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

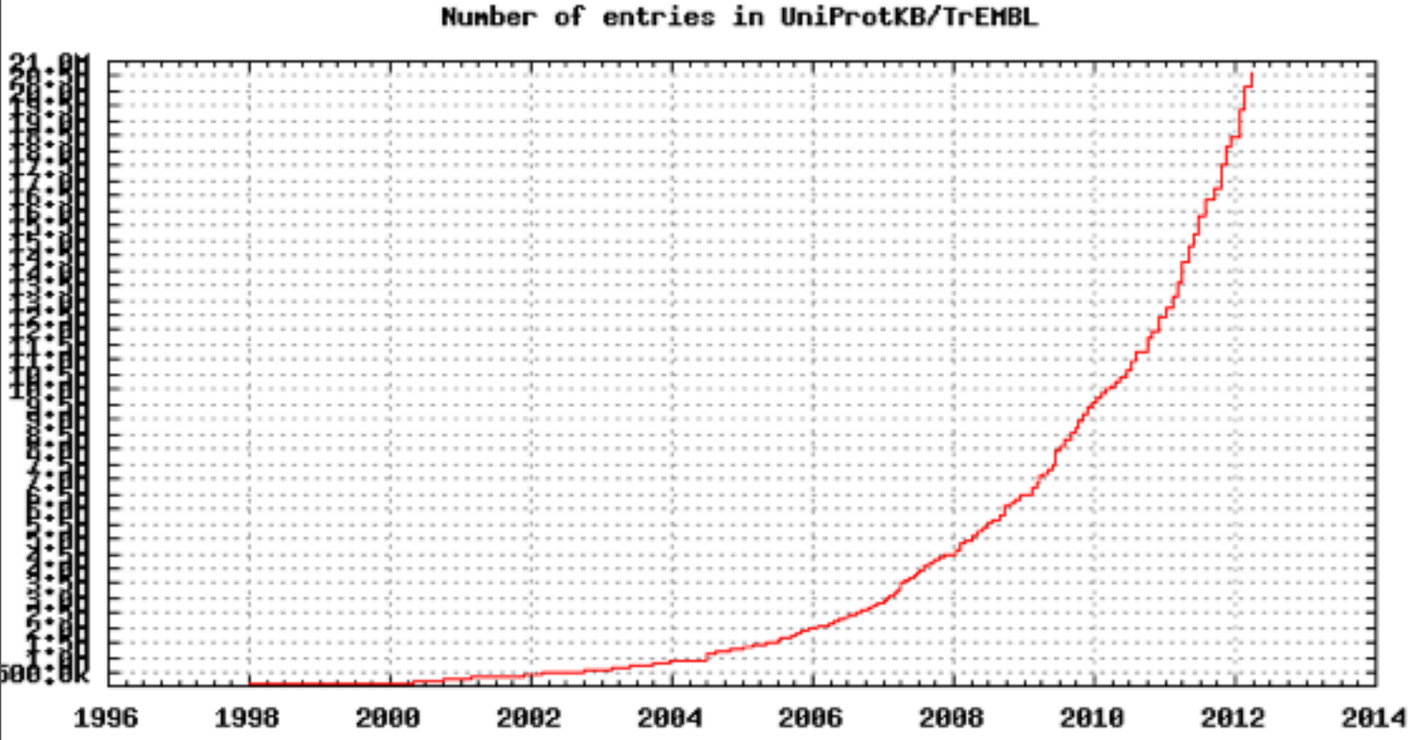
Documents at: http://uv.ulb.ac.be and

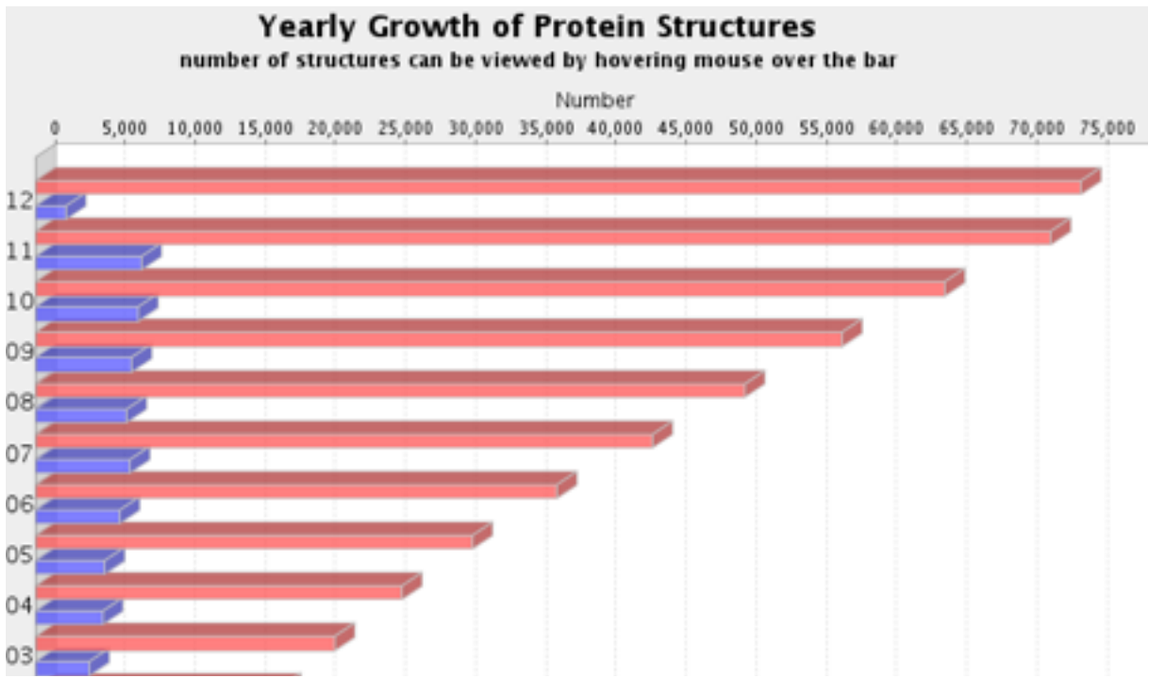
http://babylone.ulb.ac.be/~dgilis/MA1bioinfo.php

Part 5

1. Introduction

The different genome sequencing research programs generate a huge number of protein sequences, proteins whose structure and function are in general unknown.





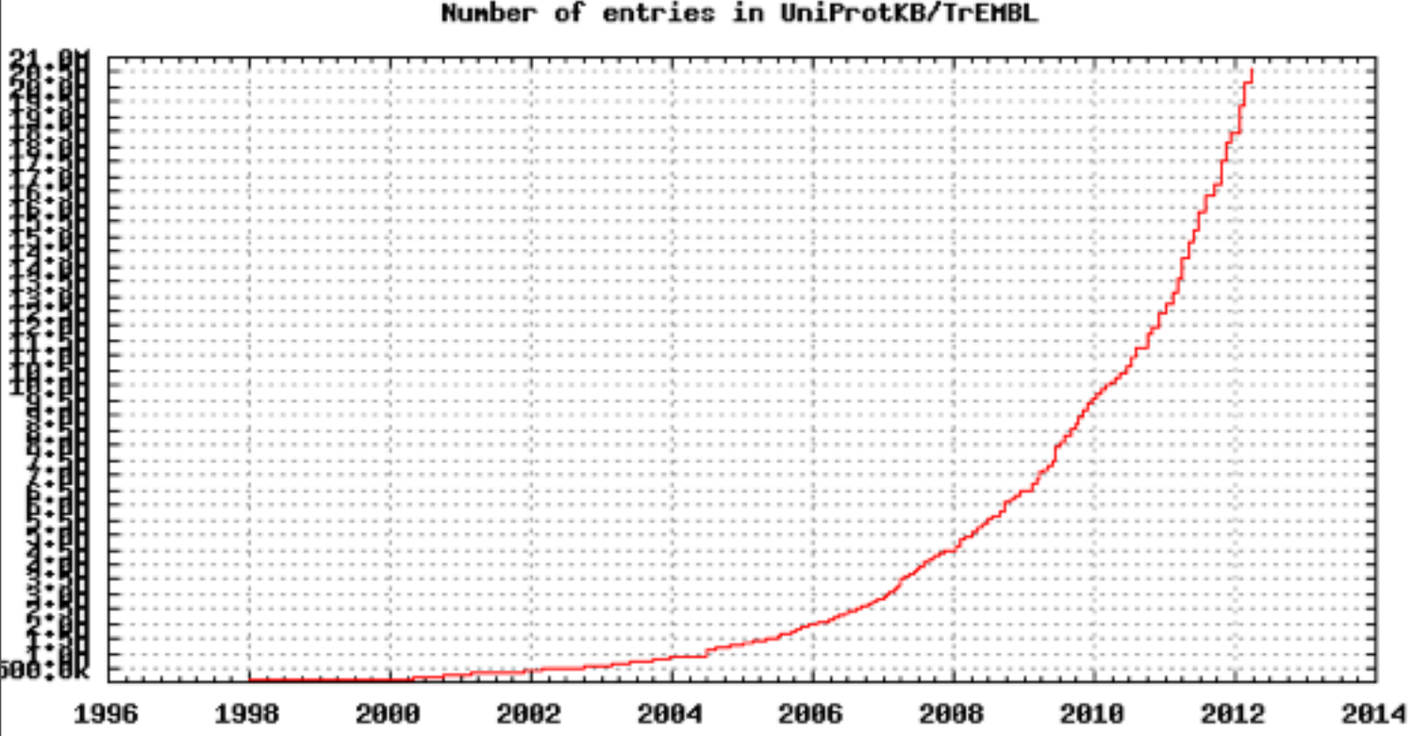
http://www.ebi.ac.uk/uniprot/TrEMBLstats/

1. Introduction

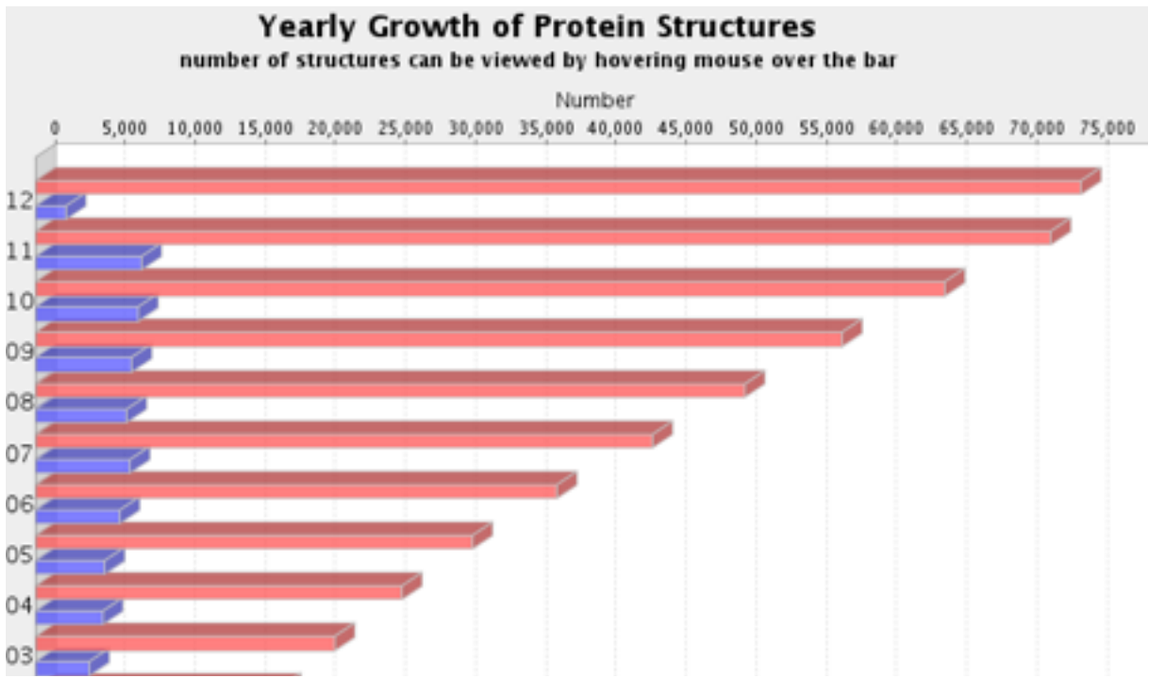
The different genome sequencing research programs generate a huge number of protein sequences, proteins whose structure and function are in general unknown.

Number of sequences
April 2012: 20.639.311
July 2013: 41.451.118
(UniProtKB/TrEMBL)

Number of protein structures
April 2012: 74.612
September 2013: 86.646
(PDB)



http://www.ebi.ac.uk/uniprot/TrEMBLstats/



http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=molType-protein&seqid=100

Experimentally resolved structures (September 2013, PDB)

Exp. Method	Proteins	Nucleic acids	Protein/NA complexes	Other	Total
X-Ray	77139	1481	4059	3	82682
NMR	8829	1044	193	7	10073
Electron microscopy	466	45	128	0	639
Hybrid	51	3	2	1	57
Other	150	4	6	13	173
Total	86635	2577	4388	24	93624

http://www.pdb.org/pdb/statistics/holdings.do

Why do we need the structure of a protein?

- Study of its properties

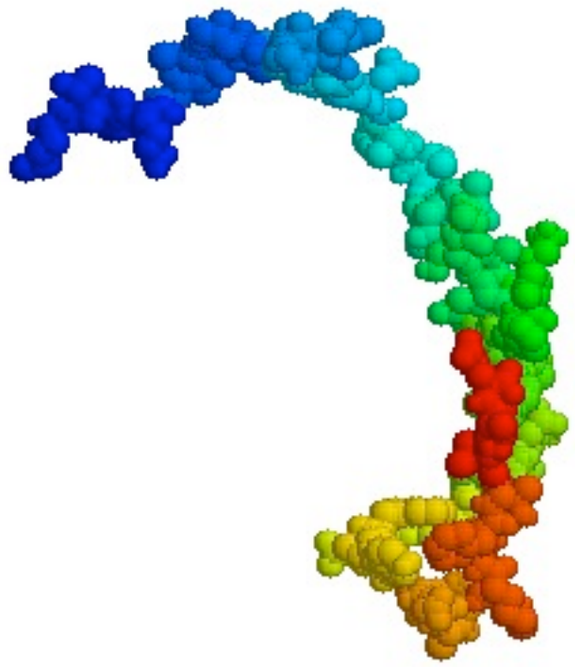
- Rational modification of its physico-chemical characteristics (solubility, thermodynamic or thermal stability, activity, specificity, ...)

- Design of ligands

- ...

Some facts about protein folding

* There are a very huge number of possible conformations: consider a small protein composed of 100 amino acids and 4 possible conformations per amino acid => $4^{100} \sim 1{,}6 \bullet 10^{60}$ conformations.



* The native 3D structure corresponds, in general, to the conformation with the lowest free energy.

* The time required to fold: 1milli second - 1 second.

How to obtain the 3D structure of a protein ?

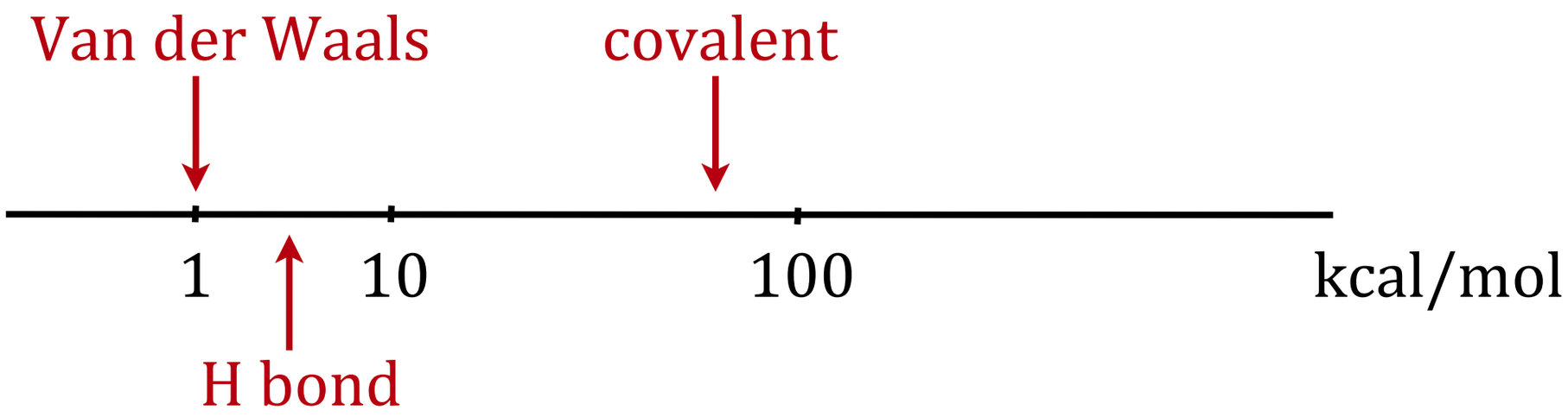
Experimental techniques (NMR, X-Ray crystallography)

In silico prediction (bioinformatics methods)

Interatomic interactions lead to protein folding and to its 3D structure. The free energy difference between the folded and unfolded states is about 20-60 kJ/mol (5-15 kcal/mol).

Types of interactions

- electrostatics;
- Van der Waals;
- hydrogen bonds;
- hydrophobic effect;
- ...

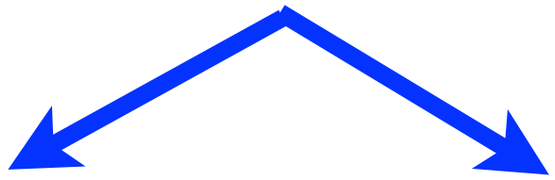


hydrophobic effect: water is a polar solvent that makes hydrogen bonds. The side-chain of non-polar (hydrophobic) amino acids can not form hydrogen bonds with water molecules. The introduction of a non-polar molecule in water causes the disruption of some water hydrogen bonds, and the building of new water hydrogen bonds leading to a "cage" of water around the non-polar molecule. This leads to a decrease of water entropy, which is unfavorable in terms of free energy. The non-polar molecules aggregate together, leading to a decrease of their accessible area to water and to a minimization of the unfavorable effects.

Energy and scoring functions are used to evaluate the compatibility between a protein sequence and a structure.

2. Energy functions

There are mainly two types of energy functions

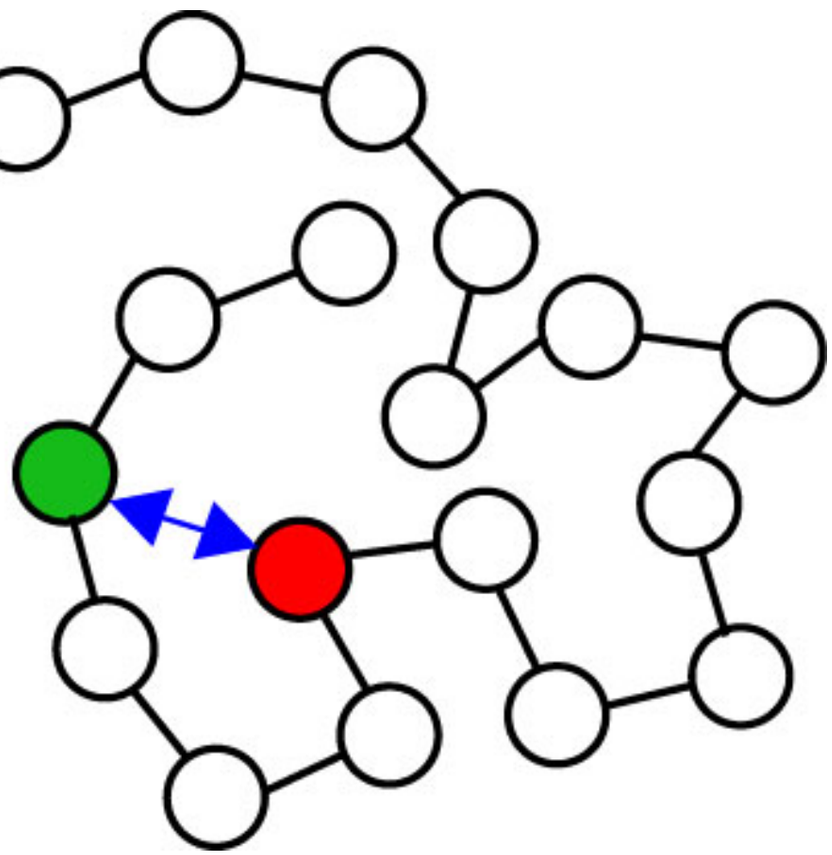


Semi-empirical potentials

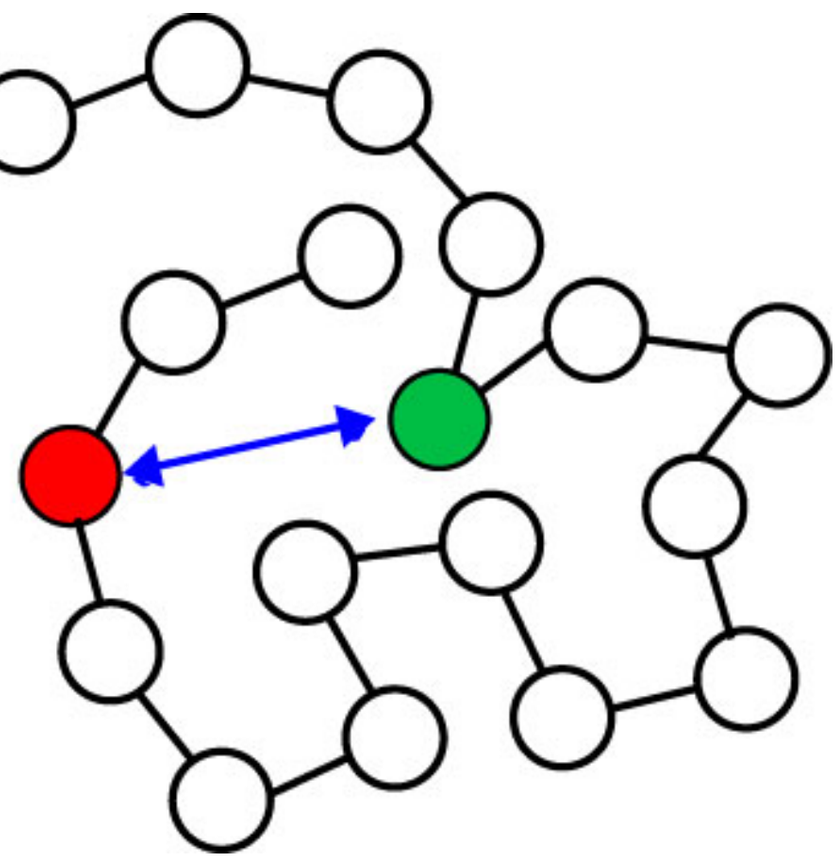
Potentials that are derived from a database of known protein sequences and structures

How are defined local and non-local interactions ? Local interactions are inyeractions between residues that are close along the sequence, whereas non-local interactions involve residues that are distant along the sequence but spatially close.

Local



Non-local



2.1. Semi-empirical potentials

These potentials correspond to an analytical expression that describes the different inter-atomic interactions.

The interactions that are included in the model are chosen.

The analytical expression that describes is chosen.

The general equation is:

$$E = E_{bonded} + E_{angle} + E_{torsion} + E_{non\text{-}bonded} + E_{other}$$

Exemple: CHARMM

$$\boldsymbol{E} = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} +$$

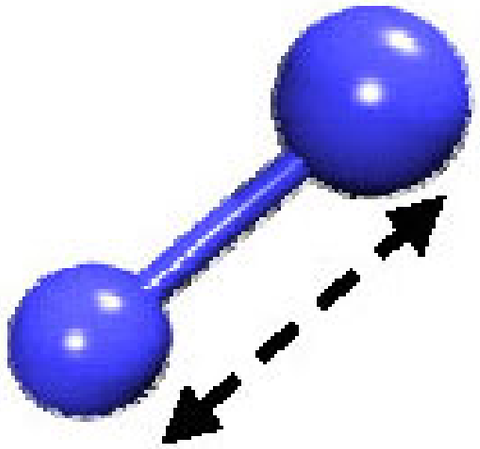
$$\underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} +$$

$$\underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

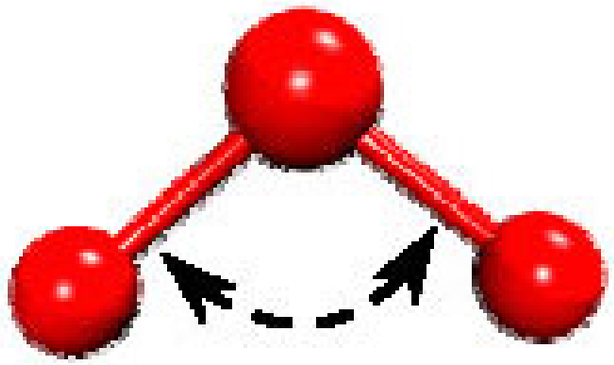
Two types of terms: between bonded and non-bonded atoms.

2.1.1. Between bonded atoms

Bond stretching



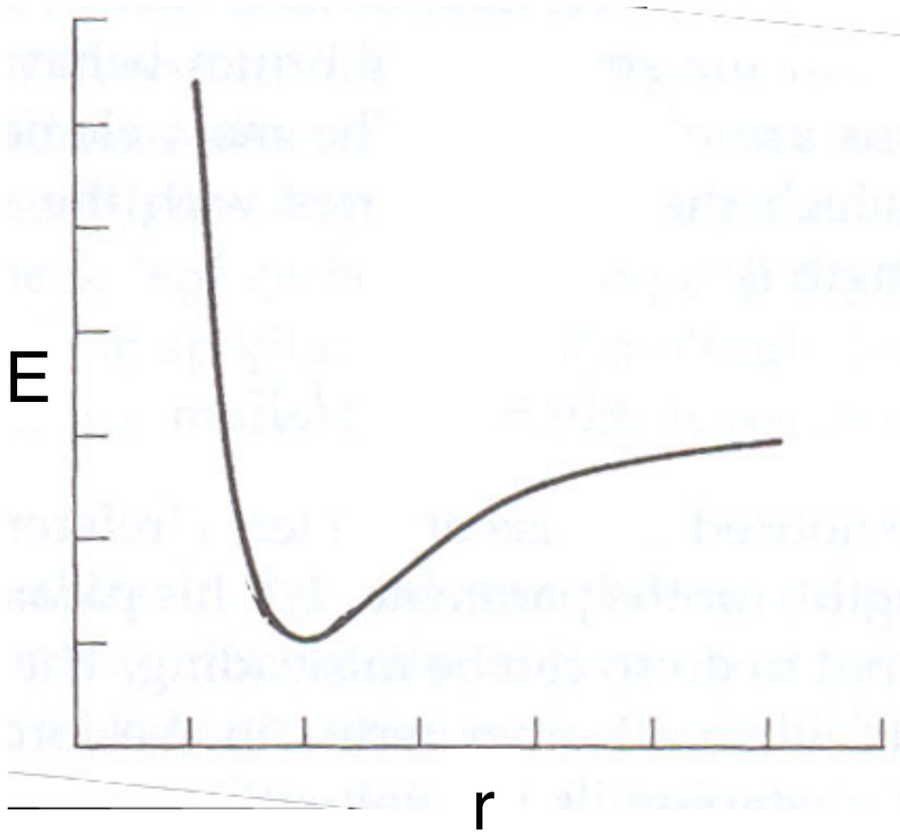
Angle opening



How to evaluate the variation in the energy when the bond is stretched, for instance ?

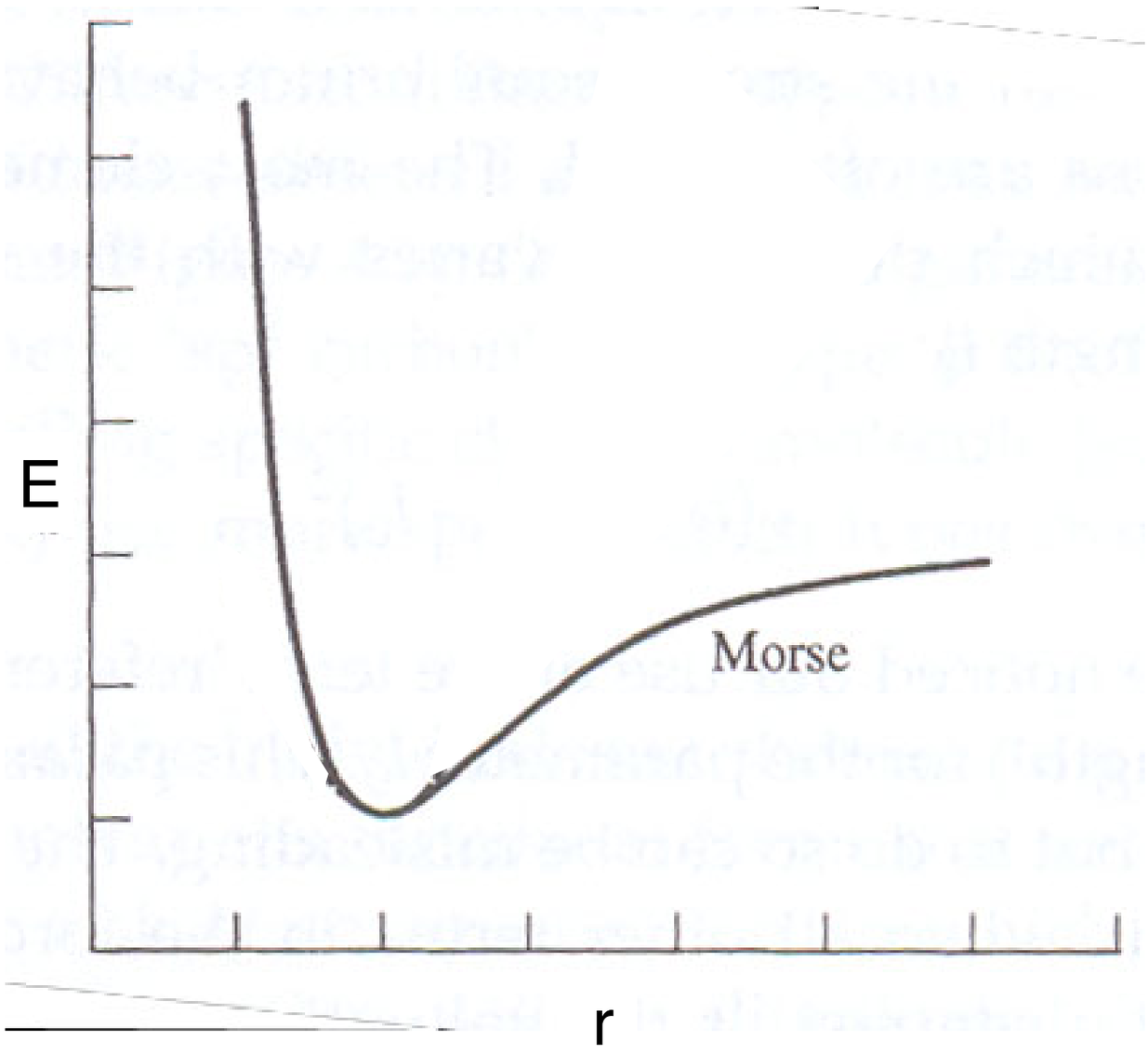
Here is the curve that represents the potential energy of the bond versus the distance.

How to model this curve ?



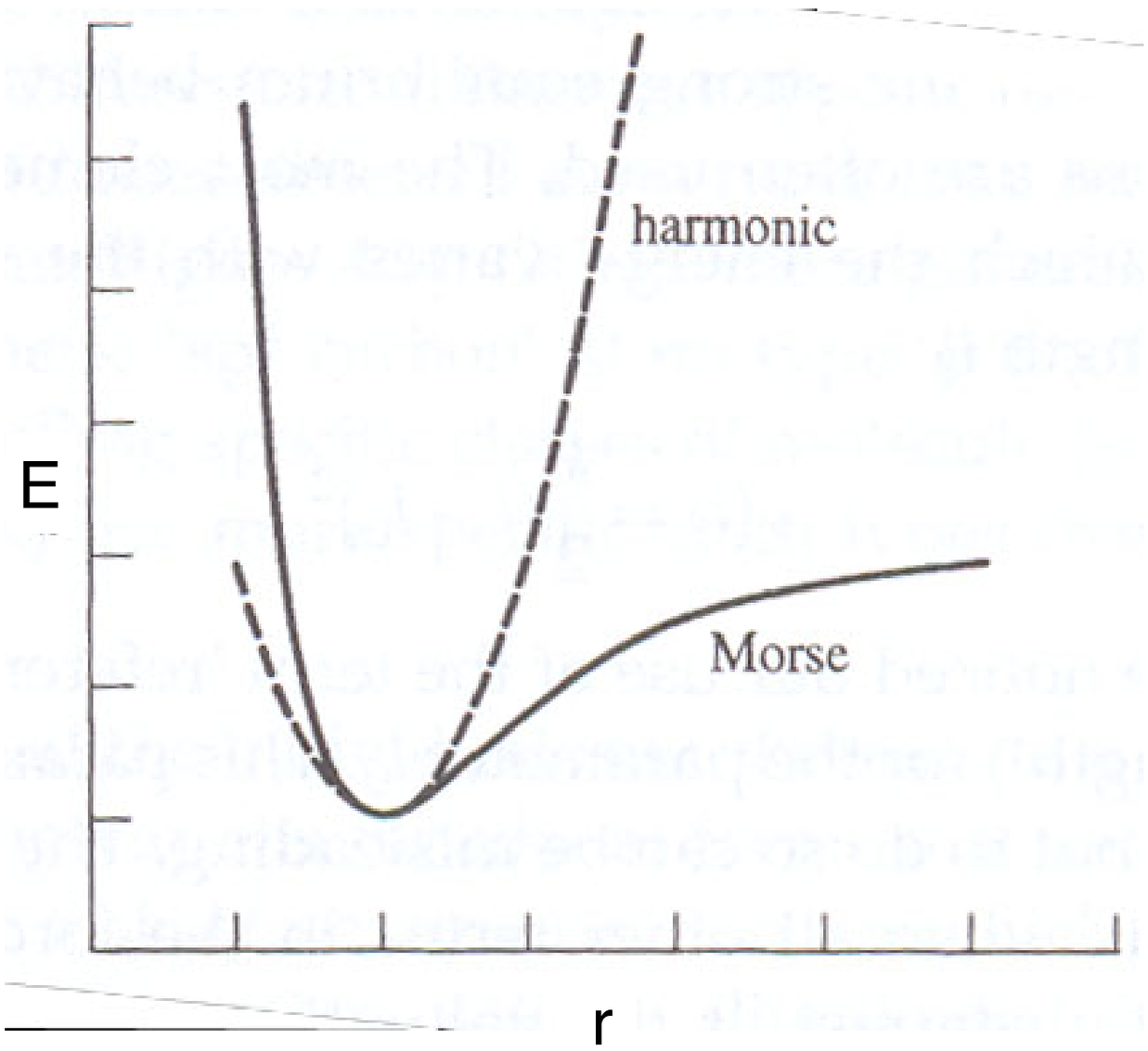
A first possibility, the Morse potential: $E=D_e \{1 - \exp [-a(r-r_0)]\}^2$

Where r is the bond length, D_e, r_0 and a are parameters.



The Morse potential contains 3 parameters per bond type. The harmonic potential is another possibility to approximate this curve:

$E = k\,(r-r_0)^2$



This approximation is suitable for small bond length changes around the equilibrium position r_0.

This analytical formulation is used in Charmm:

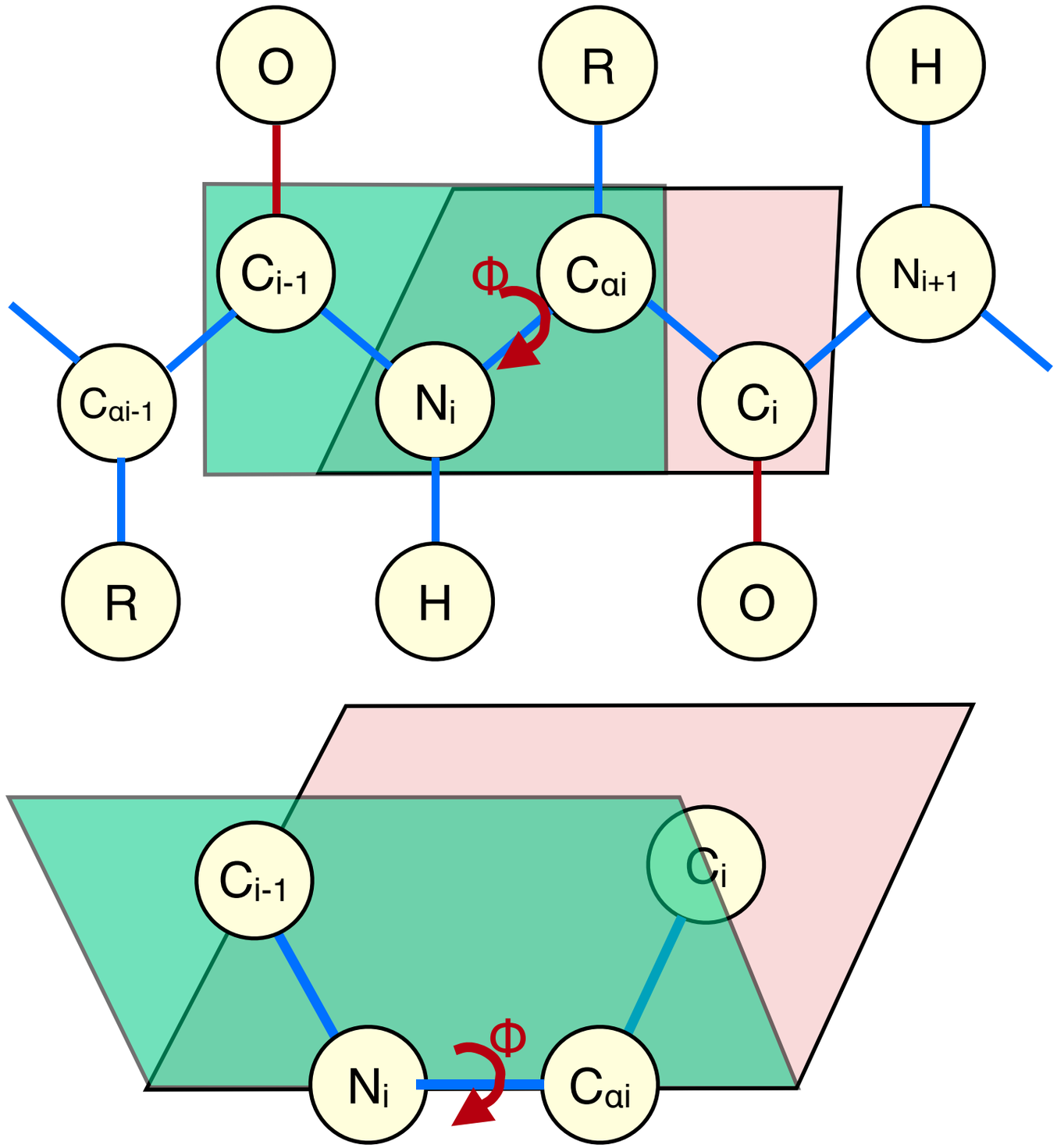
$$\underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}}$$

There are only two parameters, that depend on the atom types of the bond: k et r_0. A similar mathematical expression is used to model the opening of a bond angle:

$$\underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}}$$

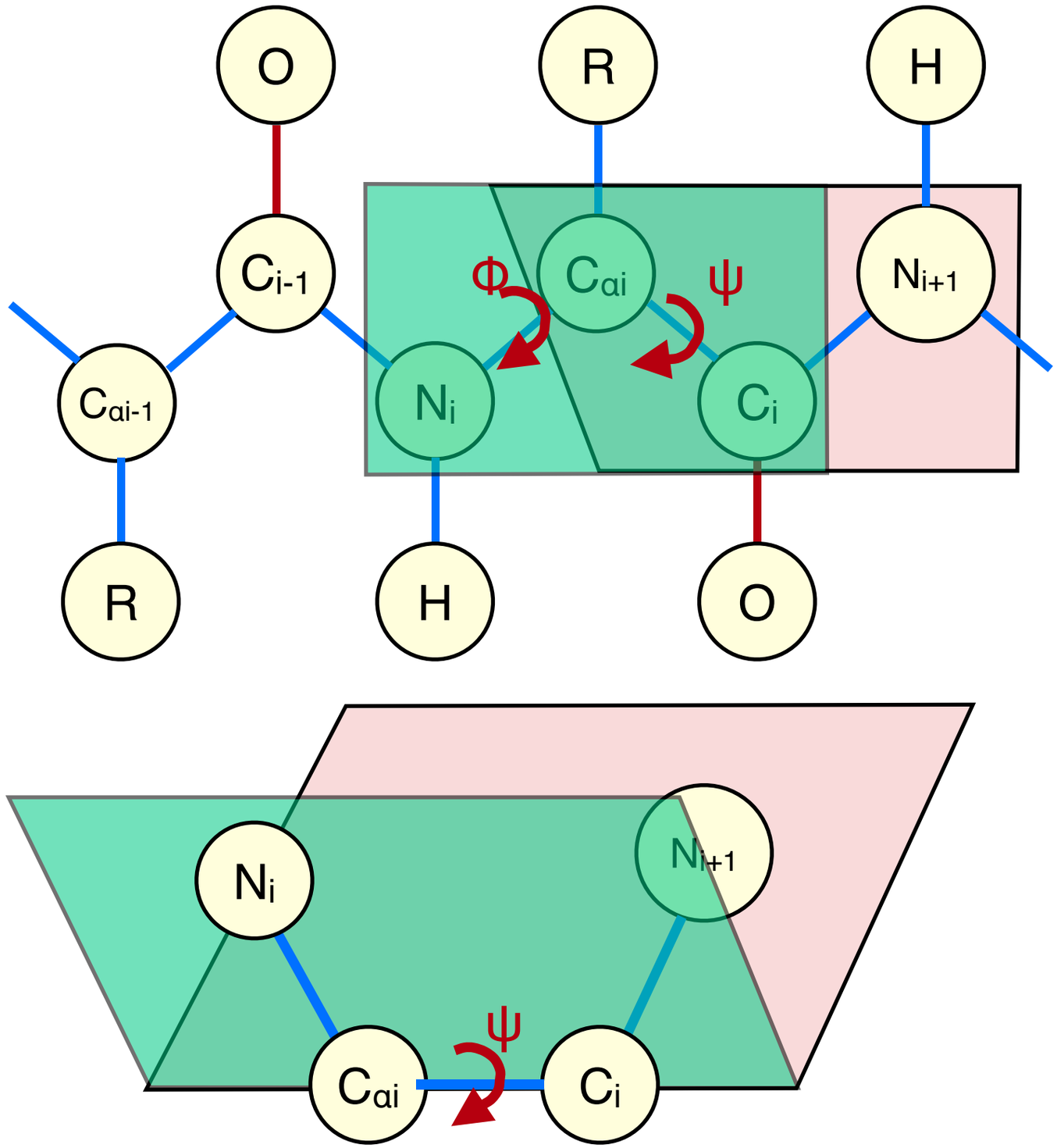
2.1.2. Between non bonded atoms

Definition of the main chain dihedral angles



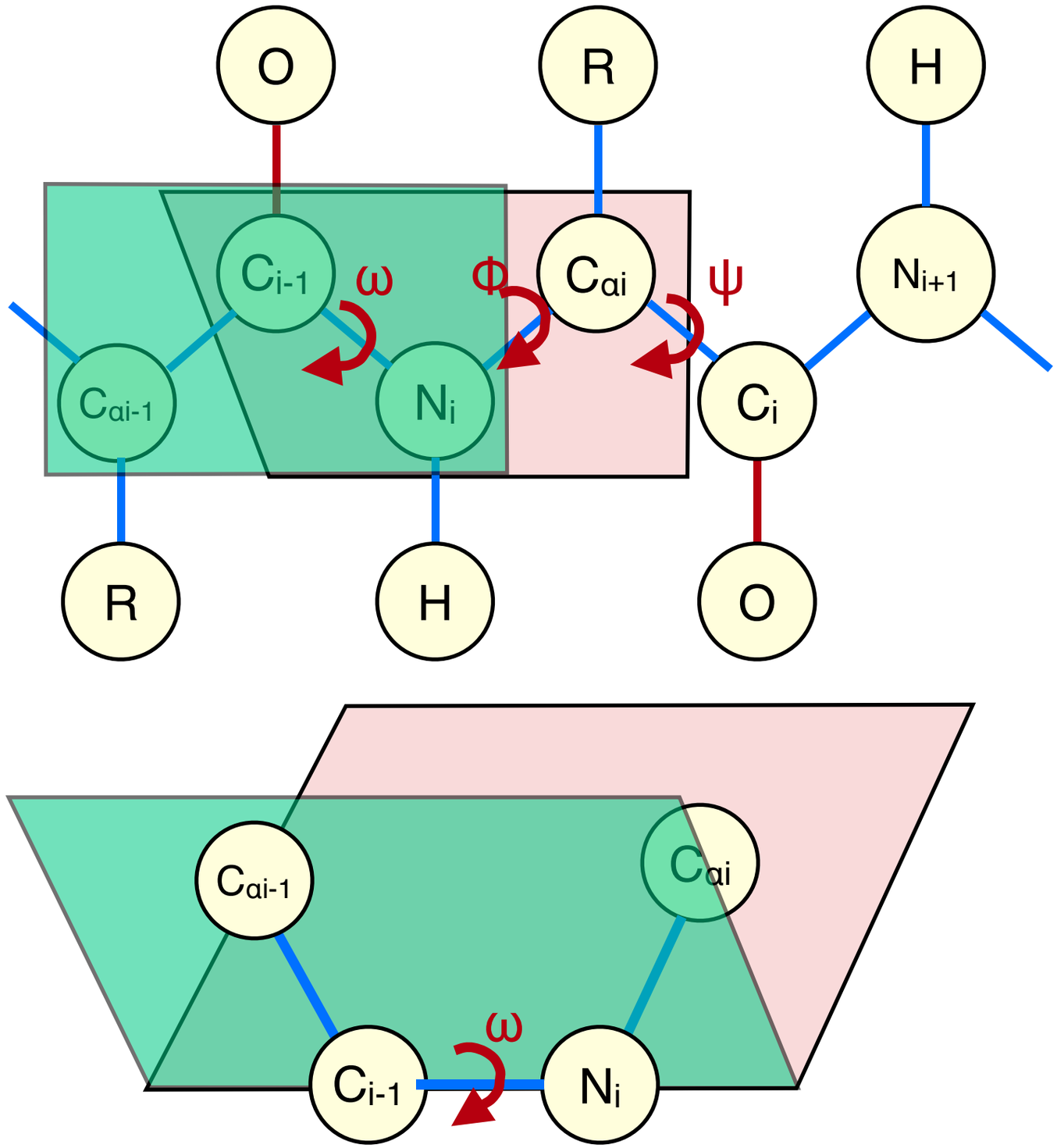
2.1.2. Between non bonded atoms

Definition of the main chain dihedral angles



2.1.2. Between non bonded atoms

Definition of the main chain dihedral angles

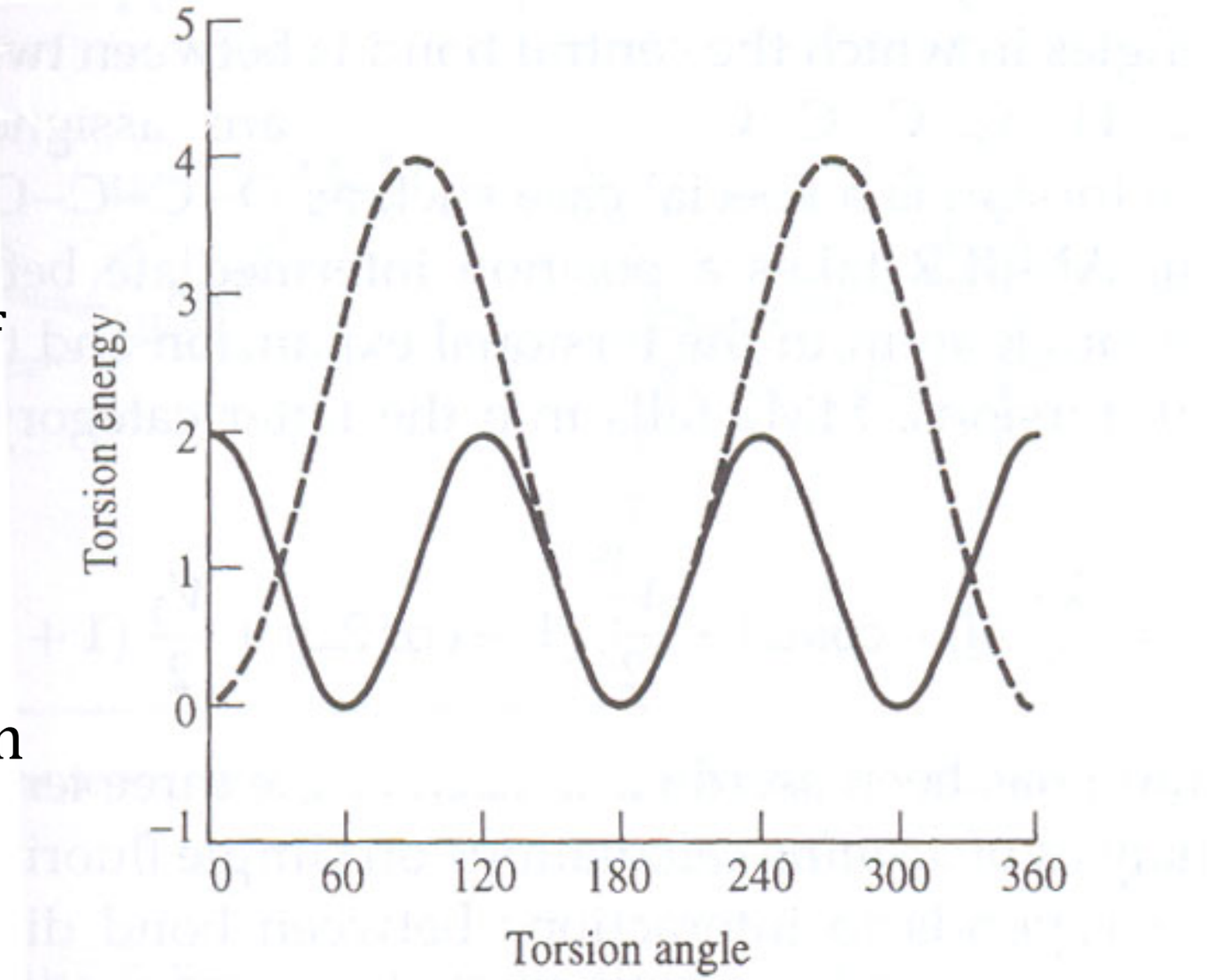


2.1.2.1. Torsion potential

$$\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]$$

k is the amplitude, *n* the number of knots, δ corresponds to the phase and Φ is the torsion angle.

This potential models the energy barriers when considering rotation around a bond.



2.1.2.2. Electrostatic interactions

$$\sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}$$

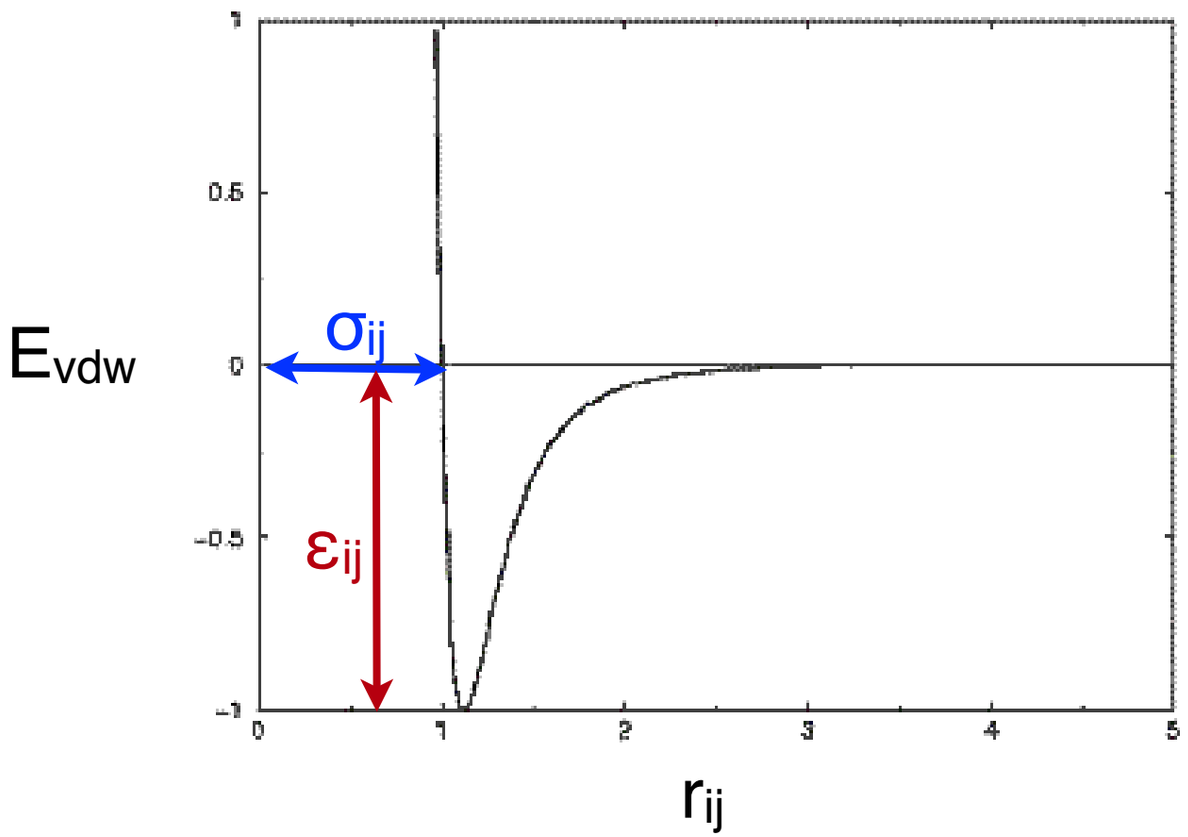
q_i and q_j are the charges, r_{ij} is the distance between the charges and ε corresponds to the dielectric constant of the medium. ε is different at the surface and in the core of the protein.

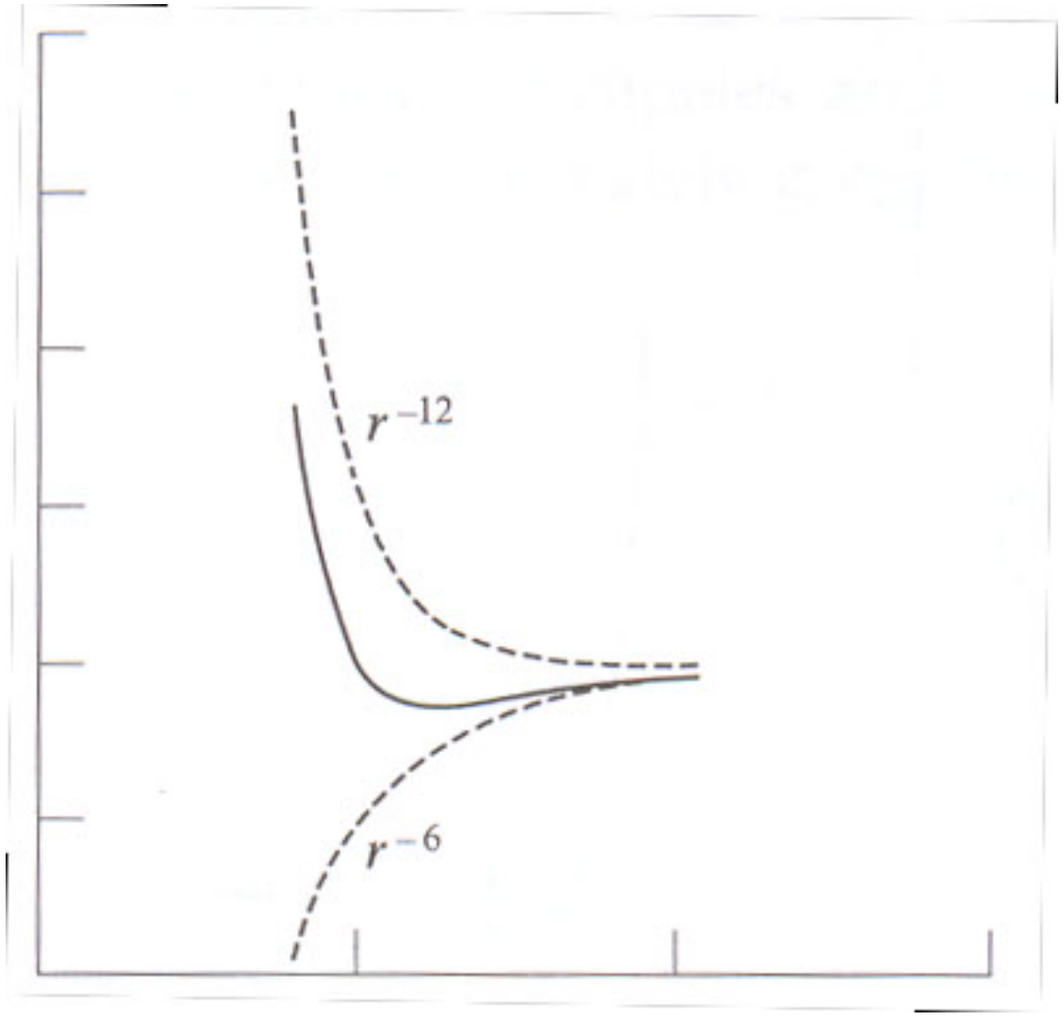
2.1.2.3. Van der Waals interactions

$$\sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

These interactions are computed with the Lennard-Jones potential.

$$\sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$





The $1/r^6$ term is an attractive term: this analytical form has a clear physical justification coming from the average of all the induced dipole-induced dipole geometries.

The $1/r^{12}$ term is the repulsive part of the interaction. This analytical formulation has no physical justification. $1/r^{10}$ or $1/r^9$ terms are found in some potentials.

$$\boldsymbol{E} = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} +$$

$$\underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} +$$

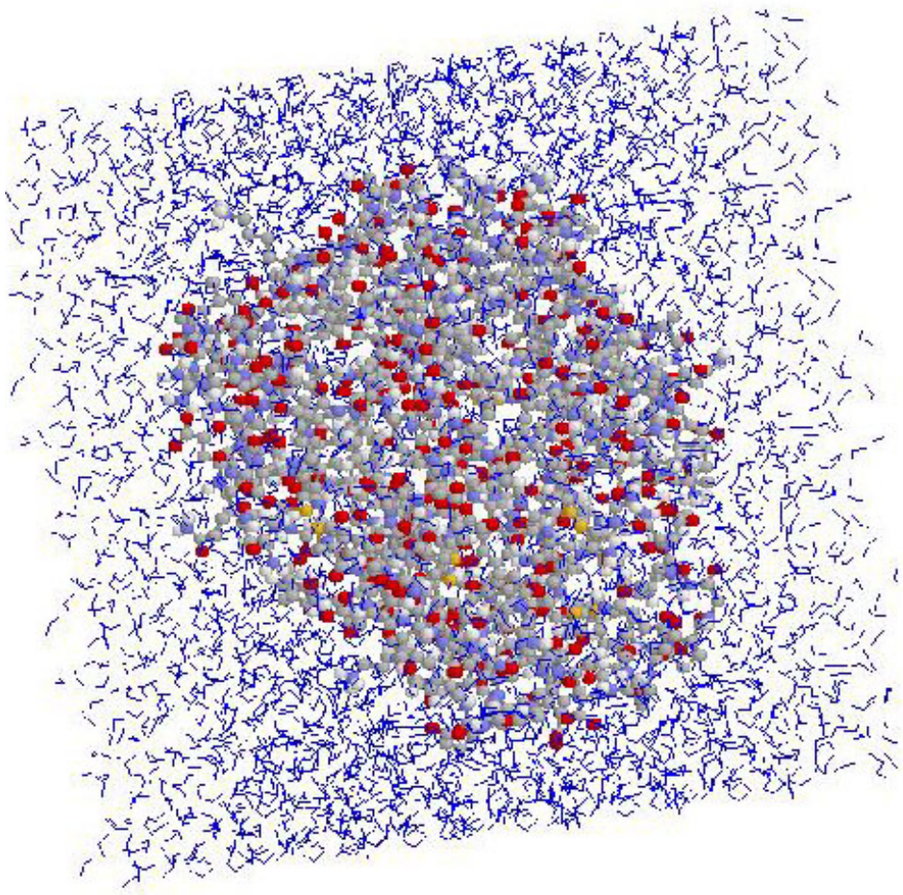
$$\underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

In this kind of potential, some terms can be added or removed. The analytical formulation of the terms can also be modified.

2.1.3. Solvent contribution

How to take into account the solvent contribution in a semi-empirical potential ?

◉ Explicit model. Water molecules are added in the system and the energetic contribution of these molecules is computed. Given that a large number of atoms is added, the computational cost increases.



◉ Implicit model. The solvent is considered as a perturbation compared to the gaz phase. The effect of the solvent is described by an analytical formulation.

2.1.4. Parametrization of semi-empirical potentials

The identification of the parameters of the potential, that leads to the most reliable calculation of the energy is not a trivial task.

$$\boldsymbol{E} = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} +$$

$$\underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} +$$

$$\underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

◉ Empirical estimation: the parameters that lead to the best correspondence between measured and computed energies.

◉ Computation of the parameters by quantum approaches.

Be careful about the transferability of the parameters: some potentials have been optimized to compute the energy of different types of molecules (nucleic acids, sugars, proteins, ...)

Some examples of semi-empirical potentials: AMBER, CHARMM, Gromos, OPLS, ...

2.2. Effective potentials

The electrostatic and Van der Waals potentials are computed between pairs of atoms:

$$\sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}$$

$$\sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

The total energy is obtained by summing all the pair contributions. BUT: pair interactions are under the influence of the other atoms.

Is it necessary to take into account all the triplets, quadruplets, ... ??

Number of pairs: N (N-1)/2
Number of triplets: N (N-1) (N-2)/6
...

=> Effective potential: the effective potential takes into account the presence of the other atoms through the parameters. This kind of potential does not correspond to the «real» interaction energy between two isolated atoms, but is parametrized to take into account the effects of the other atoms in the pair energy.

2.3. Database-derived potentials

Semi-empirical potentials

Potentials derived from a database of proteins whose sequences and structures are known

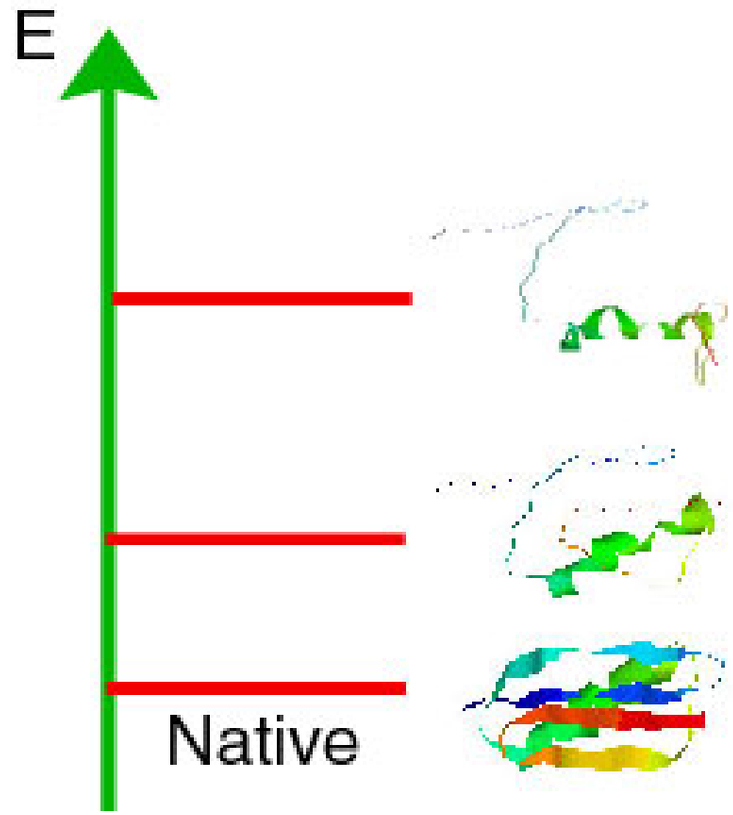
2 approaches

Analytical formulation that describes the interactions + optimization of the parameters of this function by using the database of known protein structures.

Derivation of the frequency of association between sequence and structure elements ; in the statistical mechanics framework, these frequencies are converted into free energies.

In statistical mechanics, the probability distribution of the conformations of a molecule obeys the Boltzmann law:

$$P(C_i) = \frac{exp\ (-E(C_i) / kT)}{\sum\ exp(-E(C_j) / kT)}$$



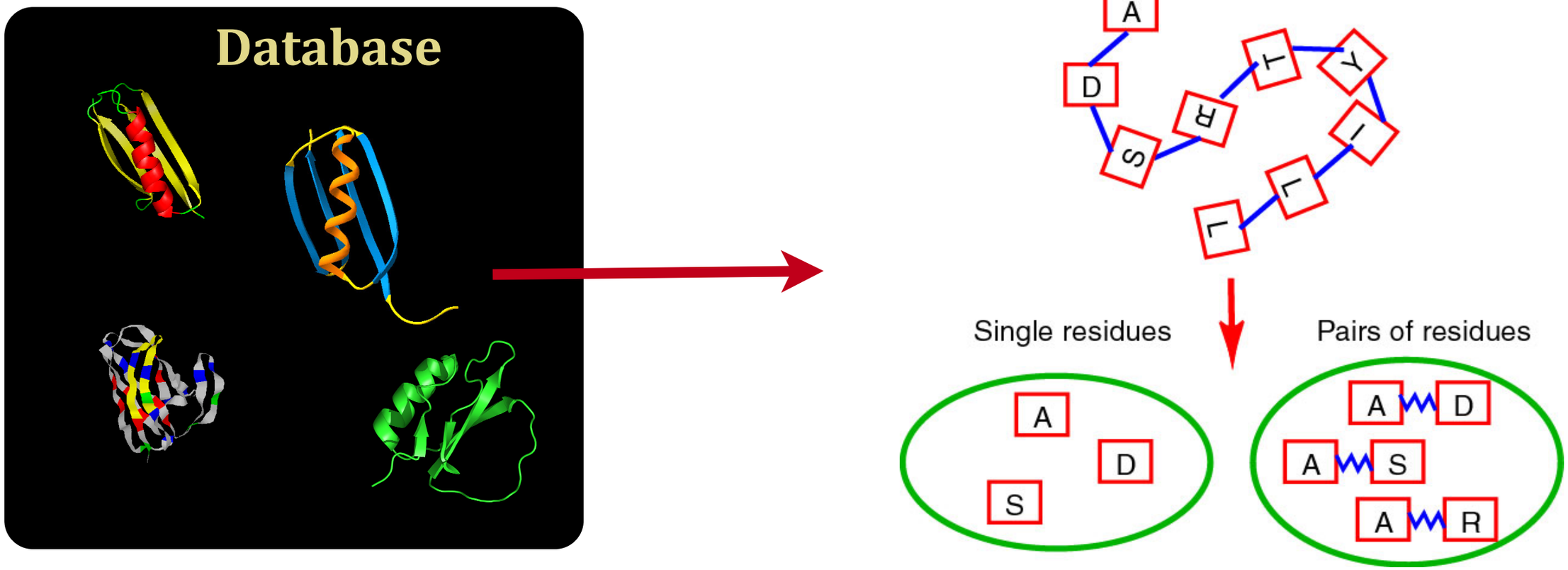
where $P(C_i)$ is the probability of conformation C_i, $E(C_i)$ is the energy of conformation C_i, k is the Boltzmann constant and T is the temperature.

Only the native conformation of a protein is known. The conformations C_j are not characterized.

First step: to build a database of proteins whose sequence and structure are known:

- Low sequence similarity between these protein (<25%) to avoid bias;
- Structures well resolved and refined (resolution < 2Å)
- Large database to have reliable statistics.

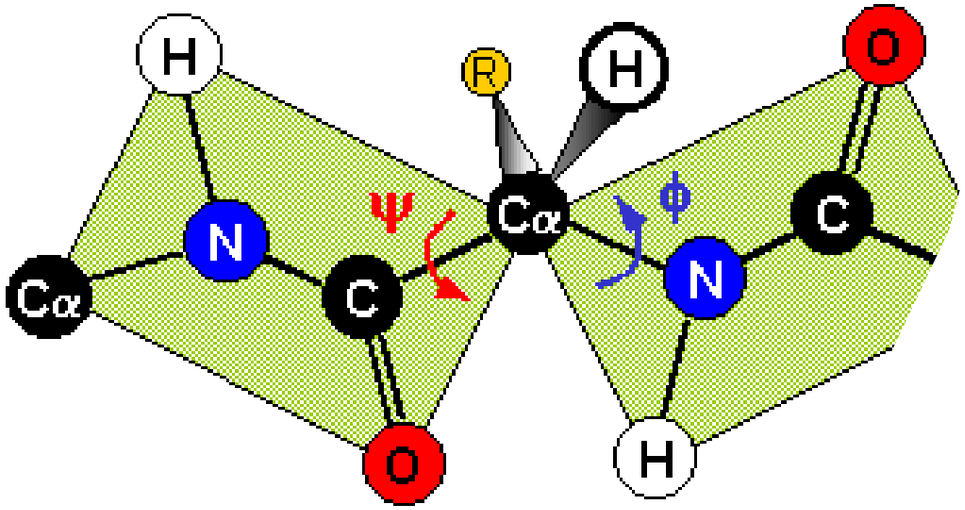
Second step: to divide the sequences and the structures into sequence and structure elements.

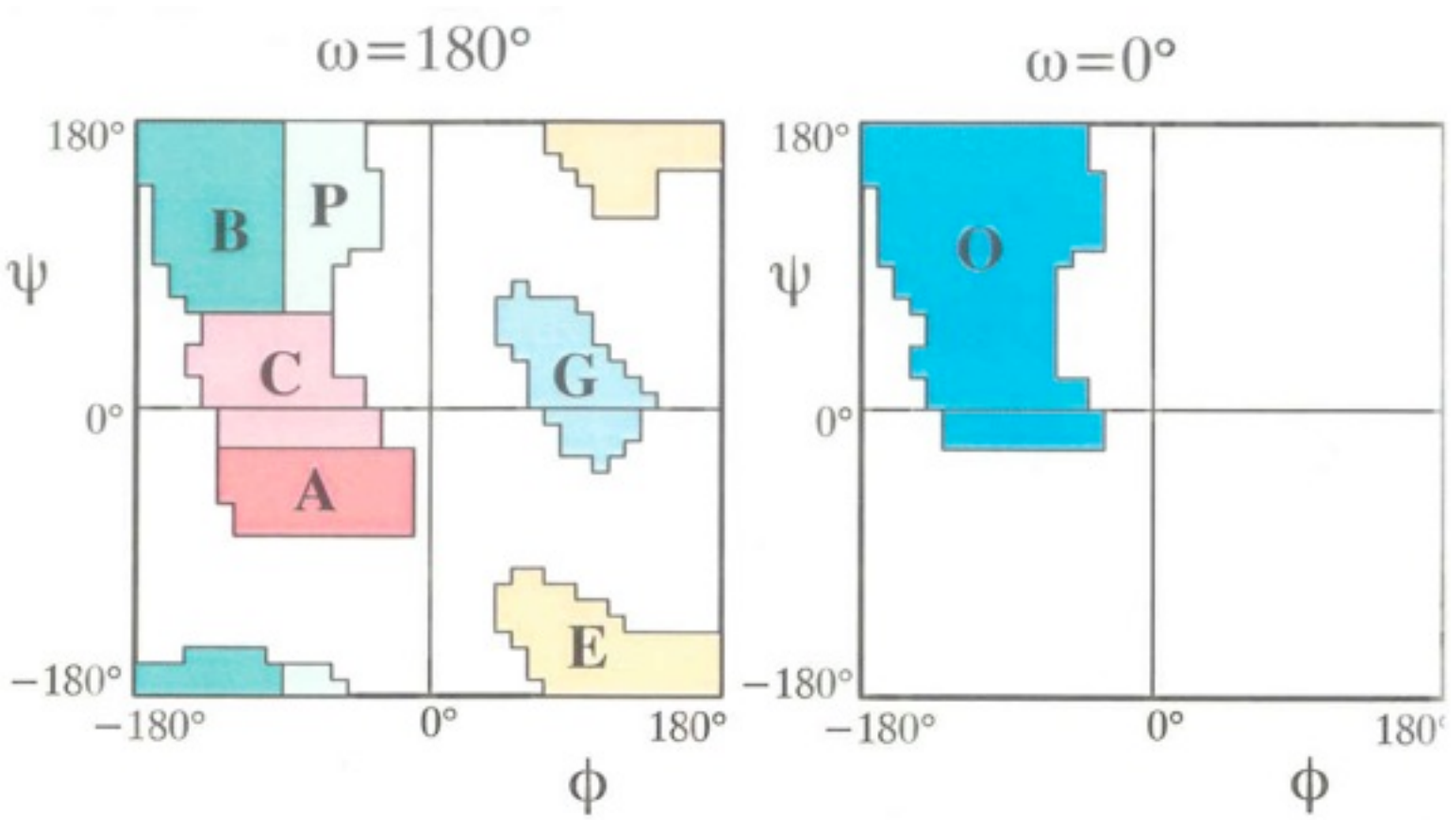


Sequence elements

Structure elements: parameters that describe protein structures.

- Distance between amino acids: this distance can be computed between atoms (atomic description) or between Cα's or side-chain geometric centres, ...
- Main-chain torsion angles





It is possible to group these angles into domains (the figure shows an example of 7 torsion angle domains:
A=α helix, C=3_{10} helix, B= β type extended conformation,...)

2.3.1. Distance potentials

In statistical mechanics, the pair potential of mean force $w^{(2)}$ between two particles at positions $\mathbf{r_1}$ et $\mathbf{r_2}$ is:

$$\exp[-w^{(2)}(\mathbf{r_1},\mathbf{r_2})/kT] = \frac{P^{(2)}(\mathbf{r_1},\mathbf{r_2})}{P^{(1)}(\mathbf{r_1})\ P^{(1)}(\mathbf{r_2})}$$

where $P^{(1)}(\mathbf{r_1})$ is the probability to have a particle at position $\mathbf{r_1}$ and $P^{(2)}(\mathbf{r_1},\mathbf{r_2})$ is the probability to have a particle at position $\mathbf{r_1}$ and another at position $\mathbf{r_2}$.

If particles of different types coexist in the same system, the pair potential of mean force $W^{(2)}$ between two particles of types s_1 et s_2 at positions $\mathbf{r_1}$ et $\mathbf{r_2}$ is:

$$\exp[-W^{(2)}(\mathbf{r_1},\mathbf{r_2};\ s_1,s_2)/kT] = \frac{P^{(2)}(\mathbf{r_1},\mathbf{r_2}|s_1,s_2)}{P^{(1)}(\mathbf{r_1}|s_1)\ P^{(1)}(\mathbf{r_2}|s_2)}$$

The difference $\Delta W^{(2)}$ corresponds to the potential of mean force of a system thant contains several particles, compared to a system of reference that contains only one type of particles:

$$\Delta W^{(2)}(\mathbf{r_1},\mathbf{r_2};\, s_1,s_2) = W^{(2)}(\mathbf{r_1},\mathbf{r_2};\, s_1,s_2) - w^{(2)}(\mathbf{r_1},\mathbf{r_2})$$

The potential of the whole system can be approximated as the sum over all pair interactions:

$$\Delta W^{(n)}(\mathbf{r_1},\ldots,\mathbf{r_n};\, s_1,\ldots,s_n) = \sum_{i,j=1;i<j}^{n} \Delta W^{(2)}(\mathbf{r_i},\mathbf{r_j};\, s_i,s_j)$$

This is equivalent to assume that the probability to have n particles at positions $\mathbf{r_1},\ldots,\mathbf{r_n}$ corresponds to the product of the probability of pairs.

When this formalism is applied to proteins: s_1 et s_2 are the amino acid types, $\mathbf{r_{12}}$ is the distance between them, and water molecules are implicitly taken into account (they are included in the statistical average).

The potential of the reference state, $w^{(2)}$, represents a globular state where amino acids are identical (no distinction between the different types of amino acids), that could correspond to a denatured state.

The probabilities P are computed from frequencies of observation, F, of amino acid pairs separated by a spatial distance comprised between r_{12} and $r_{12}+\Delta r_{12}$ in a database of known structures.

The following equation is obtained:

$$\Delta W^{(2)}(r_{12};\, s_1,s_2) \cong -kT \ln \frac{F(r_{12}|s_1,s_2)}{F(r_{12})}$$

W is a free energy that contains entropic contributions due to the statistical averages, to the implicit presence of water molecules and to the discretization of the conformationnal space.

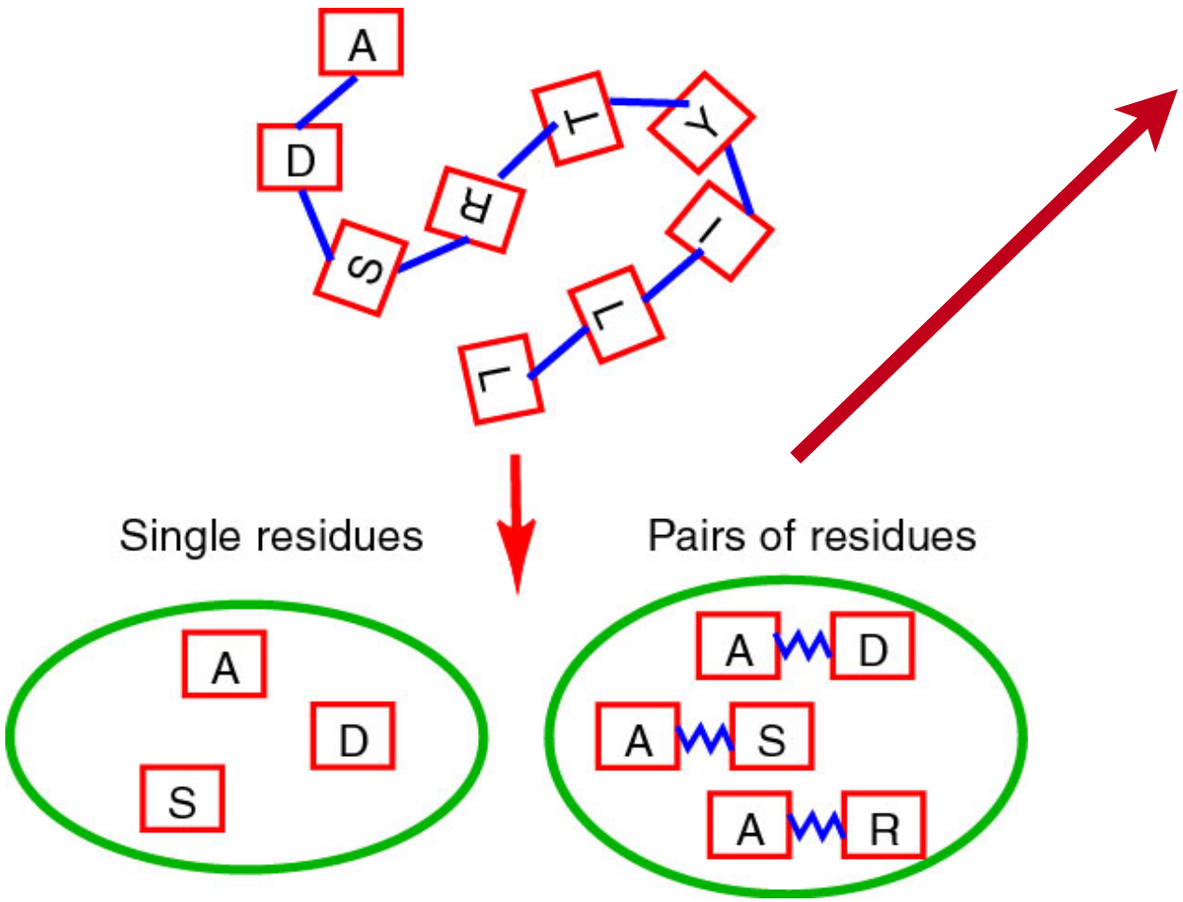
Practically:

- The distance are computed between C^{α}, C^{β} ou side chain average centroids (C^{μ}). They can also be computed between atoms.

- The distances are divided in domains (0,2Å wide for instance), or can correspond to contact / non-contact, ...

- It is possible to smooth the potentials, by comining the frequencies computed on the neighbouring domains.

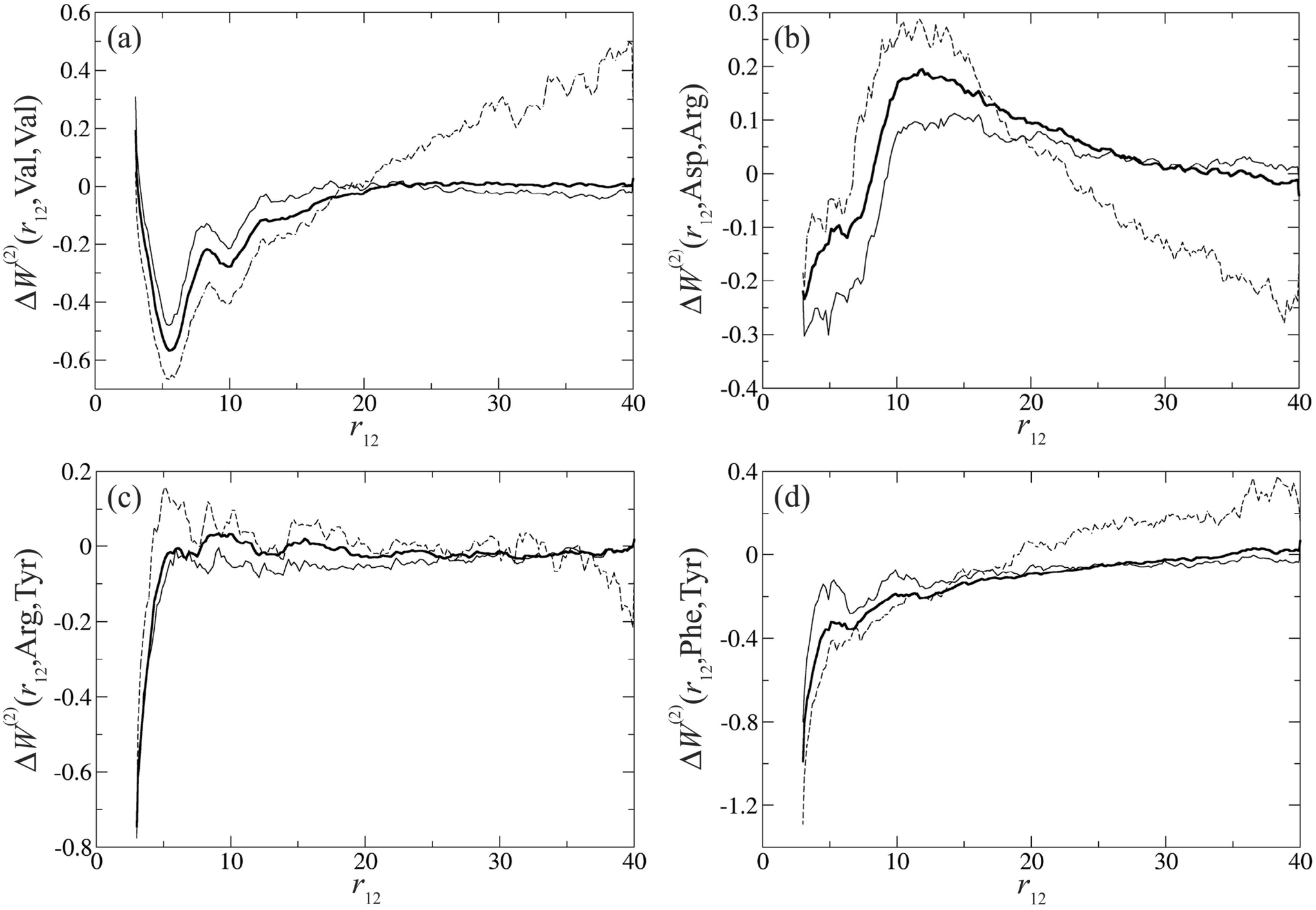
- It is possible to compute distinctly the frequencies for amino acids separated by 2 to 8 amino acids along the sequence and for amino acids separated by more than 8 amino acids (local and non-local potentials) .



Frequency of association between amino acid pairs and distance domains.

Example: number of observations of a pair A-G separated by a distance between 3,2 et 3,4 Å.

!!!!: these potentials depend on the average size of the proteins of the database.



..... :small proteins —— :large proteins —— :all

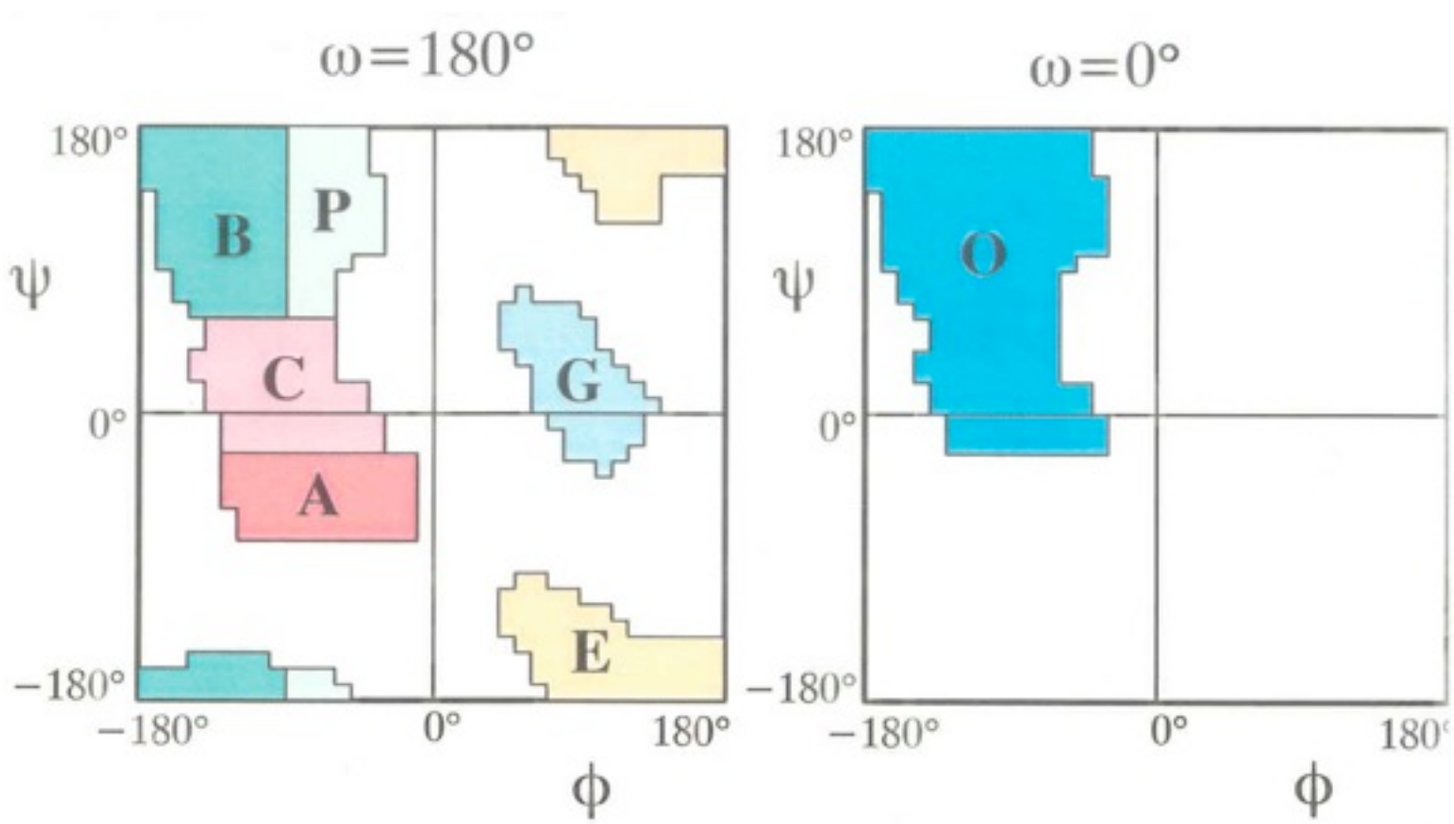
The free energy $\Delta W^{(2)}$ of the sequence S adopting the conformation C computed by a distance potential is (N amino acids in the sequence):

$$\Delta W^{(2)} \cong \sum_{i,j=1}^{N} \Delta W^{(2)}_{|i-j|}(r_{ij};\, s_i, s_j) \cong -kT \sum_{i,j=1}^{N} \ln \frac{F^{|i-j|}(d_{ij} | s_i, s_j)}{F^{|i-j|}(d_{ij})}$$

This distance potential is a mean force potential that includes the coupling between several types of interactions.

2.3.2. Torsion potentials

In a torsion potential, the structure elements are main chain torsion angle domains.



$F(t_i|s_j)$, $F(t_i,t_j|s_k)$, $F(t_i|s_j,s_k)$ is computed (t: torsion domain, s: amino acid type)

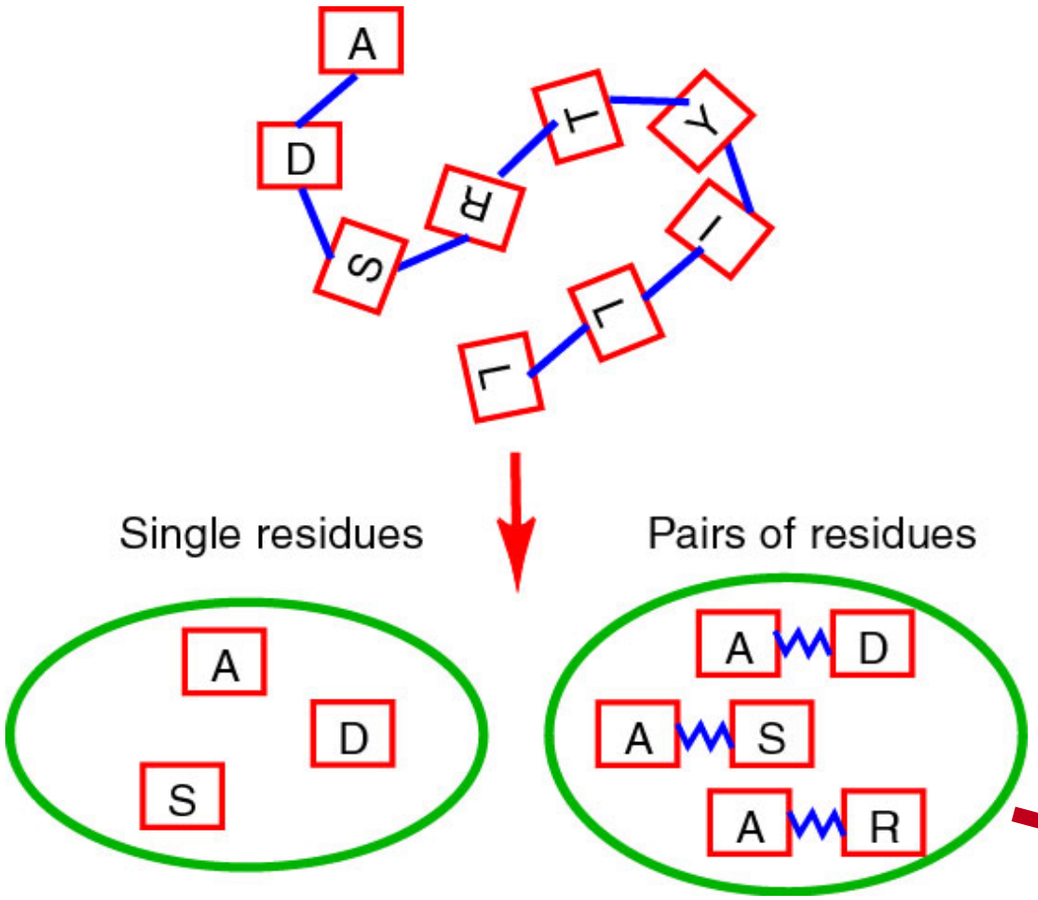
Type 1: influence of a residue pair on a torsion domain
$i-8 \leq j \leq k \leq i+8$:

$$\Delta W^{(2)} \cong \sum_{i,j=1}^{N} \Delta W^{(2)}(t_i; s_j,s_k) \cong -kT \sum_{i,j=1}^{N} \frac{1}{\zeta_i} \ln \frac{F(t_i|s_j,s_k)}{F(t_i)}$$

ζ is a normalization factor, to count each contribution one time.

Type 2: influence of a residue on pair of torsion domains

$$\Delta W^{(2)} \cong \sum_{i,j=1}^{N} \Delta W^{(2)}(t_i,t_j;s_k) \cong -kT \sum_{i,j=1}^{N} \frac{1}{\zeta_i} \ln \frac{F(t_i,t_j|s_k)}{F(t_i,t_j)}$$



Frequency of association between amino acids and backbone torsion angle domains.

2.3.3. Hydrophobicity potential

In this potential, the solvent accessibility of each residue is calculated and the frequency of association between residues and solvent accessibility domains is computed.

$$\Delta W^{(2)} \cong \sum_{i=1}^{N} \Delta W^{(2)}(h_i;s_i) \cong -kT \sum_{i=1}^{N} \ln \frac{F(h_i|s_i)}{F(h_i)}$$

All these potentials are dominated by different types of interactions.

2.4. Comparison semi-empirical potentials / database-derived potentials

Semi-empirical

↗ Correspond to well defined interactions with a clear physics background

↘ Solvent and entropic effects

Database-derived

↗ Simplified protein representation is possible.

↗ Take into account entropic effects.

↗ Take into account the solvent.

↘ Contribution of the different interactions is less obvious.

↘ Depend on some characteristics of the database.

2.5. Evaluation of the performances of energy functions (for protein structure prediction)

In protein structure prediction, energy functions must discriminate the native structure from all non-native conformations. The native structure must correspond to the energy minimum.

The performances of energy functions can be rated with decoy sets.

Decoy sets: sets of native and non-native protein structures. They are created *in silico*.

There exists several approaches to create decoy sets:

- Simulation: protein folding is simulated, decoys correspond to conformations obtained during the folding trajectory.

- Comparative modelling: it is a prediction method and decoys correspond to bad predictions.

- Sequence exchange: the sequence of a protein is «mounted» on the structure of another protein.

Good decoy set:

- Contains conformations that are close to the native structure.
- Contains structures of different proteins, belonging to different structural classes.
- Contains a large number of structures.
- Contains structures that are representative of different regions of the conformational space.

Structure close to the native: how to rate that ?

- percentage of native contacts;
- rmsd

Good energy function:

- The native structure is computed with the lowest energy.
- The energy function is able to discriminate the native structure from the decoys.
- The energy of non-native structures increases when the (structural) similarity with the native structure decreases.