Biophysics and structural bioinformatics I

D. Gilis & M. Rooman

Unité de bioinformatique génomique & structurale

UD3.203/UD3.204 (bâtiment U, campus du Solbosch)

Tél: 02/650.36.15 - 02/650.20.67

e-mail: dgilis@ulb.ac.be / mrooman@ulb.ac.be

Documents at: http://uv.ulb.ac.be

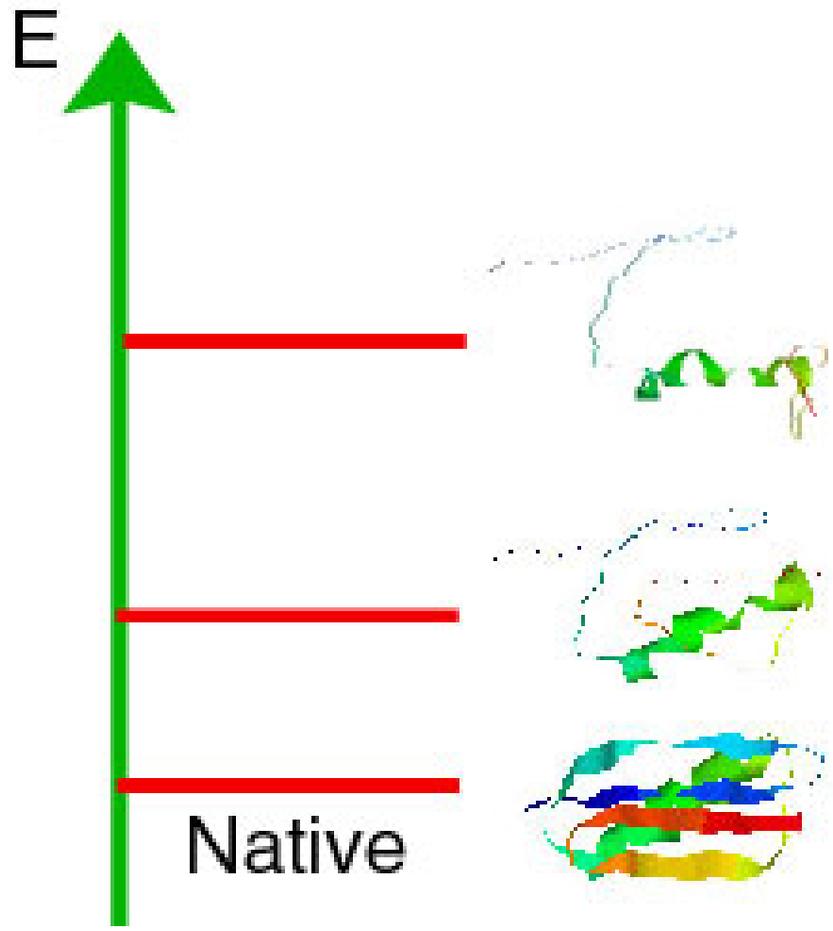
Part 7

1. Introduction

Prediction of the 3D structure of proteins

The aim is to find the 3D structure of a given protein sequence.

The energy function or the scoring function will be used to rank the different possible conformations.



The main problem is that there exists a huge number of possible conformations. The protein will adopt, in general, one conformation, that corresponds to the free energy minimum (Anfinsen experiments).

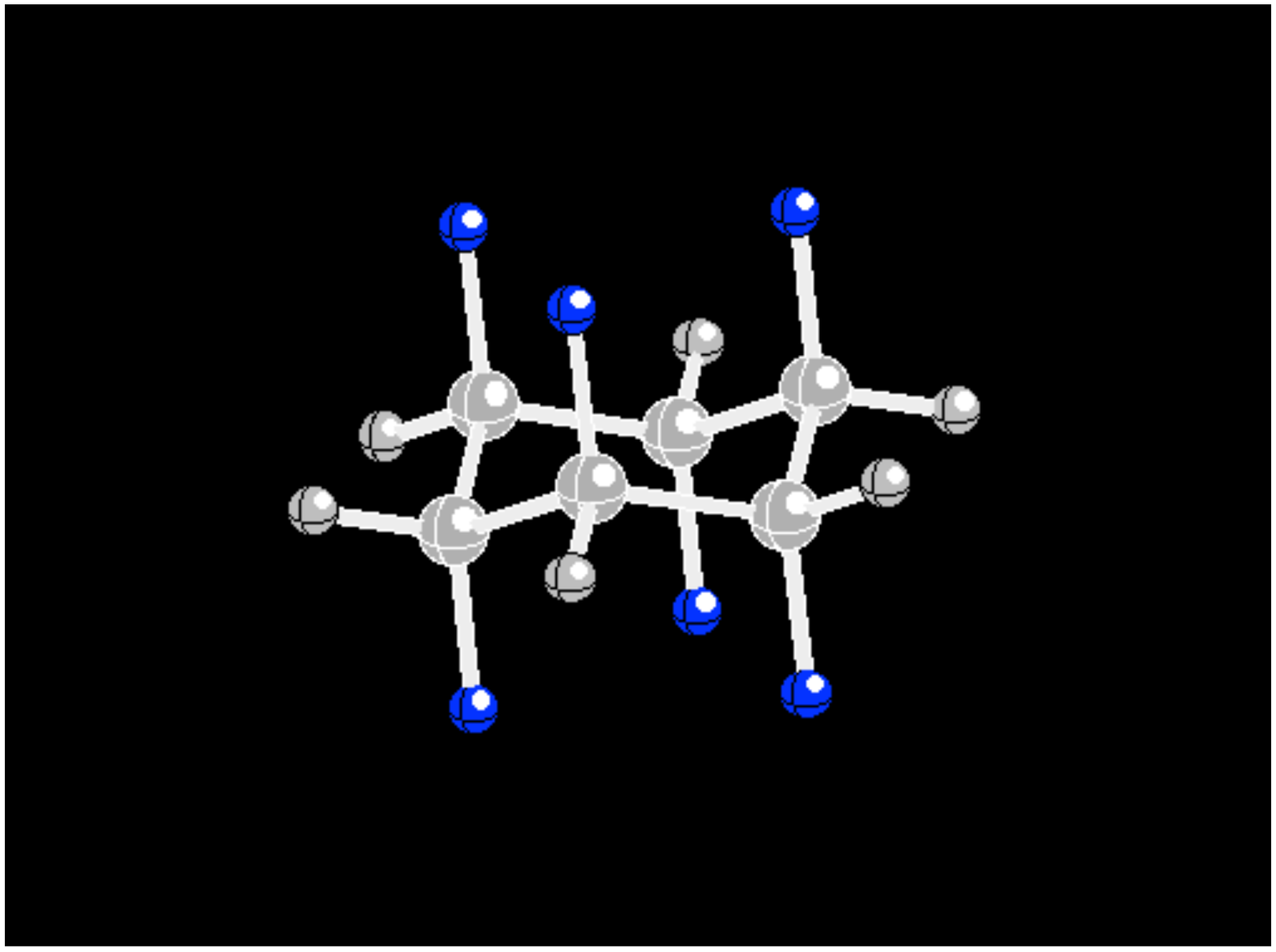
=> How to explore the different possible conformations ?

What is the conformational space ?

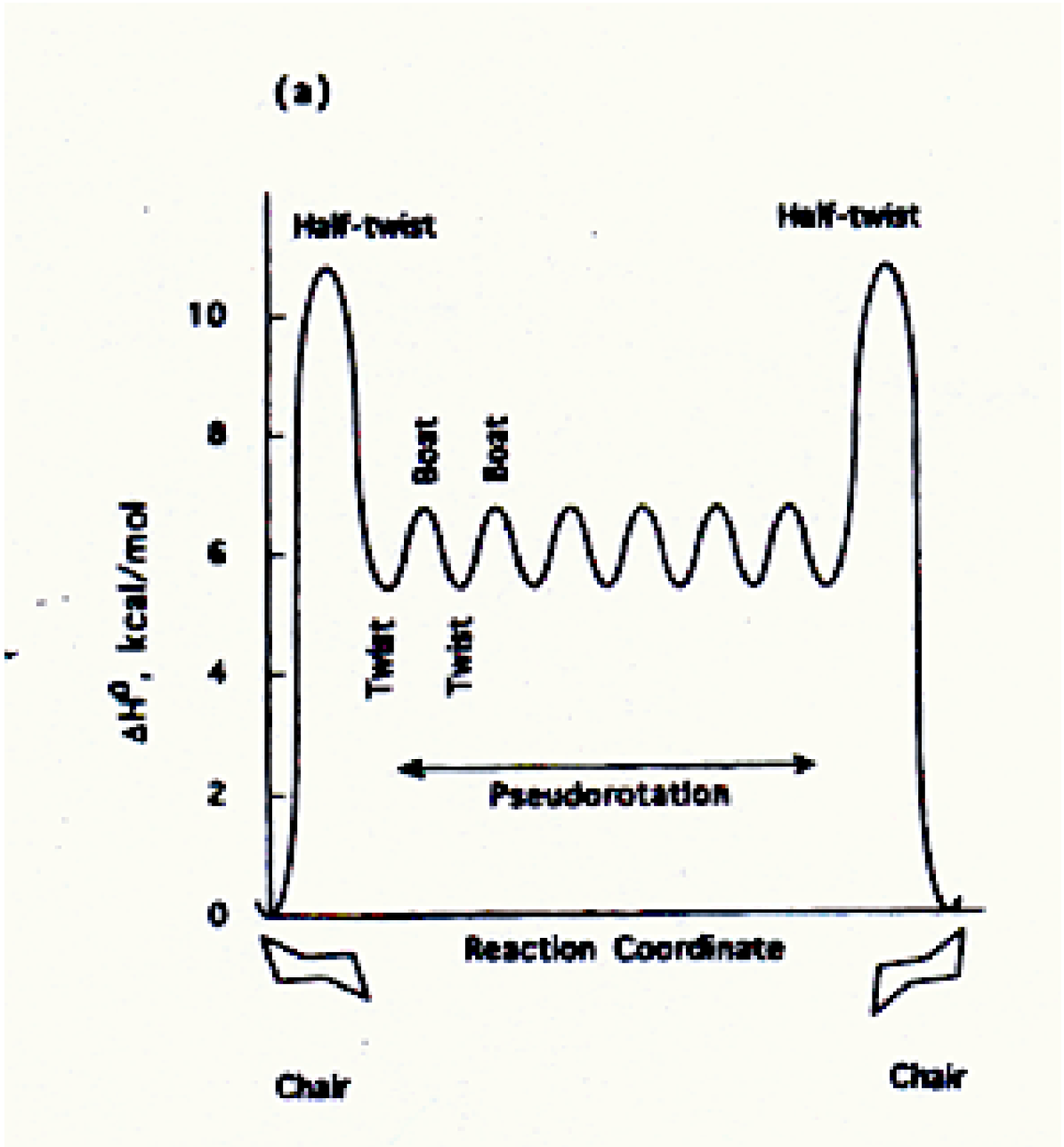
It is a multidimensional hypersurface obtained by plotting the (free) energy versus the coordinates describing the conformation of the system.

For instance, for a system composed of N atoms, the energy can be plotted versus the 3N cartesian coordinates. Generally, a projection along some coordinates is used to visualize the surface and to show the minima, the transition states, ...

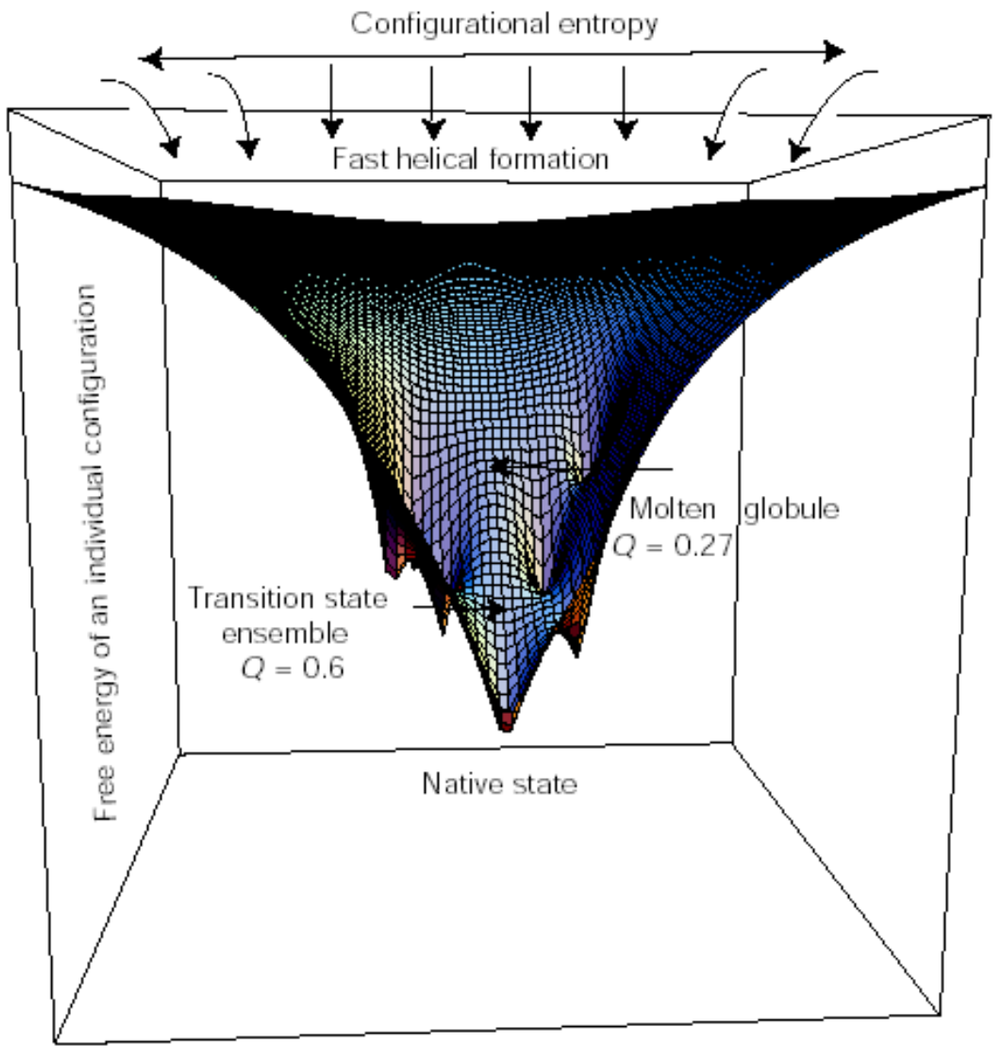
Example: cyclohexane

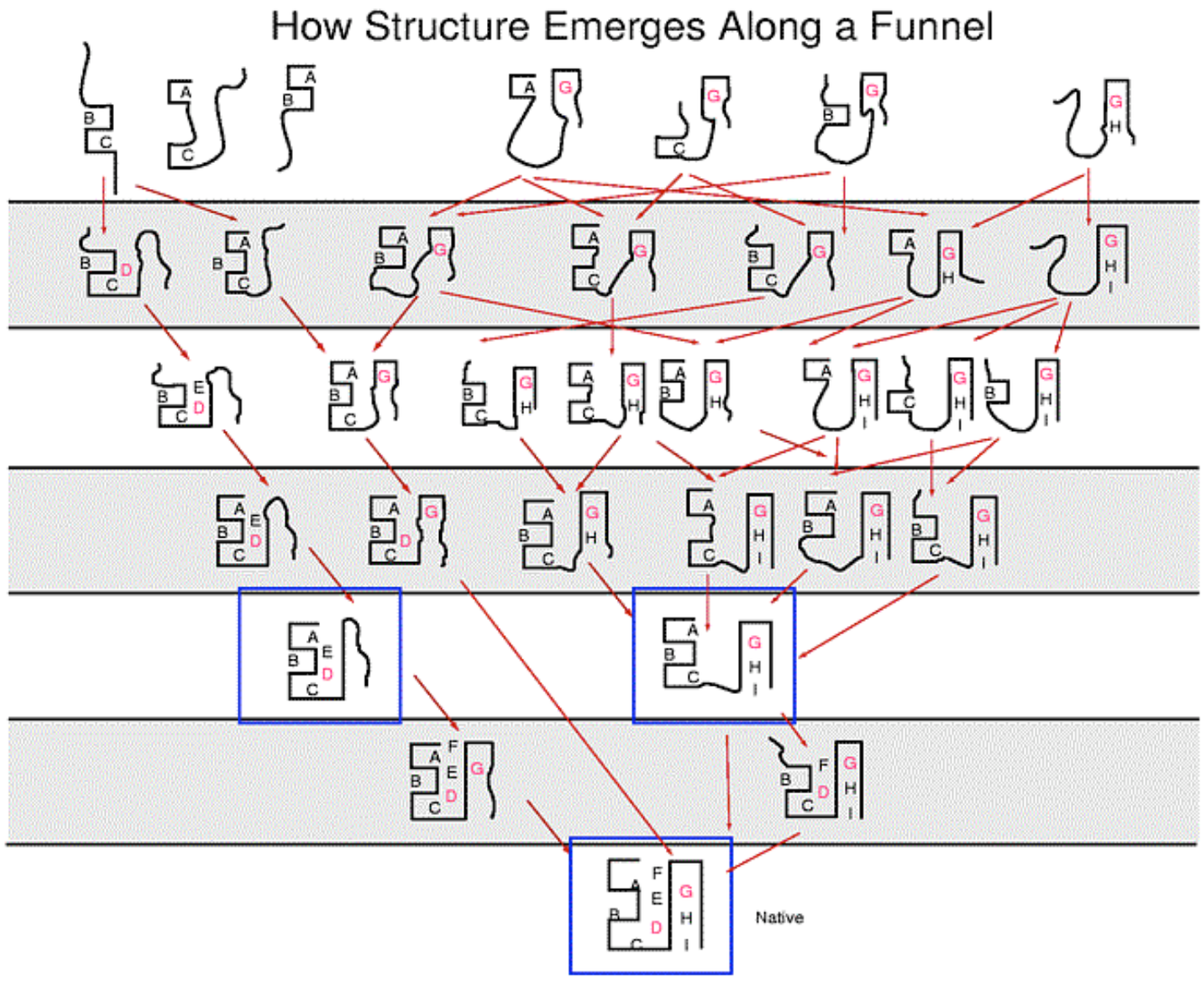


Conformational space of cyclohexane



The conformational space of proteins is more complex. It contains a large number of local minima.





The problem facing protein structure modeling:

to find the native conformation among a huge number of possible conformations, for a given sequence.

Several approaches are possible

Search for the native conformation among all the possible conformations, on the basis of energetic criteria

Ab initio modeling:

explore the conformational space

Considering that it is possible to model the structure of the protein on the basis of the existing experimentally resolved proteins

Fold recognition

Comparative modeling

avoid a search among all the possible conformations. Use the experimentally resolved structures to build a model

2. Comparative modeling

Principle:

This method is based on the fact that the space of the possible conformations is smaller than the space of the possible sequences.

Indeed, an analysis of databases of structural domains shows that similar sequences adopt a similar structure.

Let the **target** be the protein of unknown structure

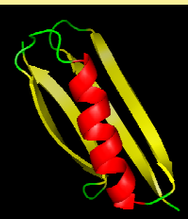
One will search in a database of experimentally resolved structures for proteins that share a "high" sequence identity with the target. These structures will be the **templates**.

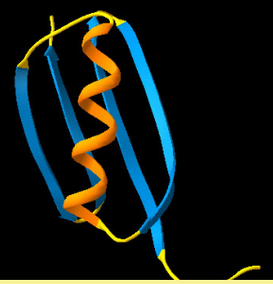
target

M V G L I W A Q A T S G V I G R G G D I P

database of possible templates

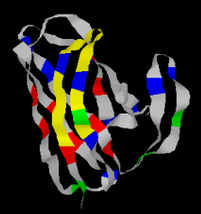
M V G F I W A N A G G E Q I S R I Q

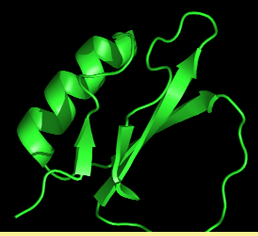




K E L L L S T L S P E Q L V L T

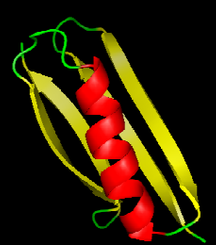
L A R N A N V A T G L E G E E N A





V V I G I D L A I A D G A F T V Y T T G

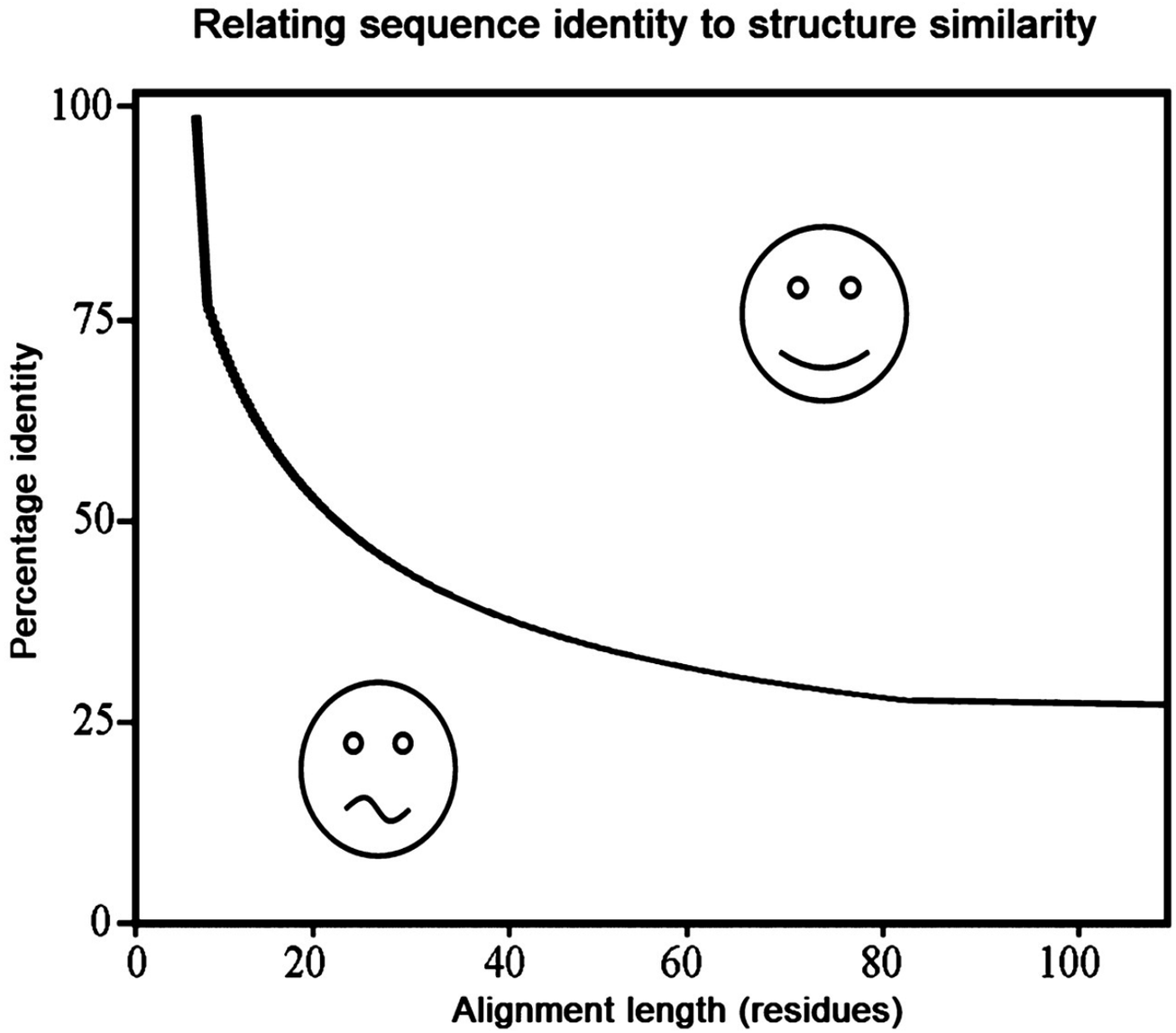
Building of a model of the structure of the target on the basis of template 1

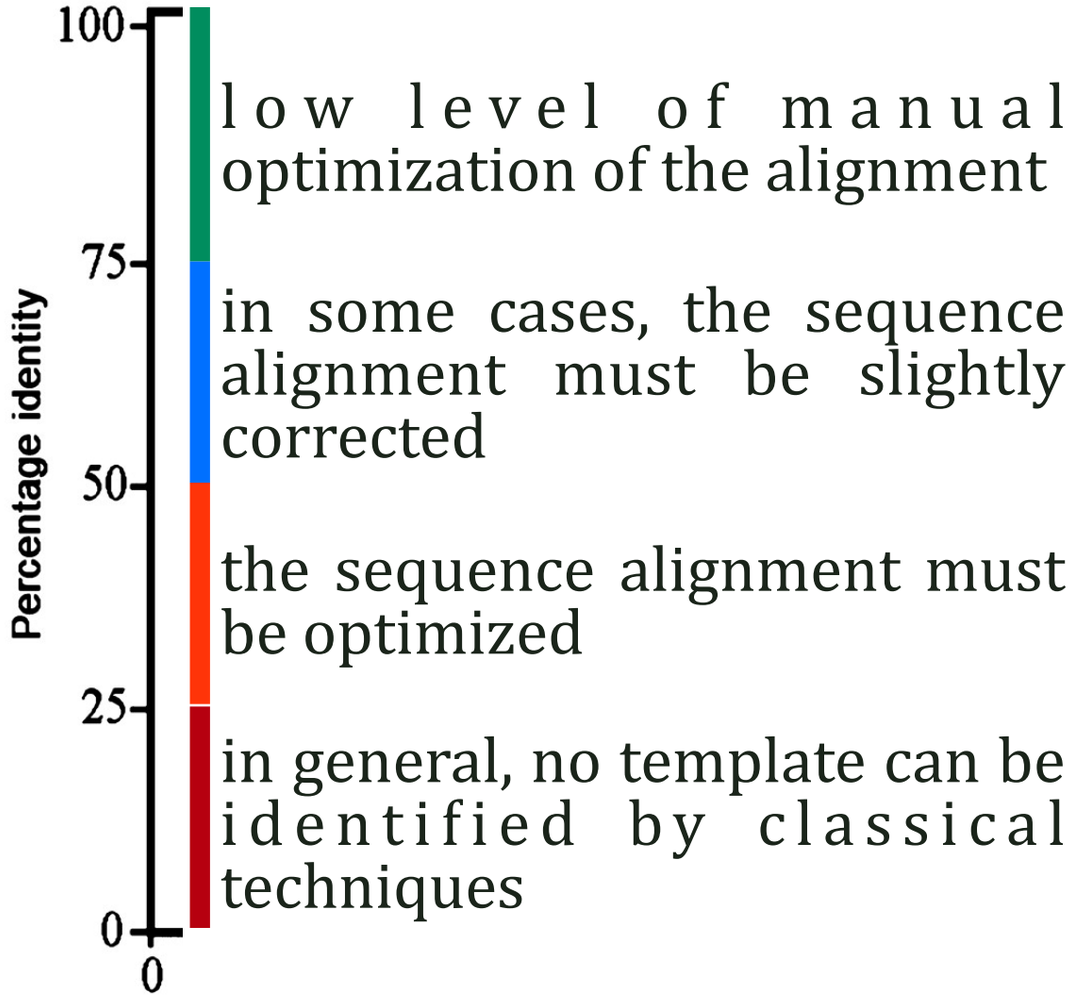


	Sequence alignments	similarity
target	-MVGLIWAQATSGVIGRGGDIP	
template1	-MVGFIWANAGGEQISR---IQ	50
template3	-LARN--ANVATGLEGE--ENA	6
template4	VVIGIDLAIADGAFTVY--TTG	17
template2	--KELLLSTLSPEQLVL---T-	5

2.1. Selection of a template and sequence alignment

The aim is to find one or several sequences whose structure is known, and that share sequence similarity with the target sequence. For that purpose, BLAST can be used for instance.

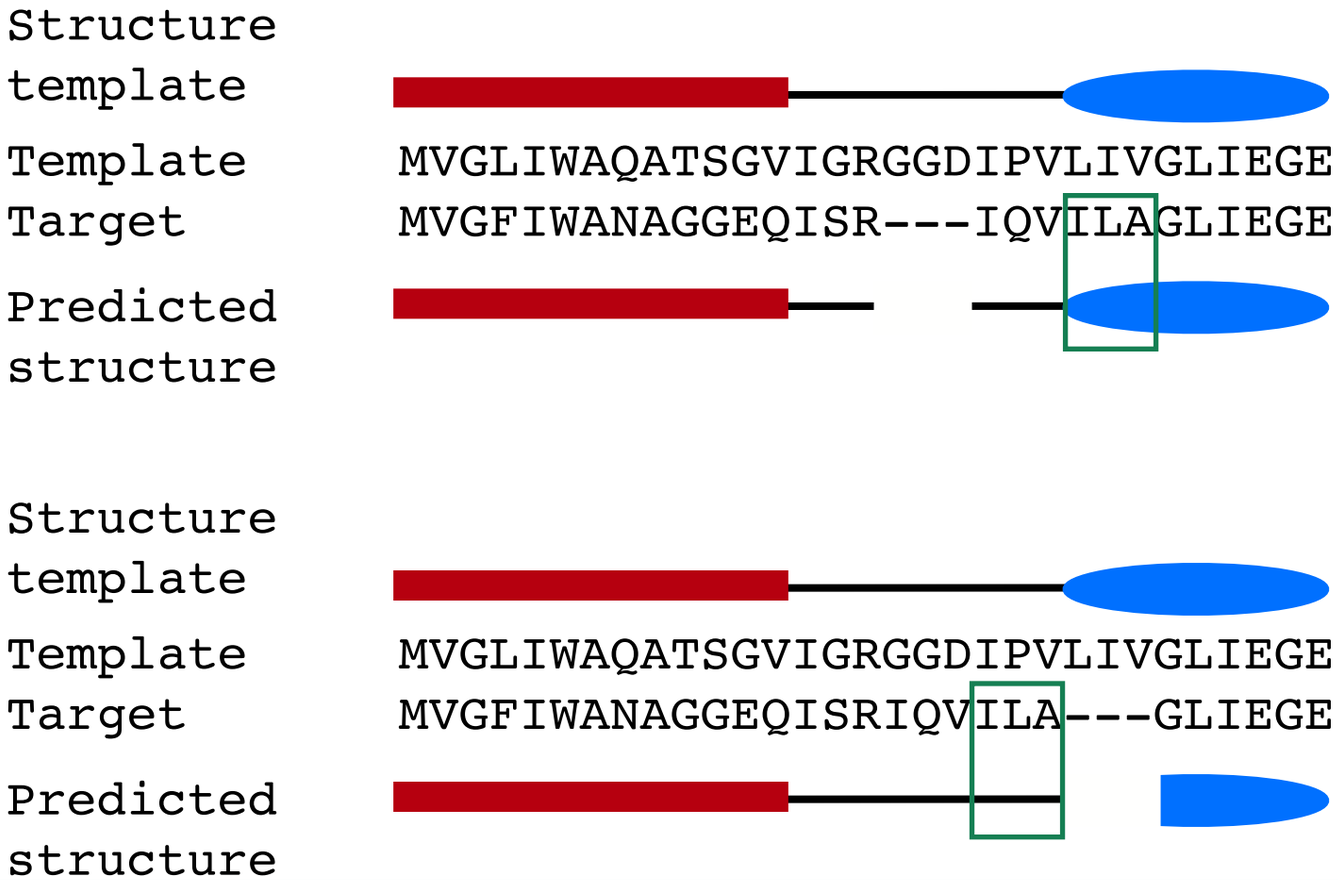




Joosten R P et al. Nucl. Acids Res. 2011;39:D411-D419

The selection of the template is based on the sequence identity between the target and the template, the quality of the experimental structure of the template, the experimental conditions used to obtain the structure of the template (pH, ionic strength, ...), in relationship with the conditions of the structure to model.

The sequence alignment is crucial: it identifies the corresponding regions between the target and the template and allows to identifies which part of the structure of the template must be used to model a given region of the target.



! Modeling of membrane proteins !

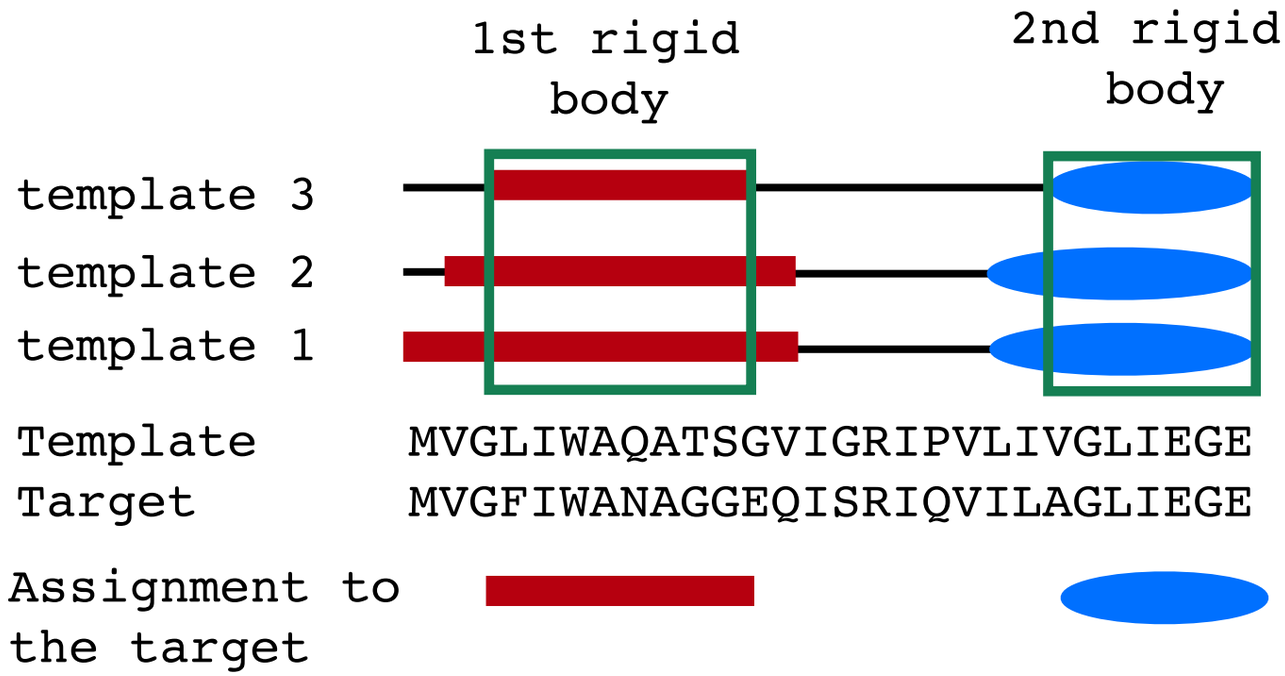
2.2. Modeling of the main chain

There exists mainly two approaches in comparative modeling:

◉ Rigid body assembly: it consists in assembling a small number of rigid bodies, that correspond to the most conserved sequence regions (between the template and the target).

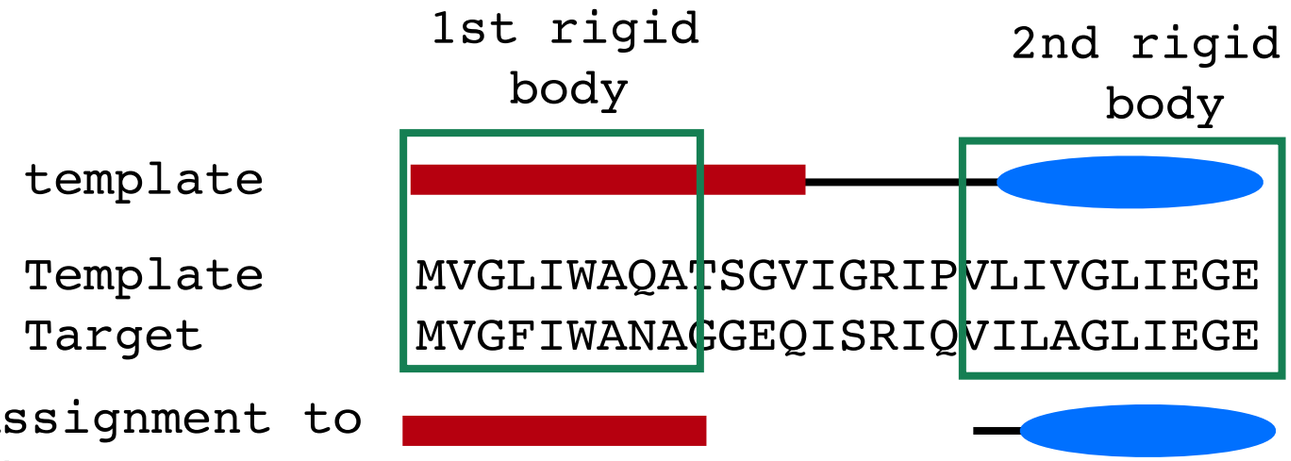
Rigid body assembly

• If several templates are selected, their structures are superimposed. The similar regions, according to the structure, will define the rigid bodies. The sequence alignment between the target and the template allows to assign these rigid bodies to the target.



Rigid body assembly

- If one template is selected, the rigid bodies are defined as the regions that present the largest sequence identity between the template and the target.

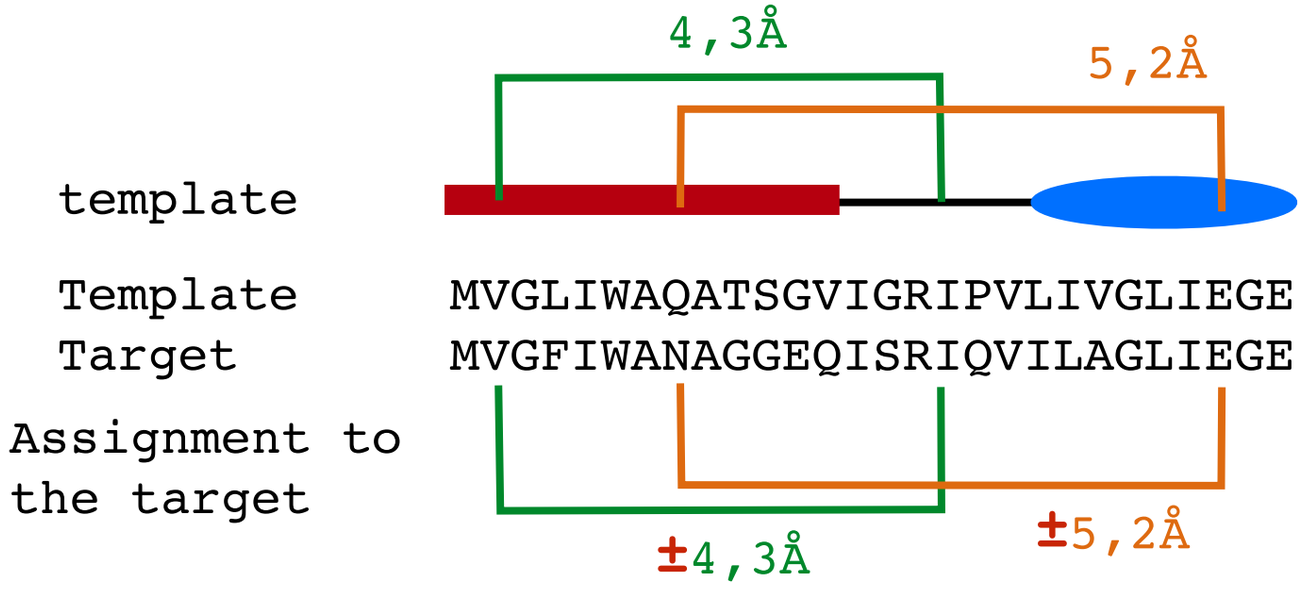


- Modeling by satisfaction of spatial restraints.

The restraints are derived from the structure(s) of the selected template(s); they correspond to bond lengths, angles values, interatomic spatial distances, ...

The sequence alignment is used to transfer the restraints from the template(s) to the target.

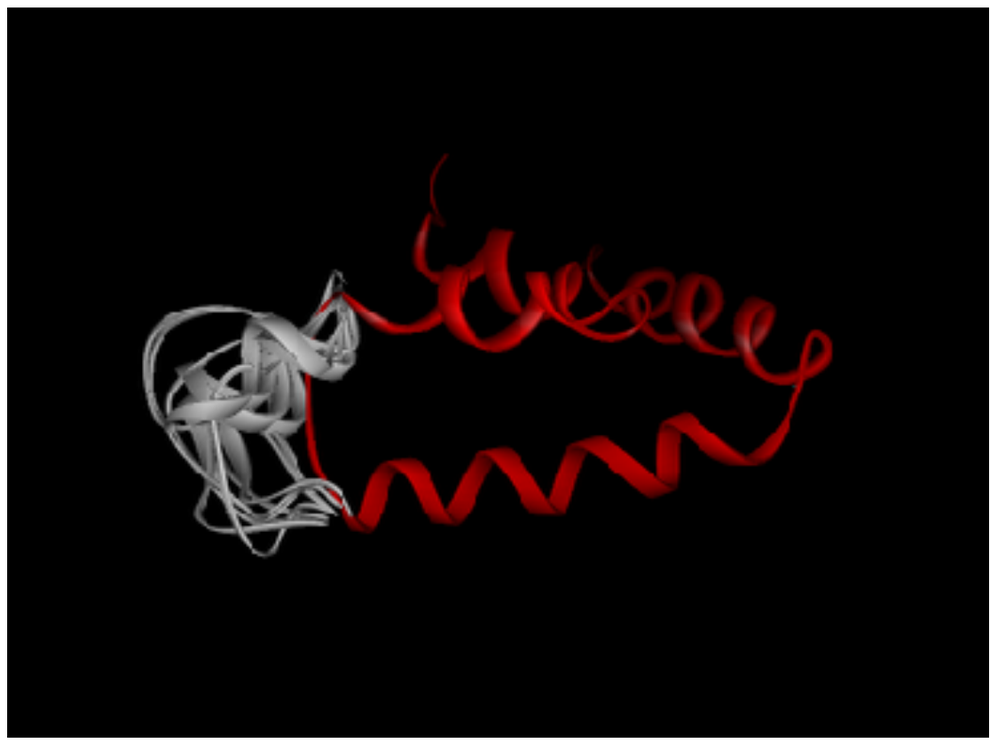
The structure(s) that fit best these restraints is (are) computed.



The advantage of this approach is that it is possible to mix the restraints obtained from different sources: experimental restraints (obtained by NMR, or other techniques), restraints obtained from the template, ...

2.3. Modeling of the loops

This is a difficult step: the loops at the surface of a X-ray experimental structure could be constrained due to crystal (impact on the loop structure of the template), some loops are flexible, ...
It is difficult to model loops larger that ± 8 residues.



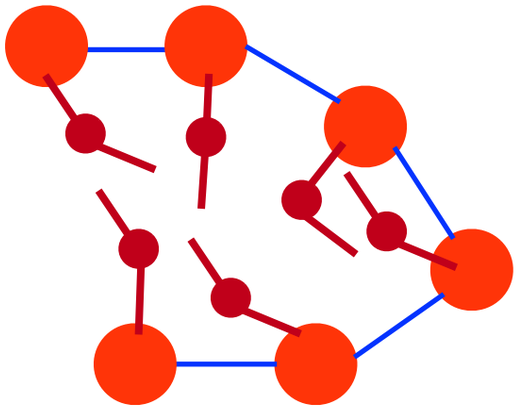
KELV---------LVLYDY QEKSPRELSQTI KKGDILTLLN STNKDWWKVE
KDLVGNDRLVLYDY QDKSIREL-----TI KTGDILTLLN STQKDWWKVH

Insertion / deletion: loops of the template and of the target present different lengths.
The main chain of these loops must be modeled.

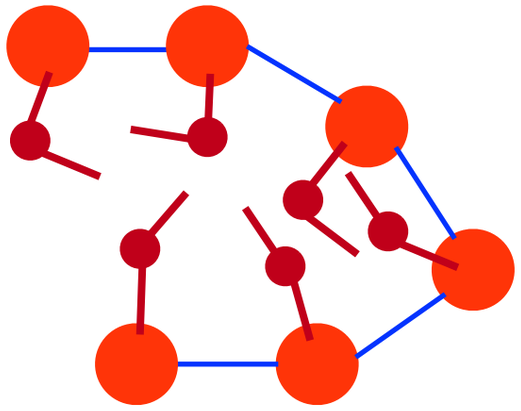
- Search in a database of known loop structures for loops having a sequence length and a distance between both extremities that are compatible with the loop to model.

- *Ab initio* techniques as molecular mechanics, molecular dynamics, Monte Carlo methods can be used (see "*Ab initio* prediction methods" chapter for the description of the methods).

2.4. Modeling of the side chains



ou



ou .....

: main chain

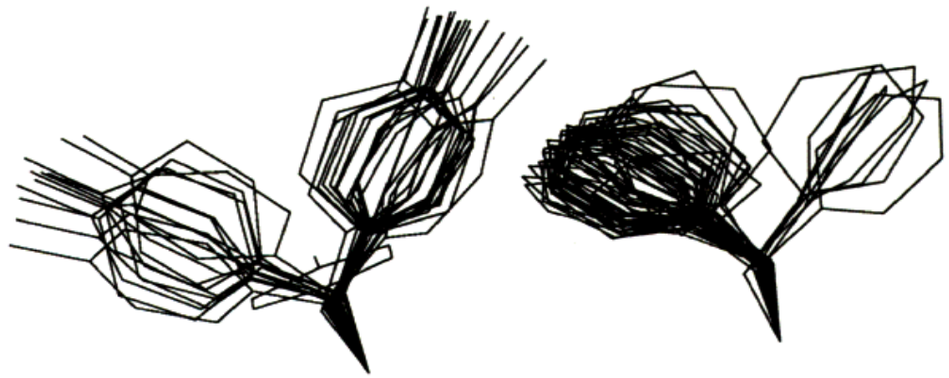
: side chain

An energy function is needed to select the best solution.

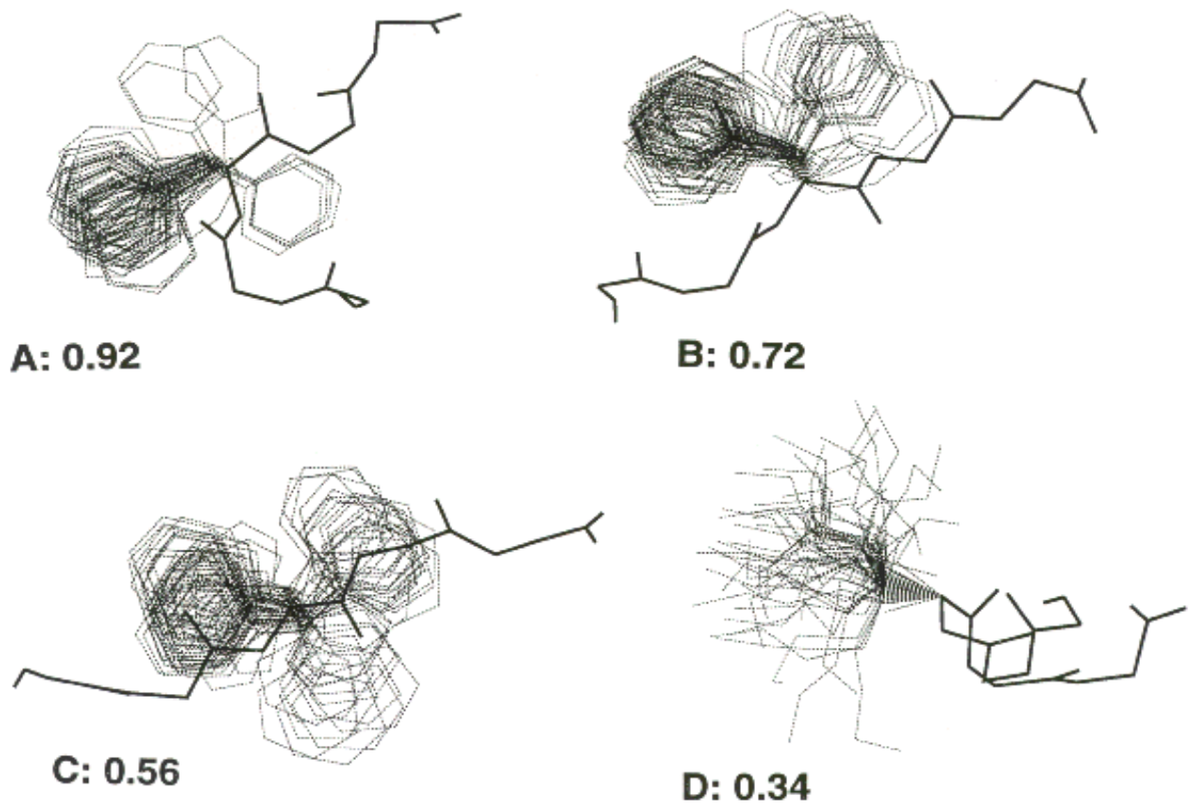
The side chains of the residues that are different in the template compared to the target are modeled.

The side chains of the identical residues in the template and in the target are modeled again.

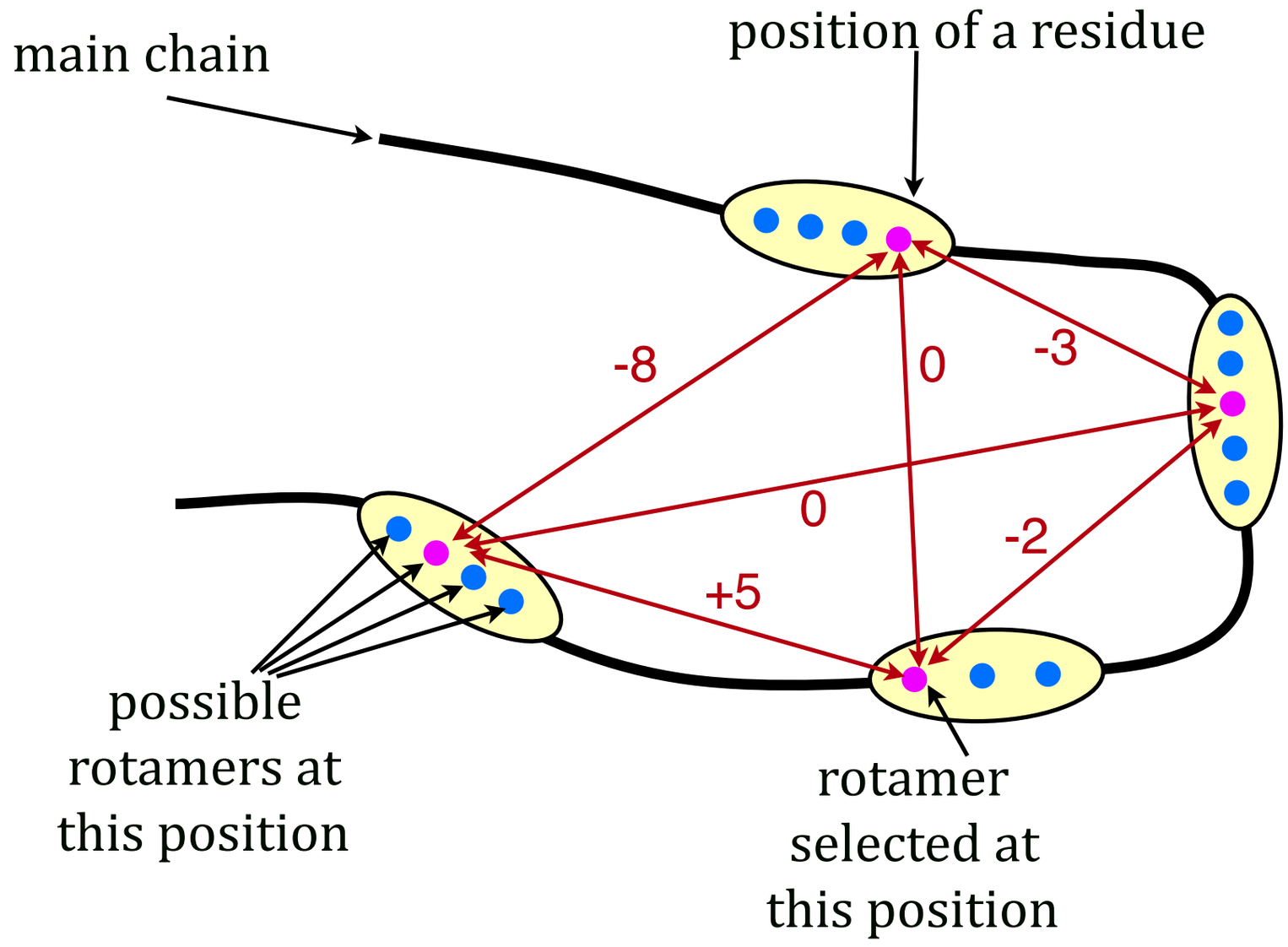
Side chain conformations observed in proteins of known structure are used: rotamer library.



The main chain limits in general the degrees of freedom of the side chain
=> side chain rotamer library depending on the main chain conformation.



The prediction of the rotamer state of each residue is a combinatorial problem:



2.5. Errors in a model obtained by comparative modeling

The amount of errors will depend on:

- the percentage of sequence identity between the template and the target.

If the sequence identity is larger than 90%, the quality of the model could be as good as an experimental X-ray structure.

If the sequence identity is between 50% and 90%, the global rmsd between the predicted structure and the "real" structure could be about 1,5 Å (and sometimes larger locally).

Below 25%, the main difficulty is the sequence alignment. Errors in the sequence alignment will have repercussion on the predicted structure.

- the amount of errors in the template structure.

Different error types:

- errors in the side chains packing: this is critical if it happens in the functional regions of the protein;

- errors in regions without template (insertion, loop) ;

- errors due to a wrong sequence alignment;

- wrong template. This can happen if the sequence identity between the template and the target is low (< 25%).

3. Fold Recognition

If no protein of known structure and that share a large sequence identity with the target is found, fold recognition can be tested.

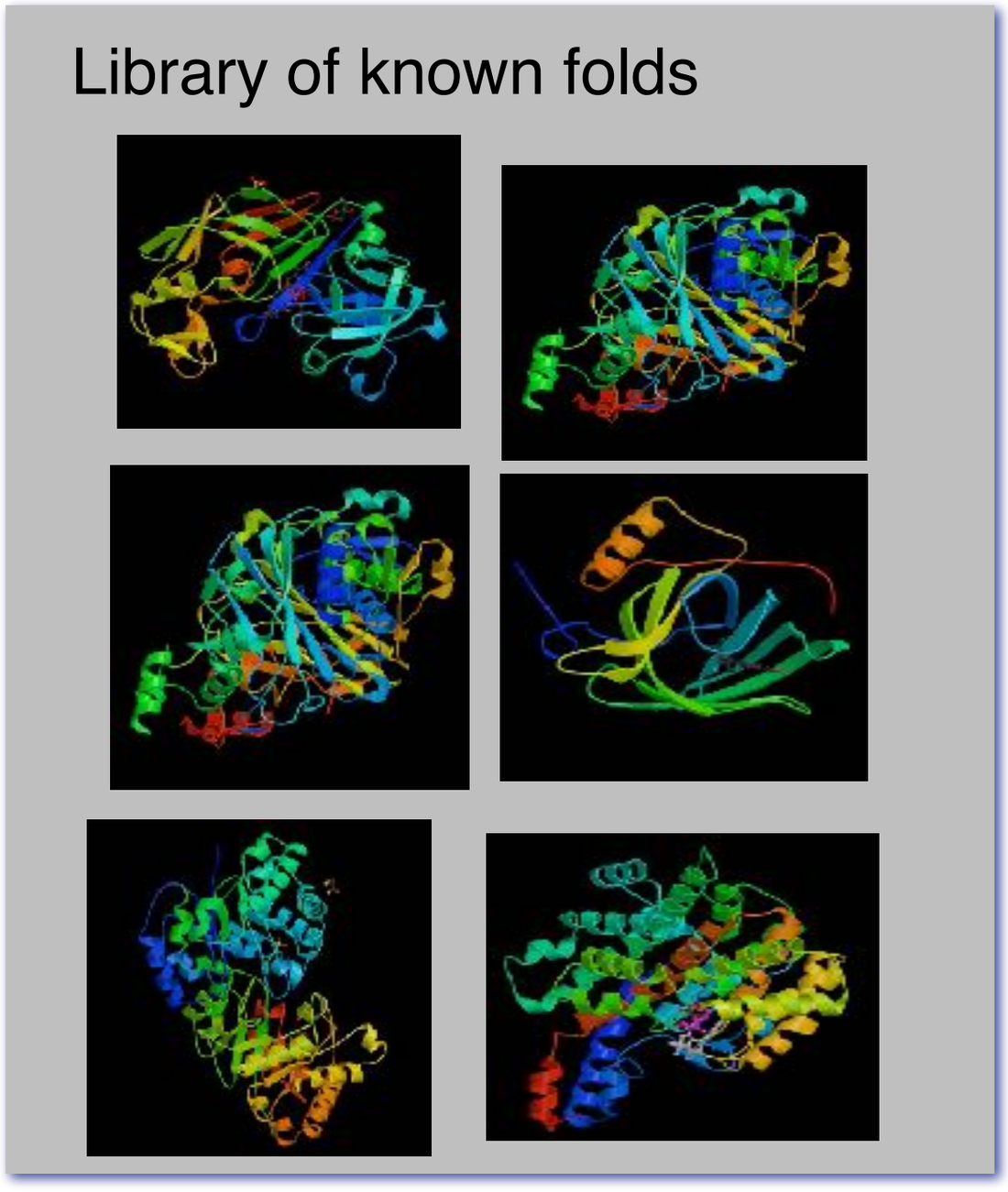
One will consider that the structure of the target could be obtained from a known fold.

target

M V G L I W A Q A T S G V I G R G G D I P

The sequence of the target will be threaded on the structures from a library of known folds.

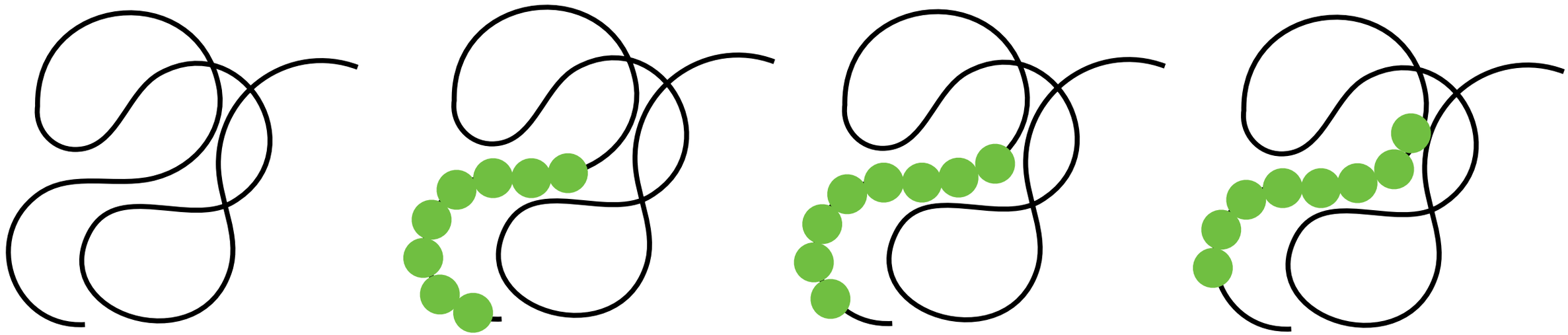
The energy of all the sequence-structure associations is computed.



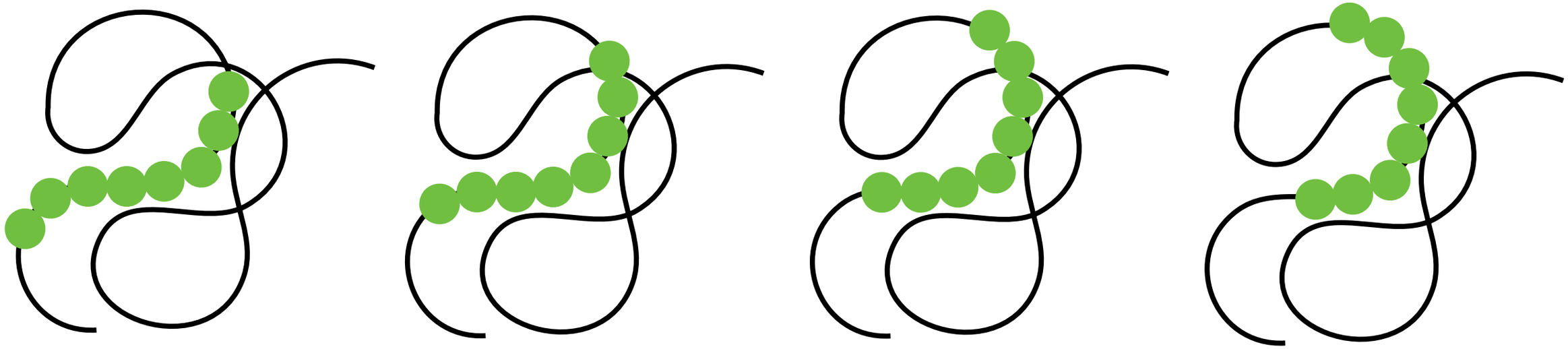
—— : structure

●●● : target sequence

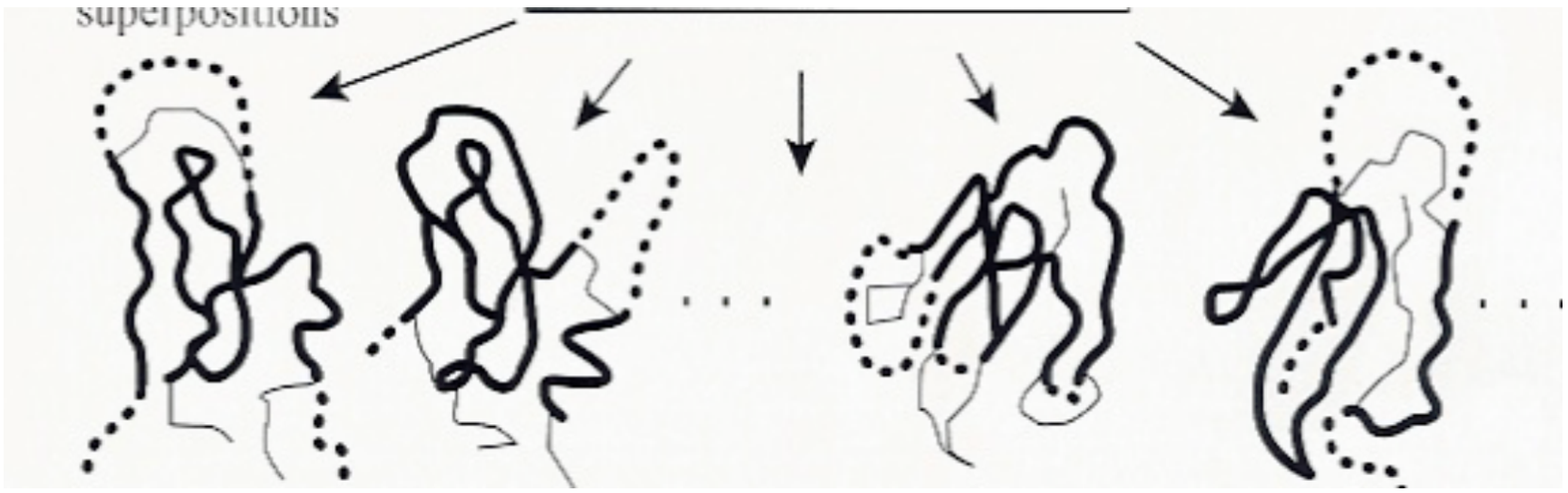
E : energy of the sequence mounted on the structure



E_1 E_2 E_3



E_4 E_5 E_6 E_7



When the sequence is mounted on the structures, gaps can be taken into account, but the evaluation of the energy is then more complicated.

The sequence/structure associations are ranked according to the energy. The scores are in general normalized:

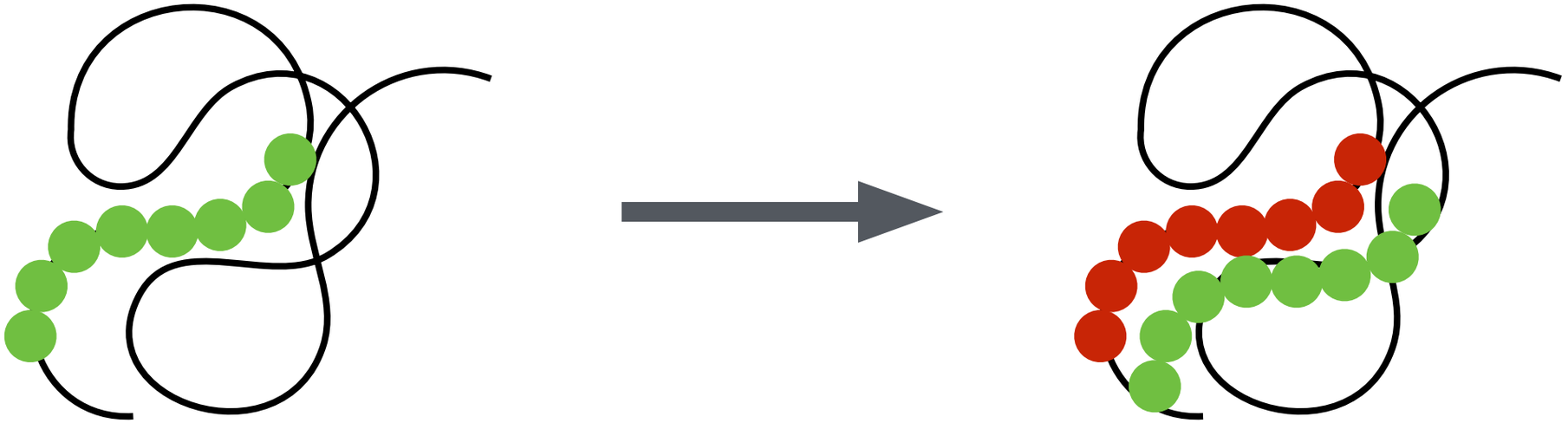
$$z_{score} = \frac{S - \langle S \rangle}{\sigma_S}$$

When the template has been selected (let be E3 the lowest energy):

: template structure

: target sequence

: template sequence

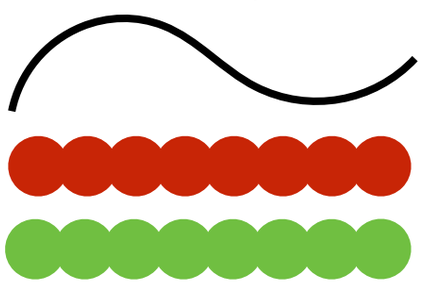


Best target sequence / template structure association

Alignment between target sequence / template sequence

template structure

Aligned template and target sequences:



Build the model with comparative modeling techniques

4. Validation

4.1. Quality of a model

There exists several methods to evaluate the quality of a model.

- on the basis of the stereochemistry: bond length, angles values, planarity of aromatic cycles, chirality, main chain torsion angles values, ... The Procheck program, for instance, do that.

- 3D profile methods or statistical potentials: the environment of a residue is compared to that found in well resolved structures obtained by X-ray crystallography.

Quality of a model and its use:

- design of a ligand that binds the active site

- rationalization of the functional differences between proteins

→ active site must be defined with an error < 1Å

- rationalization of the effects of mutations

→ the mutated region or the binding region must be defined with an error <1Å

- characterization of binding surfaces

- analysis of surface properties

→ model must be globally correct

4.2. Performances of a prediction method

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Blind evaluation of the performances of 3D protein structure prediction methods. Scientific papers describe this evaluation and the results.

Two categories:

- template-based modeling: targets with detectable evolutionary similarities to experimental structures.
- free modeling, *ab initio* methods.