, z coordinates of every matched atom within an aligned residue from the template to the target molecule.

2. The backbone atoms are then joined together to form peptide bonds at the correct angles.

3. The modeling of backbone and side chains occurs simultaneously. At this stage insertions and deletions have not yet been modeled and the core structure is a set of discontinuous chains (see arrows).

4. The modeled core is checked for misfits now, before loop construction and energy minimization.
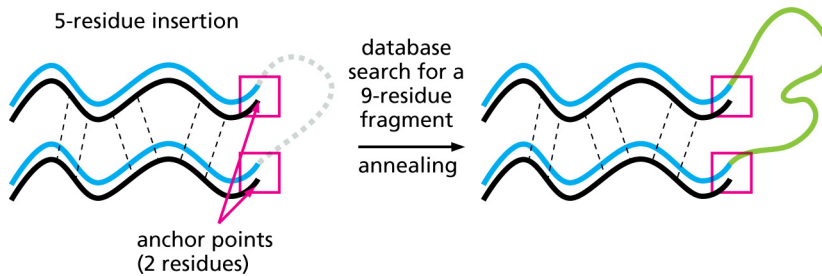
If insertions and deletions will disrupt the core or secondary structure elements, it is necessary to check and change the alignment.

Homology modeling

5. Modeling of the loops. Loops are often functionally important (binding sites).

Template: V L V A **T Y**   **H D** F V L I . . .
Target: V L I I **S Y** <u>F G N S G</u> **R E** F V I L . . .

5-residue insertion

database search for a 9-residue fragment

annealing

anchor points (2 residues)

Database search method for building a loop.

A database search of hign resolution fragments is performed for a fragment nine residues long: five for the insertion and four for the anchor points.

Ten loops have been selected on the basis of lowest RMSD for further evalutaion depending on their conformation (core-disruptive potential) and sequence homology.
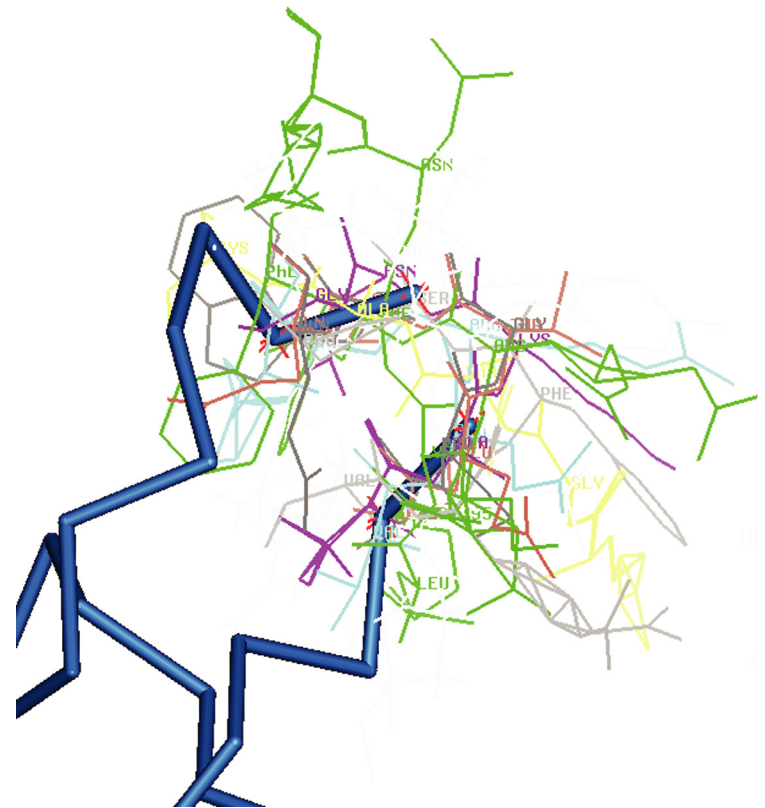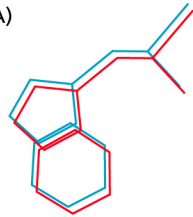
Homology modeling

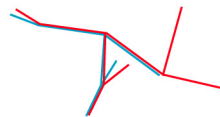6. Non identical amino acids sidechains are modeled mainly by using rotamer libraries.
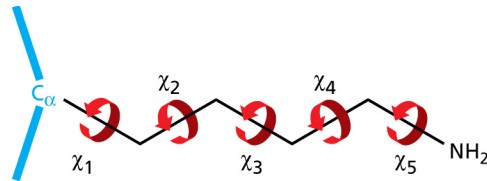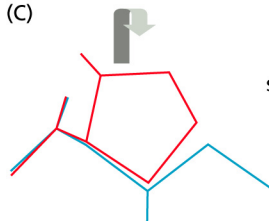

(A)

same side chain ⟶ conformer taken from template

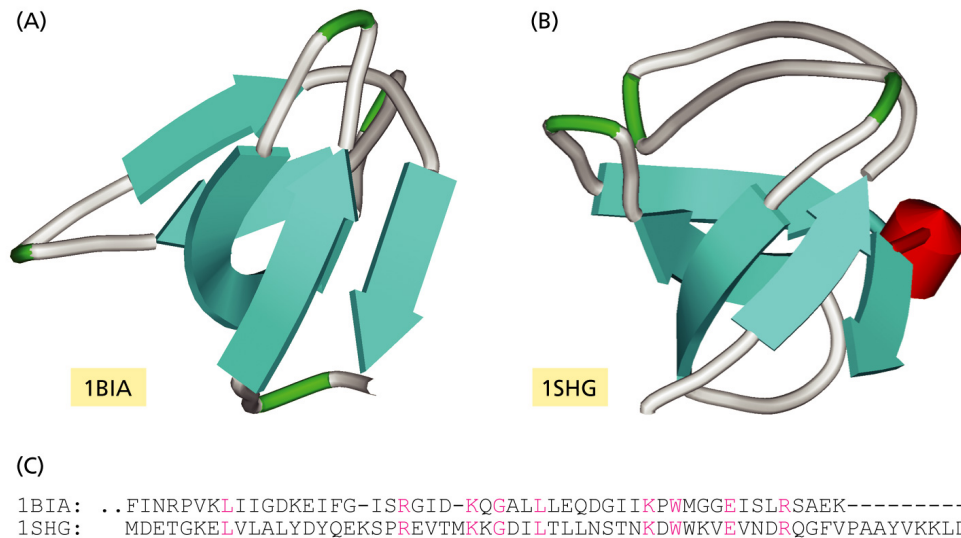(B)

partial similarity ⟶ most of side chain built on template

(C)

substitution ⟶ built based on rotamer library and energetics

Threading or protein fold recognition

If no homologous proteins with an experimentally solved structure can be found to match the target sequence, structure prediction methods that do not depend on homology have to be used.

The basis:

1. The same secondary structure elements can be formed by different many sequences.

2. This is also true for tertiary structure.

(A)

(B)

1BIA

1SHG

(C)

```
1BIA: ..FINRPVKLIIGDKEIFG-ISRGID-KQGALLLEQDGIIKPWMGGEISLRSAEK---------
1SHG:   MDETGKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
```

The ribbon representation of the structures of an SH3 domain. A) Dihydrofolate reductase (1BIA) and B) a kinase (1SHG). The sequence identity of these two domains is only 14.5%. Normal sequence alignment programs would not identify these structures as having a similar fold. C) A sequence alignment based on the structural superposition.

Threading or protein fold recognition

Because fold recognition techniques **DO NOT DEPEND PRIMARLY ON SEQUENCE COMPARISON**, a structural relationship between proteins may be recognised even if the sequence similarity is very low or non existent. This conservation of structures can be due to:

- common ancestry

- physical constraints limit the number of folds that proteins can adopt.

Therefore, **THE SAME FOLD CAN OCCUR IN A WIDE VARIETY OF DIFFERENT PROTEINS.**
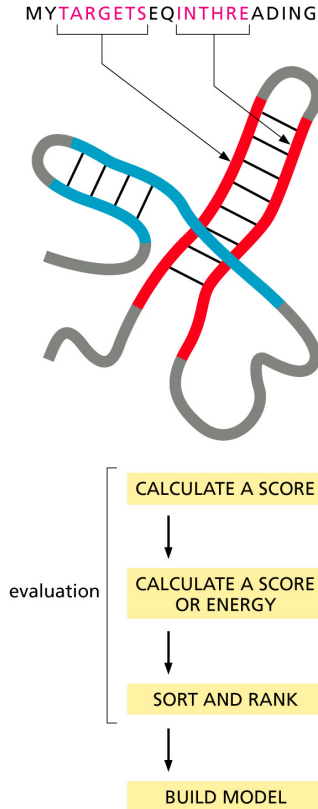
Diagramatic representation of the threading procedure.

1. First, segments of sequence are structurally aligned (threaded) on to a fold and a score/energy is obtained for each alignment.

2. A dynamic programming technique is used to find the alignment that has the best score/energy. This is done for each fold in the fold library, and the results are ranked.

3. The folds giving the best scoring results are then selected for use in modeling the query sequence.

Threading or protein fold recognition
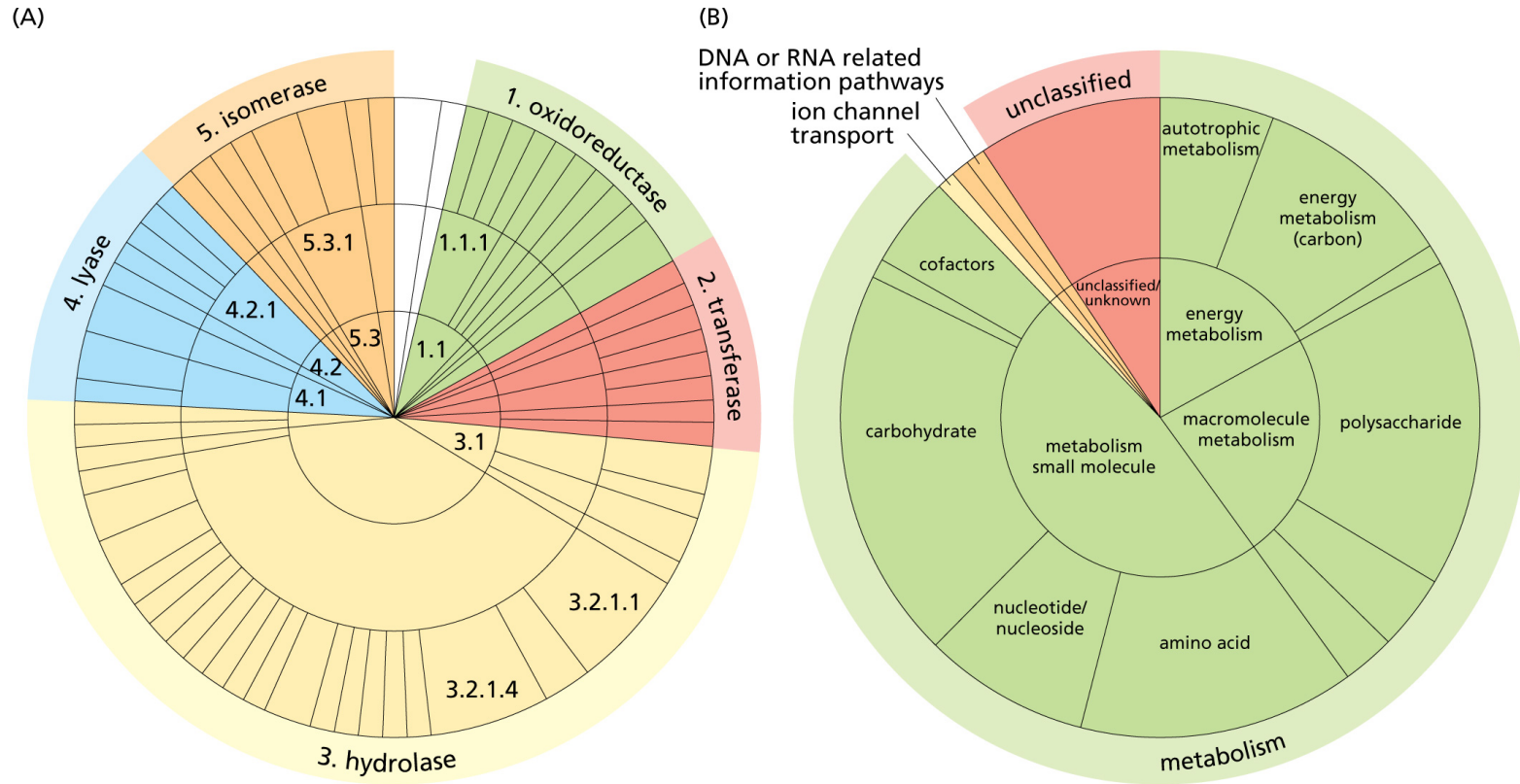
The different types of TIM barrel function illustrated. A) The wheel shows the distribution of TIM functions and the type of reaction. B) representation of the biological /pathway function of the various TIM proteins.

Protein fold databases

The PDB is not used as it includes homologous structures and proteins with similar fold.

Libraries of protein folds have been developed to reduce the number of structures to be explored.

CATH classifies protein folds according to 4 parameters:

Class, C-level Class is determined according to the secondary structure composition and packing within the structure. Three major classes are recognised; mainly-alpha, mainly-beta and alpha-beta. This last class (alpha-beta) includes both alternating alpha/beta structures and alpha+beta structures, as originally defined by Levitt and Chothia (1976). A fourth class is also identified which contains protein domains which have low secondary structure content.

Architecture, A-level

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually using a simple description of the secondary structure arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g the beta-propellor or alpha four helix bundle).

Topology (Fold family), T-level

Structures are grouped according to whether they share the same topology or fold in the core of the domain, that is, if they share the same overall shape and connectivity of the secondary structures in the domain core. Domains in the same fold group may have different structural decorations to the common core.

Some fold groups are very highly populated (Orengo et al. 1994); Orengo & Thornton, 2005) particularly within the mainly-beta 2-layer sandwich architectures and the alpha-beta 3-layer sandwich architectures.

Homologous Superfamily, H-level

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified either by high sequence identity or structure comparison using SSAP. Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria:

- Sequence identity >= 35%, overlap >= 60% of larger structure equivalent to smaller.
- SSAP score >= 80.0, sequence identity >= 20%, 60% of larger structure equivalent to smaller.
- SSAP score >= 70.0, 60% of larger structure equivalent to smaller, and domains which have related functions, which is informed by the literature and Pfam protein family database, (Bateman et al., 2004).
- Significant similarity from HMM-sequence searches and HMM-HMM comparisons using SAM (Hughey &Krogh, 1996), HMMER (http://hmmer.wustl.edu) and PRC (http://supfam.org/PRC).

CATH / Gene3D

26 million protein domains classified into 2,738 superfamilies

Browse »   Search »   Download »   Take the Tour »

What is CATH?

CATH is a classification of protein structures downloaded from the Protein Data Bank. We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence (FASTA)
- Search CATH by PDB structure

- Browse CATH Hierarchy
- CATH Release Notes
- CATH Tutorials

Example pages

- PDB "2bop"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HUPs"

- Functional Family
- FunFam Alignment
- Search for "enolase"
- Superfamily Comparison

Latest Release Statistics

CATH v4.0 based on PDB dated March 26, 2013	
235,858	CATH Domains
2,738	CATH Superfamilies
69,058	Annotated PDBs

Gene3D v12 released March 18, 2012	
6,131	Cellular Genomes
21,662,155	Protein Sequences
25,615,754	CATH Domain Predictions

Welcome to SCOP2!

Citation

Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, Alexey Murzin, SCOP2 prototype: a new approach to protein structure mining (2014) Nucl. Acid Res., 42 (D1): D310-D314. [PDF]

Description of the SCOP2 database

SCOP2 is a successor of Structural classification of proteins (SCOP). Similarly to SCOP, the main focus of SCOP2 is on proteins that are structurally characterized and deposited in the PDB. Proteins are organized according to their structural and evolutionary relationships, but, in contrast to SCOP, instead of a simple tree-like hierarchy these relationships form a complex network of nodes. Each node represents a relationship of a particular type and is exemplified by a region of protein structure and sequence.

In SCOP2, we try to put in use the knowledge we acquired over the past years and the lessons we have learned during the classification of protein structures. We believe that there are many peculiarities of proteins and their structures that have been missed due to the constraints of the original SCOP hierarchical schema. We hope that our users will find the new resource useful and that it could open new avenues for protein analysis and research.

Quick introduction on how to browse, search and download

SCOP2 offers two different ways for accessing data: SCOP2-browser, that allows navigation through the SCOP2 classification in a traditional way by browsing pages displaying the node information, and SCOP2-graph, which is a graph-based web tool for display and navigation through the SCOP2 classification. Both tools provide search of SCOP2 data by free text, node names, IDs, tags and keywords, as well as external identifiers associated with them, e.g. PDB and UniProt. SCOP2 data can also be retrieved via REST interface or downloaded from the SCOP2 Download page. For more information visit the About page.

Web browser compatibility check

To test whether your web browser and its settings are suitable to view SCOP2-graph and to visualize protein structures using Jmol applet click here.

Search Browser

[   ] [Search]
Add an asterisk to search free text (e.g. serine*)

Search Graph

[   ] [Search]
Add an asterisk to search free text (e.g. protein*domain)

Threading: scoring schemes

1. Scoring matrices that take into account the likelihood of a substitution given the nature of the environment.

2. Scoring matrices that include details of the structure in the vicinity of each residue, involving inter-atomic distances or numbers of residues within a specified distance.

The substitution of a lysine with an alanine would cause a cavity formation and a hydrophobic residue in a polar environment
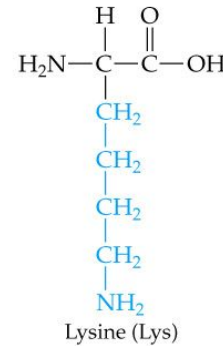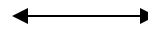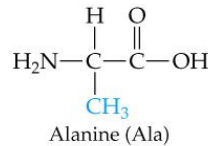
The substitution of an alanine with a lysine would cause the fold not to have a sufficient space or polar environment to accommodate the lysine
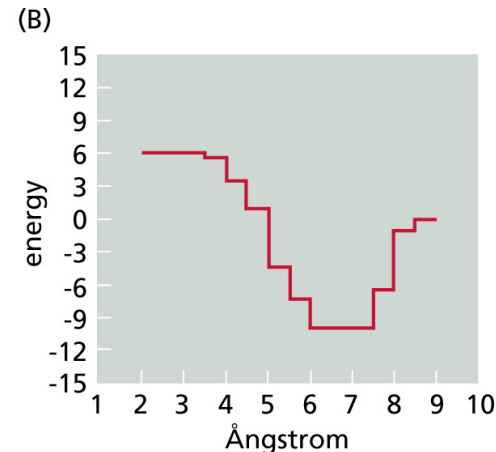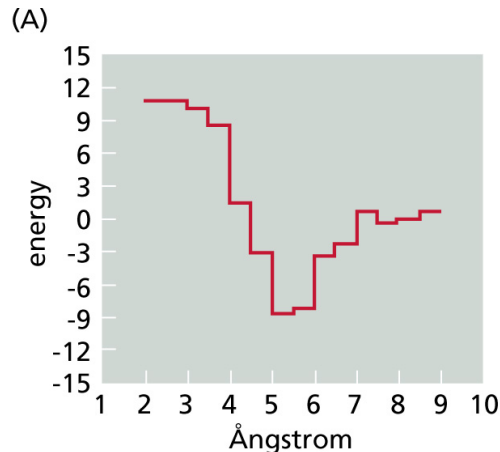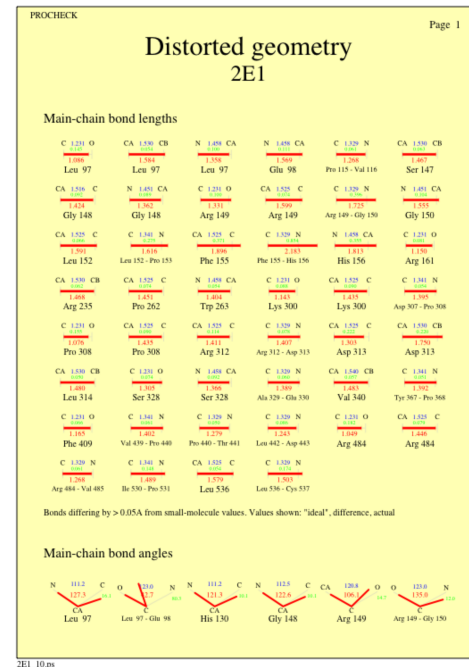
An example of energy terms derived from observed protein structures, as used in the threading programs such as LOOPP. The plots show the interaction energy for a specific pair of amino acids as a function of distance. A) Interaction energy for Val-Leu residue pairs. B) Interaction energy for Phe-Trp residues pair.

Model validation using potential energy functions

-The individual energy terms can be used to analyze the structure.

- If anything deviates from the parameters defined by the force field used, it will have a strongly unfavourable energy.

- A detailed analysis of the energy terms for a given conformation may also reveal particular interactions to be highly stabilizing or destabilsing, giving insight into the molecular function.

Procheck is a software searching for unfavorable regions based on the stereochemical geometry validation.

Model validation using potential energy functions

(A) Ramachandran plot quality assessment
(B) peptide bond planarity - omega angle xd
(C) measure of bad non-bonded interactions
(D) α carbon tetrahedral distortion
(E) hydrogen-bond energies
(F) overall G-factor
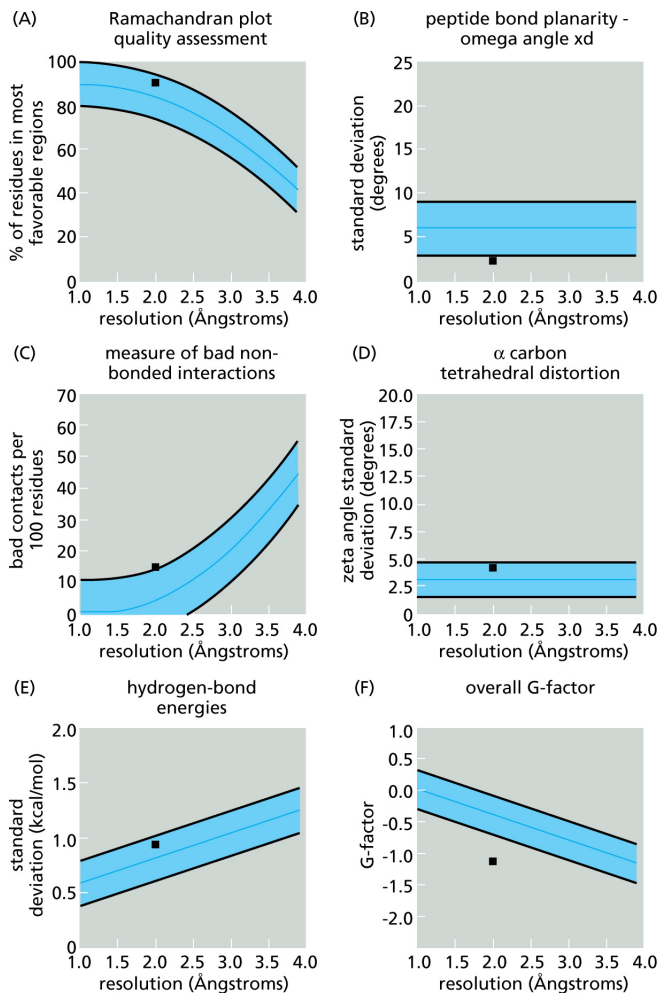
An example output of the main-chain parameters for an SH2 model from PROCHECK. Parameter values that falls within the blue band are within the expected measures for a structure at that particular resolution. The black square indicates where the predicted parameter falls.
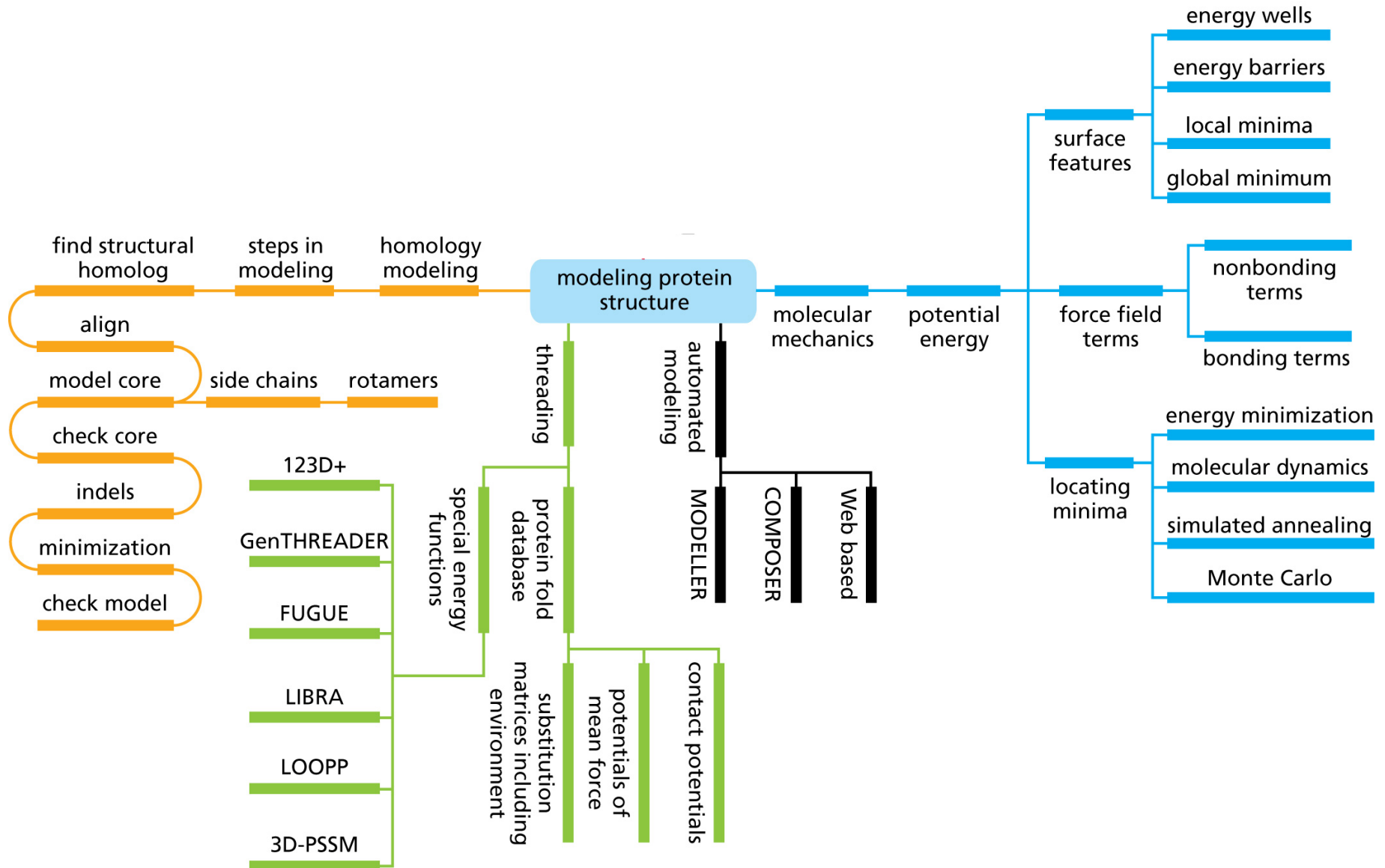
PLOT STATISTICS

| | | | comparison values | | no. of | |
stereochemical parameter	no. of data pts	parameter value	typical value	band width	band widths from mean	
(A) % residues in A, B, L	256	71.9	83.8	10.0	-1.2	WORSE
(B) omega angle at dev	281	0.6	6.0	3.0	-1.8	BETTER
(C) bad contacts/100 residues	3	1.1	4.2	10.0	-0.3	inside
(D) zeta angle at dev	262	0.9	3.1	1.6	-1.4	BETTER
(E) H-bond energy at dev	173	0.7	0.8	0.2	-0.5	inside
(F) overall G-factor	282	0.2	-0.4	0.3	2.0	BETTER

Modeling protein structures

find structural homolog — steps in modeling — homology modeling — **modeling protein structure**

align

model core — side chains — rotamers

check core

indels

minimization

check model

123D+

GenTHREADER

FUGUE

LIBRA

LOOPP

3D-PSSM

special energy functions

threading

protein fold database

substitution matrices including environment

potentials of mean force

contact potentials

automated modeling

MODELLER

COMPOSER

Web based

molecular mechanics — potential energy — force field terms

surface features
- energy wells
- energy barriers
- local minima
- global minimum

force field terms
- nonbonding terms
- bonding terms

locating minima
- energy minimization
- molecular dynamics
- simulated annealing
- Monte Carlo

Ab initio methods

If no homologous proteins with an experimentally solved structure can be found to match the target sequence, structure prediction methods that do not depend on homology have to be used.

■ They are methods that predict the structure from first principles using thermodynamic and physicochemical theory

■ They require identification of the conformation of the global energy minimum without having any prior fold information to bias the search

■ All the possible conformations of a protein sequence should be evaluated to identify the minimum energy structures.

■ In practice only a subset of conformations are sampled.

■ They require powerful computer to increase computational speed

■ Single domain proteins have been successfully predicted

■ Intermediate methods between homology modeling and ab initio methods

6.7 Å rmsd   5.6 Å rmsd

crystal structure

2.7 Å rmsd   4.5 Å rmsd
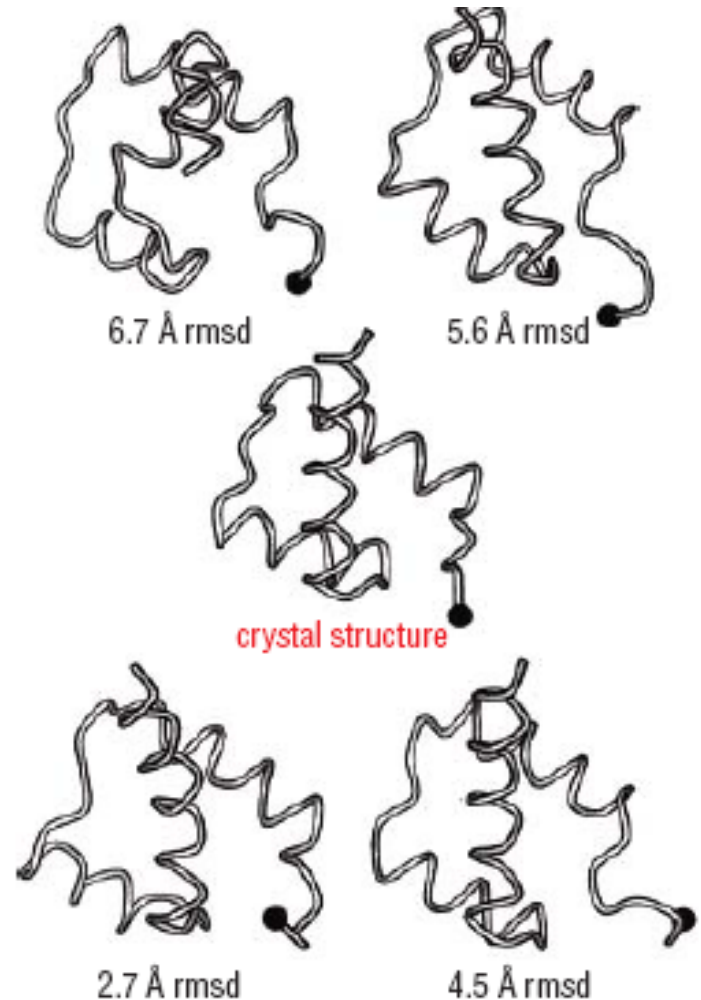
Potential energy function and force field

▪ When modeling a protein structure, the aim is to obtain a structure of **lowest possible energy** that satisfies the known stereochemical constraints on protein structures such as allowable values for backbone torsion angles ϕ and ψ and appropriate packing of side chains.

▪ The geometry of a protein conformation, in terms of its atomic coordinates, is related to its potential energy (enthalpy) by means of a collection of equations known as **potential energy functions**.

▪ They represent all the components that contribute to the overall potential energy of the protein. The combination of all these energy funtions for a given conformation is called the **force field**.

Types of force fields.

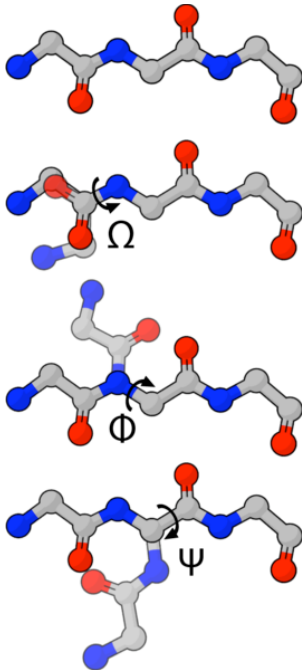1. The potential energy of a given conformation might include other molecules such as solvent

2. The potential energy of a given conformation statistically averaged environmental effects
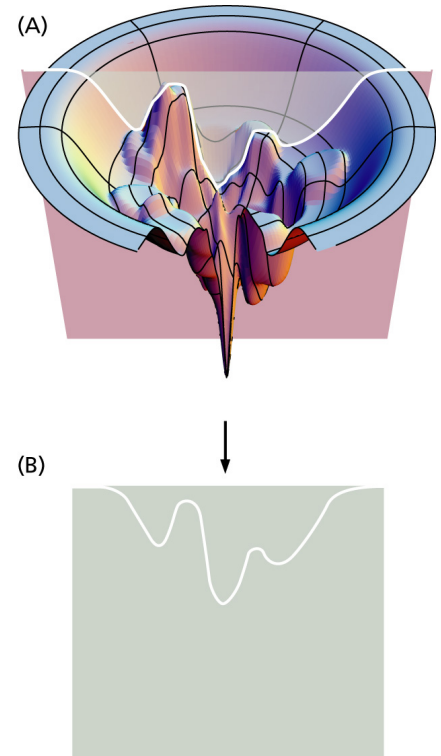
Potential energy function and force field

(A)

(B)

Molecular system will be found in those conformations that have
the lowest free energy

$$\Delta G = \Delta H - T\Delta S$$

$$\Delta G < 0$$

Ω

Φ

Ψ

The dihedral in proteins

When modeling protein structures, the entropic component is assumed
to be constant and only the potential energy is calculated.

$$E = E_{\text{bonds}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{non-bonded}}$$

$$E_{\text{non-bonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

Potential energy function and force field

$$E = E_{\text{bonds}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{non-bonded}}$$

$$E_{\text{non-bonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

The collection of algebric terms and parameters in both the bonding and nonbonding components is usually referred to as a **force field**.

The force fields terms give a relative energy so that use different terms and differ in parameterization.

### The AMBER force field is widely used for proteins

$$V(r^N) = \sum_{\text{bonds}} \frac{1}{2} k_b (l - l_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2$$
$$+ \sum_{\text{torsions}} \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \epsilon_{i,j} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

Amber is developed in an active collaboration of _David Case_ at Rutgers University, _Tom Cheatham_ at the University of Utah, Tom Darden at NIEHS (now at OpenEye), _Ken Merz_ and _Adrian Roitberg_ at Florida, _Carlos Simmerling_ at SUNY-Stony Brook, _Ray Luo_ at UC Irvine, _Junmei Wang_ at UT Southwestern, and _many others_. Amber was originally developed under the leadership of Peter Kollman.
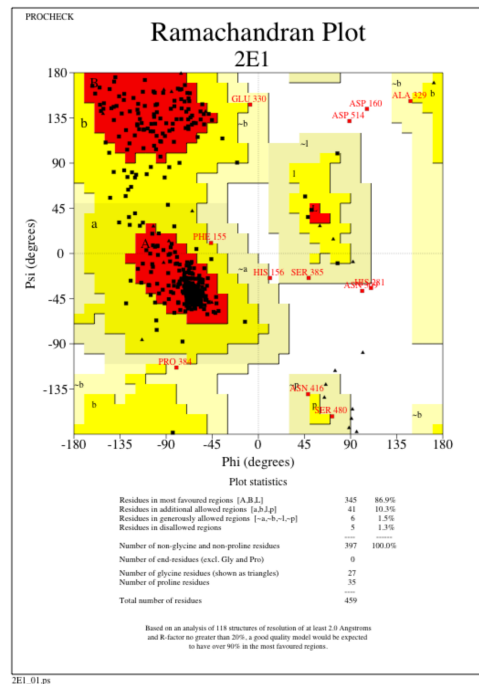
The potential energy surface

It is a surface representing the variation of the potential energy as the protein conformation varies. The Ramachandran plot is an example of potential energy surface.
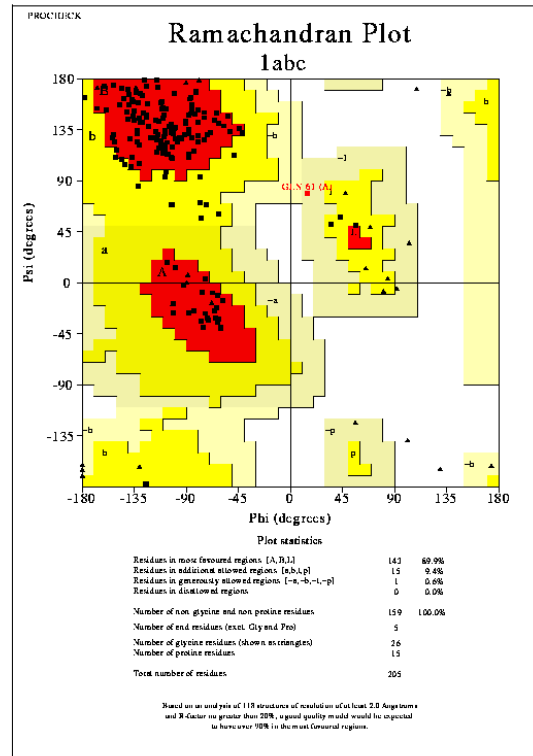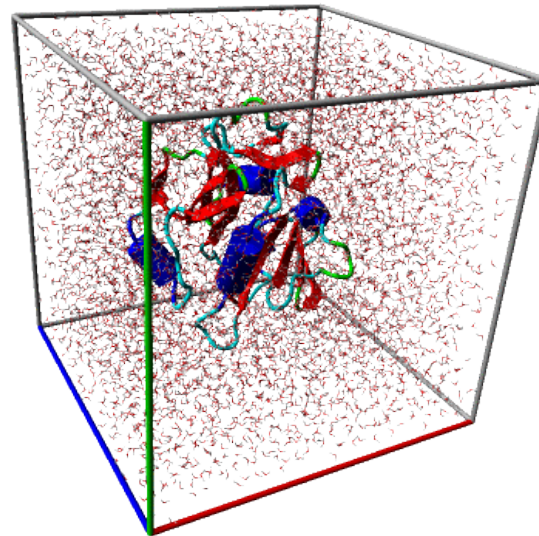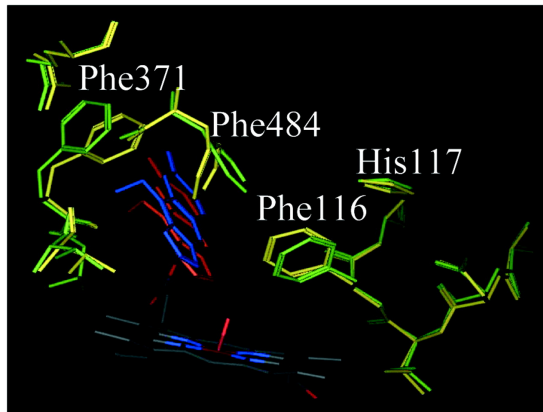
Model validation using potential energy functions

-The individual energy terms can be used to analyze the structure.

- If anything deviates from the parameters defined by the force field used, it will have a strongly unforable energy.

- A detailed analysis of the energy terms for a given conformation may also reveal particular interactions to be highly stabilizing or destabilysing, giving insight into the molecular function.

Procheck is a software searching for unfavorable regions based on the stereochemical geometry validation.
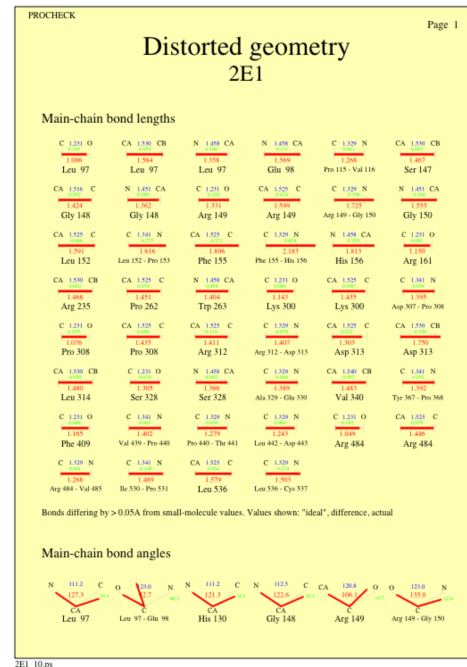
Molecular mechanics

Molecular mechanics techniques can be used on the model to remove unfavorable interactions and improve the molecular geometry identifying energy minima.

(A)

(B)

Energy gradient

Local minimum energy

Global minimum energy

Phe371

Phe484

His117

Phe116

Simulated Annealing can escape local minima with chaotic jumps

Profit

Global Maximum

Local Maximum

Parameter 2

Simulated Annealing Path Start

Jump 2

Jump 1

Parameter 1
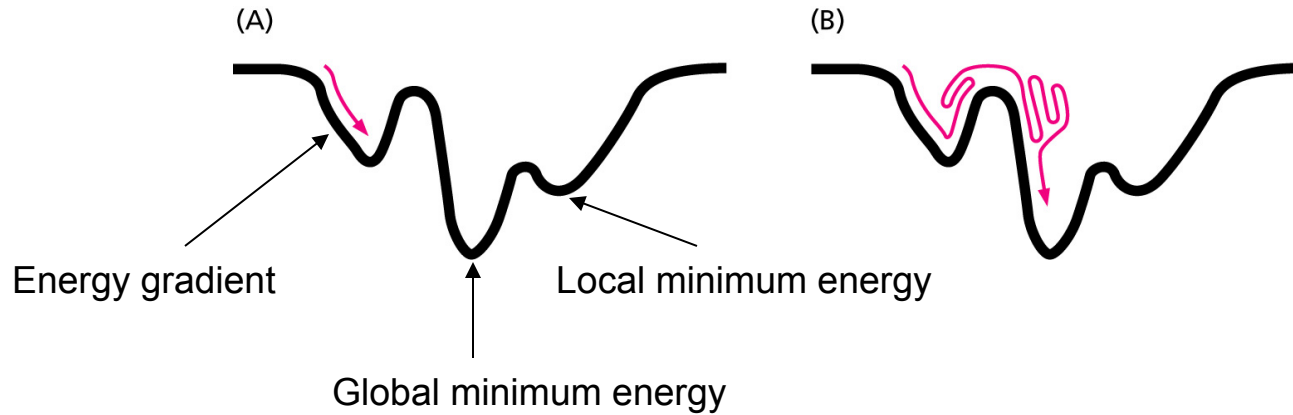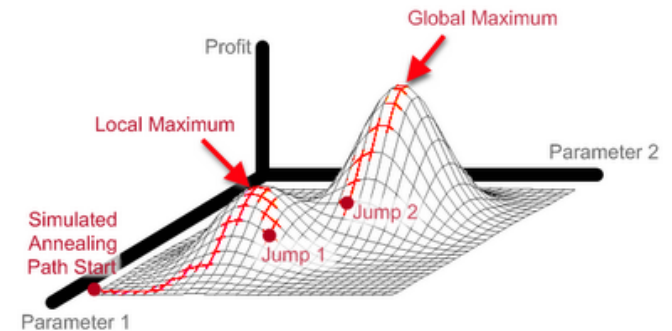
Molecular mechanics

Energy minimization: many steps allow the the modification of the protein model conformation to give a new one with lower energy.

It allows side chains in the protein core to be relaxed so they can pack together without overlapping.

Limits: it locates only local energy minima while the global energy minimumis required to reach the correct prediction.

Energy gradients are calculated

Molecular dynamics: involves solving the equations that predict the motion of the atoms over time.

Taking into account that at any point of time, the atoms in the system have defined positions and velocities, an estimate is needed to determine its position a short instant of time later.

Usually a time step of femtosecond is used and the calculation run for several tens of thousands of steps.

They allow the protein to cross small energy barriers escaping local minima.

Thus, they increase the chance of reaching the global minimum energy.

The temperature is mantained constant.

Simulated annealing: the temperature is varied during the run

First, a very high temperature is used (1000 K) so that energy barriers are crossed due to the vibrational energy of the system.

Then, the temperature is gradually decreased and the system is trapped into wells, hopefully leading to the global minimum energy.

Phe371
Phe484
His117
Phe116

Simulated Annealing can escape local minima with chaotic jumps

Global Maximum

Profit

Local Maximum

Simulated Annealing Path Start

Jump 1

Jump 2

Parameter 2

Parameter 1