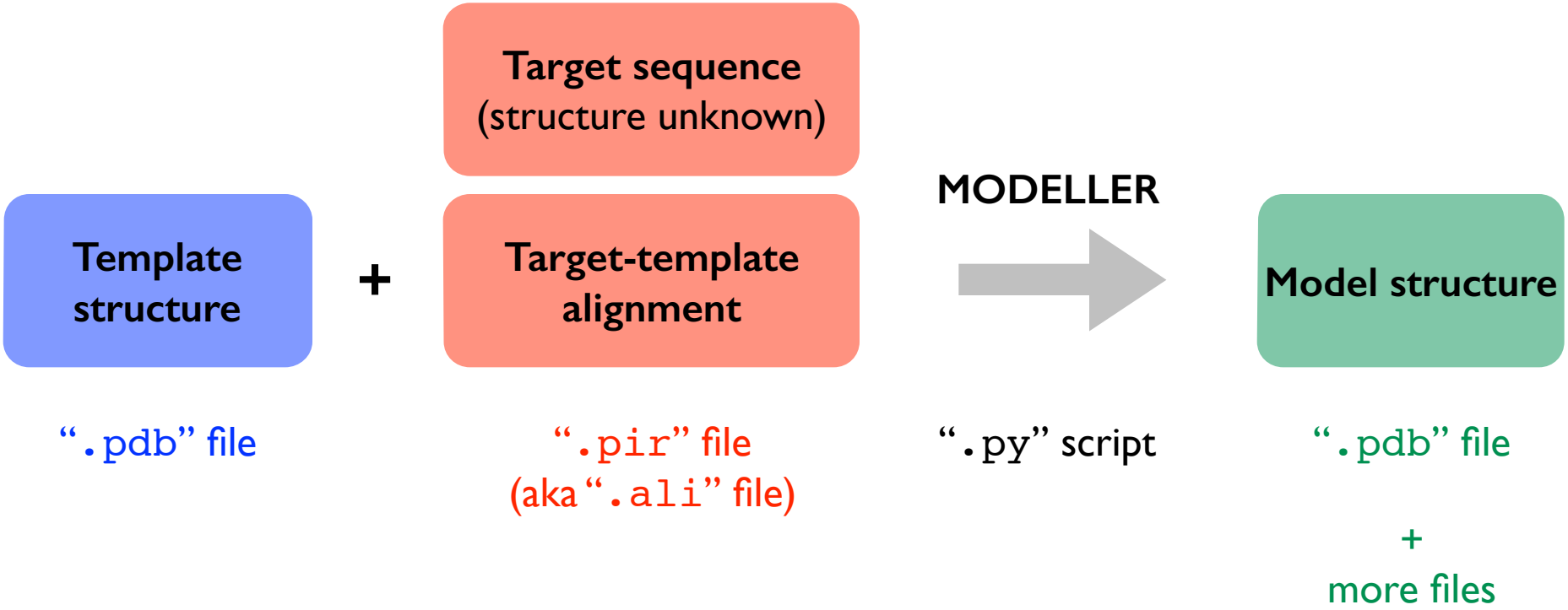
# CHEM 498Q / 630Q

## Molecular modelling of proteins

**Fall 2015 Term**

**Instructor:**
Guillaume Lamoureux
Concordia University, Montréal, Canada

# Homology modelling



# MODELLER's `automodel` command

1. Read the target-template alignment (so that we know which residue to compare to which)

2. List all the restraints that will have to be satisfied by our model

"This $C_\alpha$ should be close to that $C_\alpha$."
"This dihedral angle should be around 110°."
etc.

More about restraints :

A. Šali & T. L. Blundell. 1993. *J. Mol. Biol.* **234**, 779-815.
http://salilab.org/pdf/Sali_JMolBiol_1993.pdf

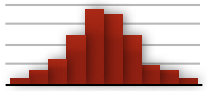
A. Šali & J. Overington. 1994. *Protein Sci.* **3**, 1582-1596.
http://salilab.org/pdf/Sali_ProteinSci_1994.pdf

3. Calculate a model that satisfies the restraints as well as possible

# Spatial restraints

Spatial restraints are described using **probability distribution functions** (pdf's).

**Feature *x*** (distance, angle, etc.) has the following **pdf** :

$$p(x) =$$ 

Examples of features (***x***) :
- $C_\alpha$–$C_\alpha$ distances
- $\phi$ or $\psi$ dihedral angles (backbone)
- $\chi_1$ $\chi_2$ $\chi_3$ $\chi_4$ torsion angles (side-chain rotamers)
- Atom overlaps (Atoms should not overlap.)
- etc.

Probability that *x* falls between $x_1$ and $x_2$ according to the pdf :

$$p(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

pdf's are defined **conditionally**.

**Feature *x*** has the following pdf when conditions ***a***, ***b***, ***c***, ***d*** are met :

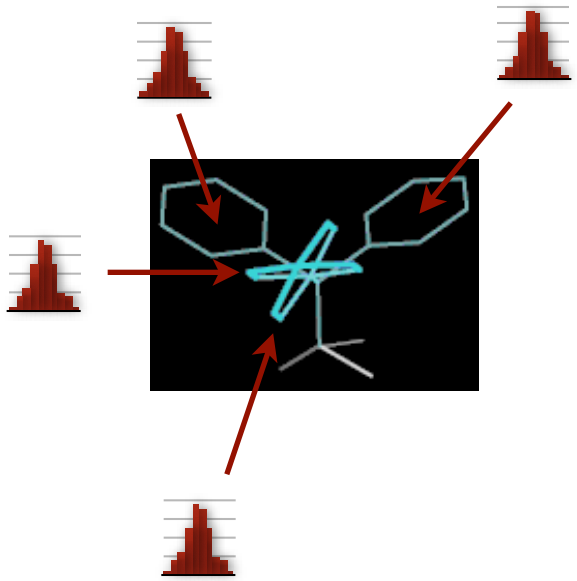
$$p(x|a, b, c, d)$$

Examples of conditions (***a***) :
"The amino acid is of a certain type."
"The backbone is in an $\alpha$-helix conformation."
"The side-chain is in a certain rotameric state."

# Example of a rotamer library

S.C. Lovell, J.M. Word, J.S. Richardson & D.C. Richardson. 2000. "The Penultimate Rotamer Library" *Proteins: Structure Function and Genetics* **40**, 389-408. http://www3.interscience.wiley.com/journal/72504897/abstract

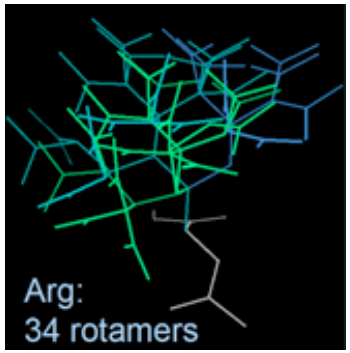
| Name | # | % | alpha | beta | other | $\chi^1$ mode | **$\chi^1$ comm.** | $\chi^2$ mode | **$\chi^2$ comm.** | $\chi^2$ range | $\chi^1$ 1/2 Width at 1/2 Height | $\chi^2$ 1/2 Width at 1/2 Height |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Phenylalanine | | | | | | | | | | | | |
| p90° | 202 | 13% | 1% | 24% | 11% | 59 | **62** | 88 | **90** | 60 to 90, -90 to-60 | 11 | 11 |
| t80° | 522 | 33% | 57% | 18% | 29% | 177 | **-177** | 80 | **80** | 20 to 90, -90 to -75 | 13 | 17 |
| m-85° | 697 | 44% | 29% | 51% | 47% | -64 | **-65** | -83 | **-85** | 50 to 90, -90 to -50 | 12 | 17 |
| m-30° | 149 | 9% | 12% | 5% | 11% | -64 | **-65** | -19 | **-30** | -50 to 0, 0 to 50 | 9 | 20 |
| | | 98% | 97% | 99% | 98% | | | | | | | |
| | 1570/1599 | | 389 | 514 | 667 | | | | | | | |



# Example of a rotamer library

S.C. Lovell, J.M. Word, J.S. Richardson & D.C. Richardson. 2000. "The Penultimate Rotamer Library" *Proteins: Structure Function and Genetics* **40**, 389-408.
http://www3.interscience.wiley.com/journal/72504897/abstract

| Name | # | % | alpha | beta | other | $\chi^1$ mode | **$\chi^1$ comm.**[a] | $\chi^2$ mode | **$\chi^2$ comm.** | $\chi^3$ mode | **$\chi^3$ comm.** | $\chi^4$ mode | **$\chi^4$ comm.** | $\chi^1$ | $\chi^2$ | $\chi^3$ | $\chi^4$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Arginine** | | | | | | | | | | | | | | 1/2 Width at 1/2 Height | | | |
| ptp85°[b] | 3 | <1%[c] | 0% | 1% | <1% | | 62 | | 180 | | 65 | | 85 | | | | |
| ptp180° | 11 | 1% | 0% | 2% | 2% | 71 | 62 | 171 | 180 | 65 | 65 | -161 | -175 | 14 | 17 | 10 | 13 |
| ptt85° | 16 | 2% | 1% | 2% | 2% | 65 | 62 | -178 | 180 | -179 | 180 | 88 | 85 | 13 | 14 | 13 | 17 |
| ptt180° | 16 | 2% | 1% | 2% | 2% | 59 | 62 | 176 | 180 | -178 | 180 | -177 | 180 | 15 | 13 | 15 | 19 |
| ptt-85° | 15 | 2% | 1% | 2% | 2% | 66 | 62 | -176 | 180 | -178 | 180 | -83 | -85 | 15 | 14 | 12 | 14 |
| ptm180° | 6 | 1% | 0% | 1% | 1% | | 62 | | 180 | | -65 | | 175 | | | | |
| ptm-85° | 5 | 1% | 0% | 0% | 1% | | 62 | | 180 | | -65 | | -85 | | | | |
| tpp85° | 11 | 1% | 3% | 1% | <1% | -178 | -177 | 57 | 65 | 57 | 65 | 85 | 85 | 13 | 13 | 12 | 15 |
| tpp180° | 8 | 1% | 1% | 0% | 1% | | -177 | | 65 | | 65 | | -175 | | | | |
| tpt85° | 20 | 2% | 3% | 2% | 2% | 177 | -177 | 64 | 65 | 180 | 180 | 86 | 85 | 14 | 13 | 15 | 14 |
| tpt180° | 15 | 2% | 3% | 1% | 1% | 179 | -177 | 60 | 65 | 178 | 180 | 163 | 180 | 13 | 17 | 14 | 17 |
| ttp85° | 33 | 4% | 5% | 3% | 3% | -179 | -177 | 177 | 180 | 65 | 65 | 83 | 85 | 14 | 17 | 13 | 15 |
| ttp180° | 25 | 3% | 5% | 3% | 1% | -178 | -177 | -178 | 180 | 65 | 65 | -162 | -175 | 14 | 16 | 14 | 26 |
| ttp-105° | 9 | 1% | 1% | 1% | 1% | | -177 | | 180 | | 65 | | -105 | | | | |
| ttt85° | 19 | 2% | 2% | 2% | 2% | -175 | -177 | 176 | 180 | 179 | 180 | 83 | 85 | 14 | 14 | 13 | 14 |
| ttt180° | 33 | 4% | 3% | 7% | 3% | -179 | -177 | 177 | 180 | -179 | 180 | 170 | 180 | 15 | 13 | 12 | 27 |
| ttt-85° | 26 | 3% | 3% | 3% | 2% | -179 | -177 | 179 | 180 | 180 | 180 | -86 | -85 | 15 | 14 | 14 | 15 |
| ttm105° | 10 | 1% | 2% | 1% | <1% | -178 | -177 | 170 | 180 | -66 | -65 | 107 | 105 | 15 | 16 | 15 | 15 |
| ttm180° | 13 | 1% | <1% | 4% | 1% | 180 | -177 | -178 | 180 | -67 | -65 | 176 | 175 | 15 | 12 | 11 | 15 |
| ttm-85° | 28 | 3% | 3% | 3% | 3% | -175 | -177 | -178 | 180 | -65 | -65 | -84 | -85 | 14 | 16 | 15 | 14 |
| mtp85° | 22 | 2% | 2% | 3% | 2% | -69 | -67 | 177 | 180 | 64 | 65 | 84 | 85 | 13 | 17 | 13 | 13 |
| mtp180° | 45 | 5% | 4% | 3% | 6% | -65 | -67 | 176 | 180 | 64 | 65 | -174 | -175 | 12 | 19 | 13 | 19 |
| mtp-105° | 7 | 1% | 0% | 2% | 1% | -62 | -67 | 179 | 180 | 67 | 65 | -113 | -105 | 11 | 15 | 13 | 15 |
| mtt85° | 34 | 4% | 4% | 4% | 3% | -67 | -67 | 178 | 180 | 179 | 180 | 83 | 85 | 12 | 19 | 13 | 19 |
| mtt180° | 89 | 9% | 9% | 5% | 12% | -67 | -67 | -178 | 180 | -177 | 180 | 174 | 180 | 14 | 13 | 13 | 21 |
| mtt-85° | 53 | 6% | 4% | 7% | 6% | -66 | -67 | -177 | 180 | -179 | 180 | -83 | -85 | 13 | 13 | 13 | 13 |
| mtm105° | 15 | 2% | 1% | 1% | 2% | -68 | -67 | -179 | 180 | -65 | -65 | 103 | 105 | 12 | 13 | 13 | 15 |
| mtm180° | 48 | 5% | 1% | 4% | 8% | -68 | -67 | 173 | 180 | -64 | -65 | 180 | 175 | 14 | 17 | 13 | 30 |
| mtm-85° | 54 | 6% | 13% | 2% | 3% | -69 | -67 | -167 | -167 | -63 | -65 | -86 | -85 | 14 | 13 | 13 | 13 |
| mmt85° | 7 | 1% | 1% | 1% | 1% | | -62 | | -68 | | 180 | | 85 | | | | |
| mmt180° | 18 | 2% | 1% | 3% | 2% | -63 | -62 | -66 | -68 | -179 | 180 | -168 | 180 | 13 | 13 | 10 | 29 |
| mmt-85° | 22 | 2% | <1% | 4% | 3% | -60 | -62 | -72 | -68 | -178 | 180 | -92 | -85 | 14 | 13 | 15 | 13 |
| mmm180° | 11 | 1% | <1% | 2% | 2% | -64 | -62 | -74 | -68 | -67 | -65 | 172 | 175 | 14 | 15 | 10 | 13 |
| mmm-85° | 22 | 2% | 2% | 3% | 3% | -62 | -62 | -64 | -68 | -61 | -65 | -82 | -85 | 14 | 13 | 15 | 13 |
| | | 82% | 79% | 81% | 84% | | | | | | | | | | | | |
| | 769/938[d] | | 234 | 146 | 389 | | | | | | | | | | | | |



## Table 5

*Definition of the $\chi_1$, $\chi_2$, $\chi_3$ and $\chi_4$ side-chain dihedral angle classes*

| $\chi$ type | Residue types | Class | Range (°) | Mean $\bar{\chi}_i$ (°) | Standard deviation $\sigma(\chi_i)$ (°) |
|---|---|---|---|---|---|
| $\chi_1$ | C, D, E, F, H, I, K, L, M, N, Q, R, S, T, V, W, Y | + | [0, 120] | 63 | 10 |
| | | *t* | [120, 240] | 180 | 10 |
| | | − | [−120, 0] | −63 | 10 |
| $\chi_2$ | E, I, K, L, M, Q, R | + | [0, 120] | 65 | 10 |
| | | *t* | [120, 240] | 180 | 10 |
| | | − | [−120, 0] | −65 | 10 |
| | D | 1 | [0, 180] | 0 | 25 |
| | N | 1 | [−180, 0] | −50 | 10 |
| | | 2 | [0, 80] | 10 | 10 |
| | | 3 | [80, 180] | 140 | 10 |
| | H, W | 1 | [−180, 0] | −75 | 10 |
| | | 2 | [0, 180] | 75 | 10 |
| | F, Y | 1 | [0, 180] | 75 | 10 |
| $\chi_3$ | K, M, R, Q | + | [0, 120] | 65 | 10 |
| | | *t* | [120, 240] | 180 | 10 |
| | | − | [−120, 0] | −65 | 10 |
| | E | 1 | [35, 85] | 60 | 15 |
| | | 2 | [85, 395] | 180 | 35 |
| $\chi_4$ | K | + | [0, 120] | 65 | 15 |
| | | *t* | [120, 240] | 180 | 15 |
| | | − | [−120, 0] | −65 | 15 |
| | R | + | [0, 70] | 45 | 10 |
| | | *t* | [70, 255] | 170 | 35 |
| | | − | [−105, 0] | −80 | 10 |

Approximate mean values and standard deviations for each peak are also given. See Šali (1991) for histograms of $\chi_i$ for each residue type that result in these definitions of the classes.

**Table from :**
A. Šali & T. L. Blundell. 1993. *J. Mol. Biol.* **234**, 779-815.
http://salilab.org/pdf/Sali_JMolBiol_1993.pdf

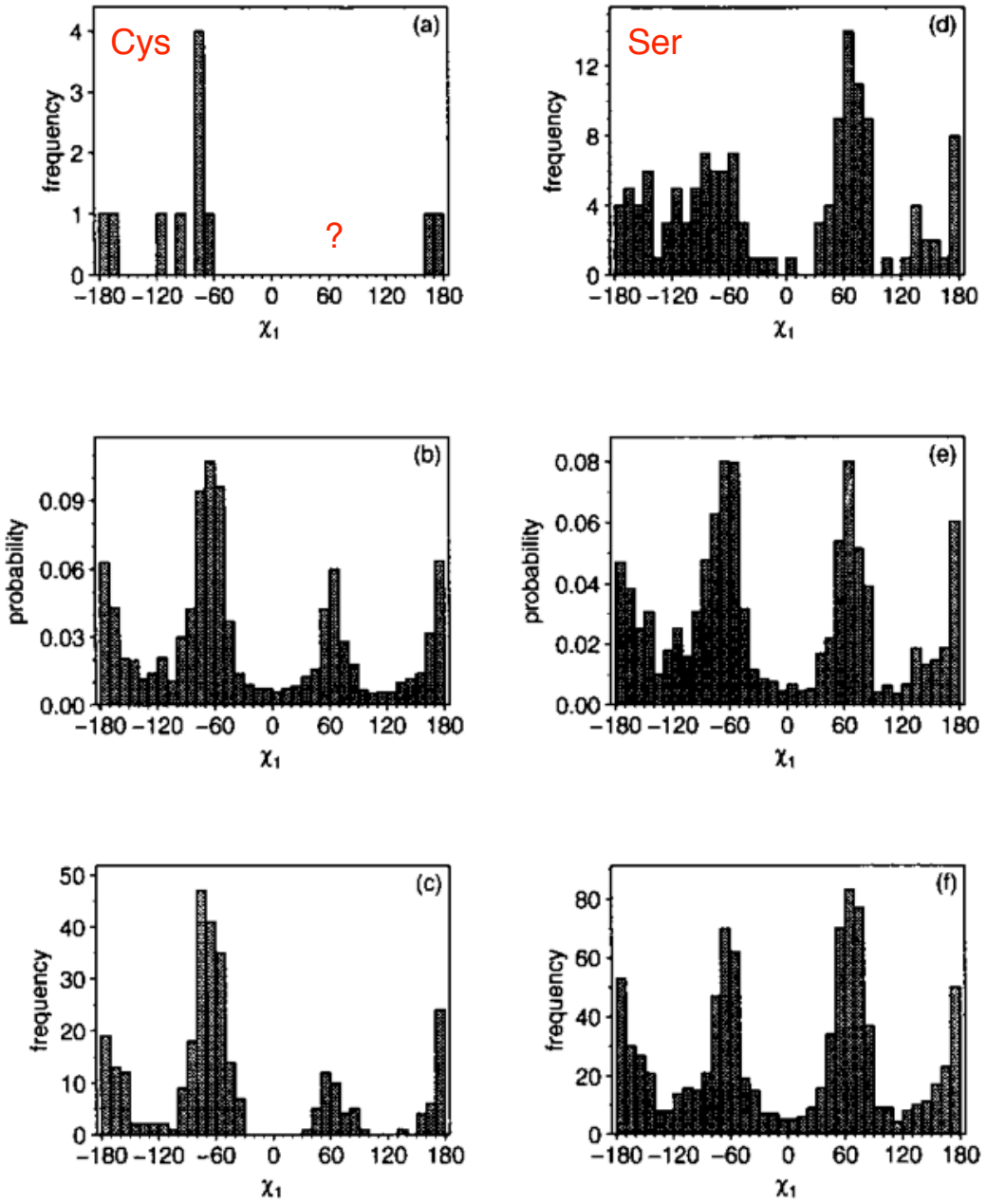



**Figure from :**
A. Šali & T. L. Blundell. 1993. *J. Mol. Biol.* **234**, 779-815.
http://salilab.org/pdf/Sali_JMolBiol_1993.pdf

**Figure 4.** Smoothing sparse distributions of $\chi_1$ side-chain dihedral angles. (a) A distribution of $\chi_1$ angles of 11 Cys residues obtained from a small database of 8 proteins, totalling 1245 residues with $\chi_1$ side-chain dihedral angles. The proteins are 4hvp, 2rspa, 4ape-n, 2app-n, 2apr-n, 3cms-n, 5pep-n, 2apr-c (see Table 1 for the protein names). (b) The smoothed Cys distribution obtained from the small database of 8 structures. The smoothing parameter $\sigma$ was 5. (c) The "reliable" non-smoothed distribution of Cys $\chi_1$ was obtained from the whole local database of 80 protein structures. This distribution consists of 297 Cys residues. (d) 138 Ser residues in the small database. (e) The smoothed Ser distribution. (f) 923 Ser residues in the large database.

# Spatial restraints

We find the list of restraints that are appropriate to apply given the template structure we use.

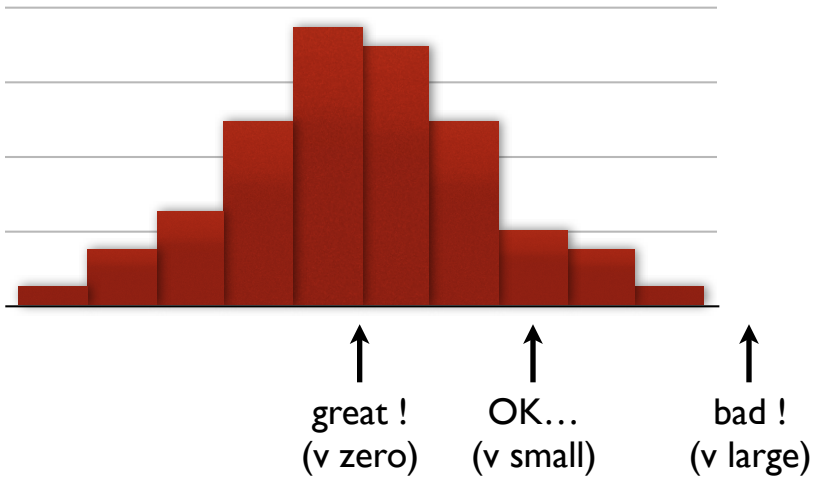
For each residue of the template structure, we pick the pdf that corresponds to the residue type and residue conformation.

If the residue does not match the target residue, we use a pdf that depends only on the residue type (and not on its conformation)

We find the model structure that best satisfies those restraints.

Easier said than done…

We calculate the level of violation for each restraint.



# Structure optimization

**Protein structure :**

$$\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots\}$$

(vector of 3*N* elements $R_i$)

**Violation function :**

$$V = f(\mathbf{R}, \text{restraints})$$

Also includes restraints to enforce a realistic stereochemistry (CHARMM force field)

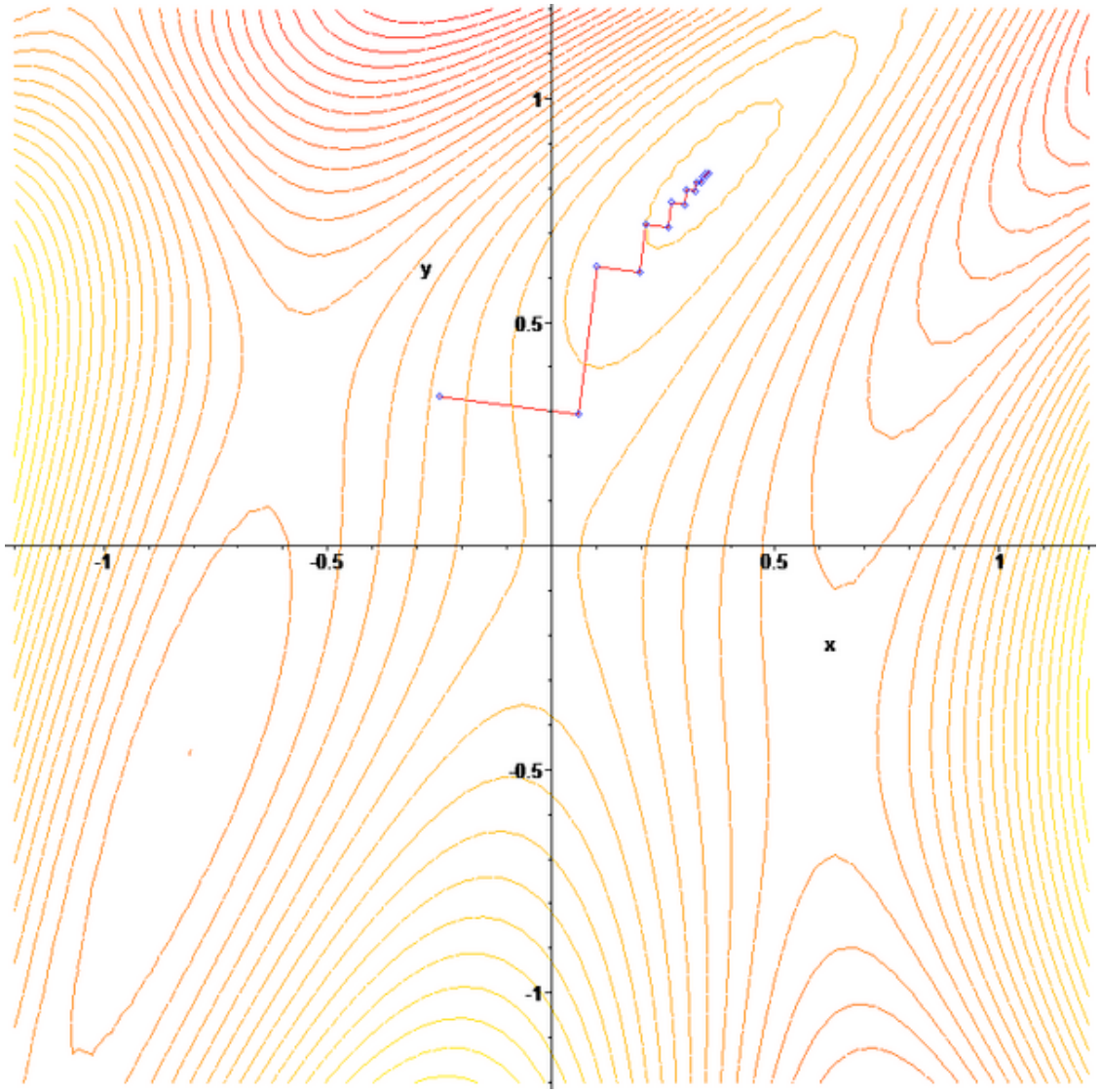



Figure from :
Wikipedia: Gradient descent

# Optimization flowchart in `automodel`

(from the MODELLER user manual: http://salilab.org/modeller/manual)

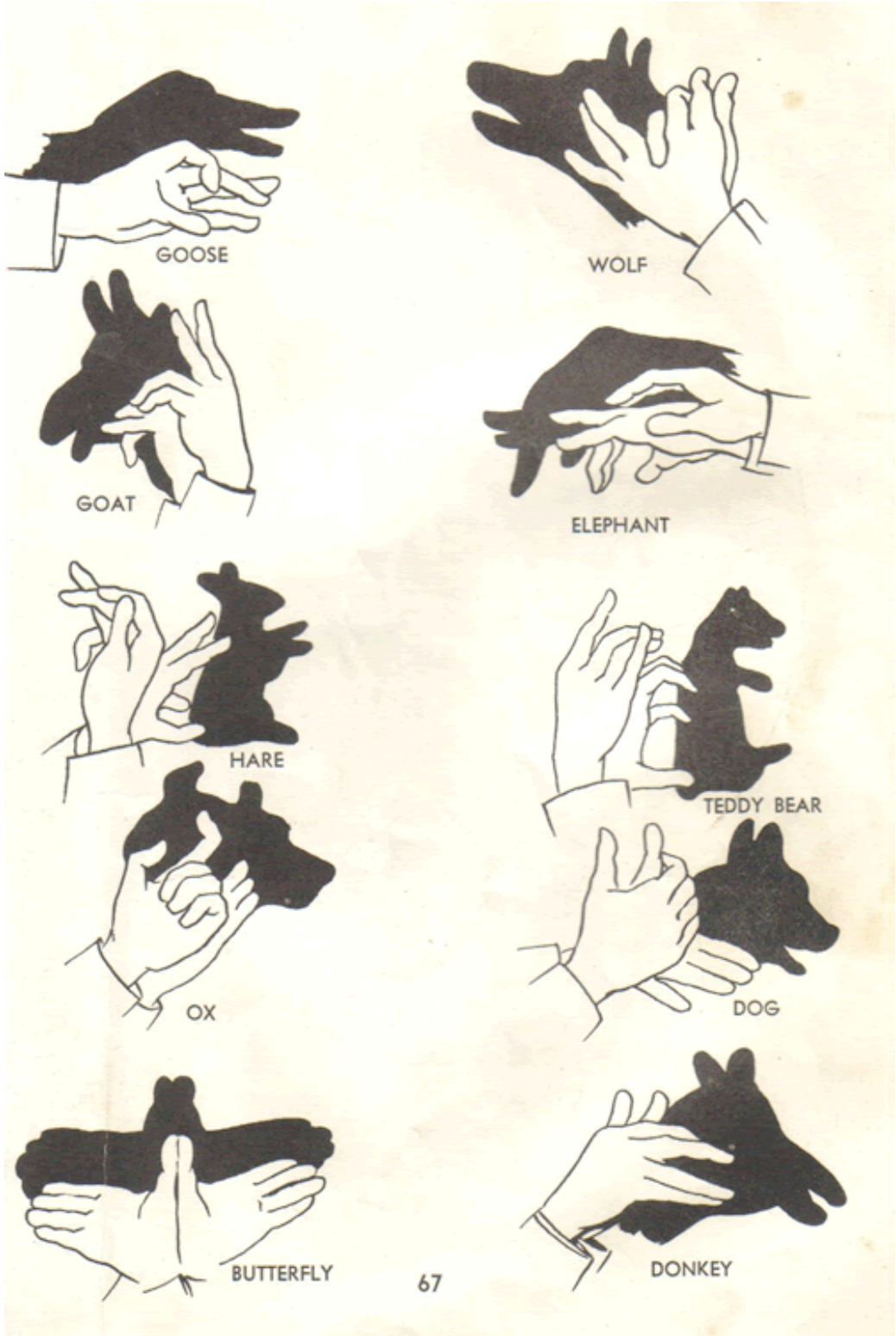
1. Generate the optimization schedule ("variable target function method", VTFM).
2. Read the initial model ("`.ini`" file).
3. Randomize the coordinates of the initial structure.

4. Optimize the model :
   - 4.1 Partially optimize the model by "variable target function method" (VTFM) :
     - Select only the restraints that operate on the atoms that are close enough in sequence, as specified by the current step of VTFM.
     - Optimize the model by conjugate gradients (CG), using only currently selected restraints.
     - Repeat as many times as specified.
   - 4.2 Refine the model by simulated annealing with molecular dynamics (MD) :
     - Do a short CG optimization.
     - Increase T in several steps and do MD optimization at each T.
     - Decrease T in several steps and do MD optimization at each T.
     - Do a short CG optimization.

5. Calculate the remaining restraint violations and write them out.
6. Write out the final model to a file "`.B9999????.pdb`".

**Think you found a good model ?**

"Not so fast."



# DOPE score

**Statistical potential** extracted from high-resolution, nonredundant protein structures.

Tested to see how good it was at scoring conformations of proteins depending on how close they were to the native conformation.

**Reference :**
M.-y. Shen & A. Šali. 2006. *Protein Sci.* **15**, 2507-2524.
http://dx.doi.org/10.1110/ps.062416606

**Statistical potential :**

A protein conformation has a lower (statistical) potential if it is found more often than expected.

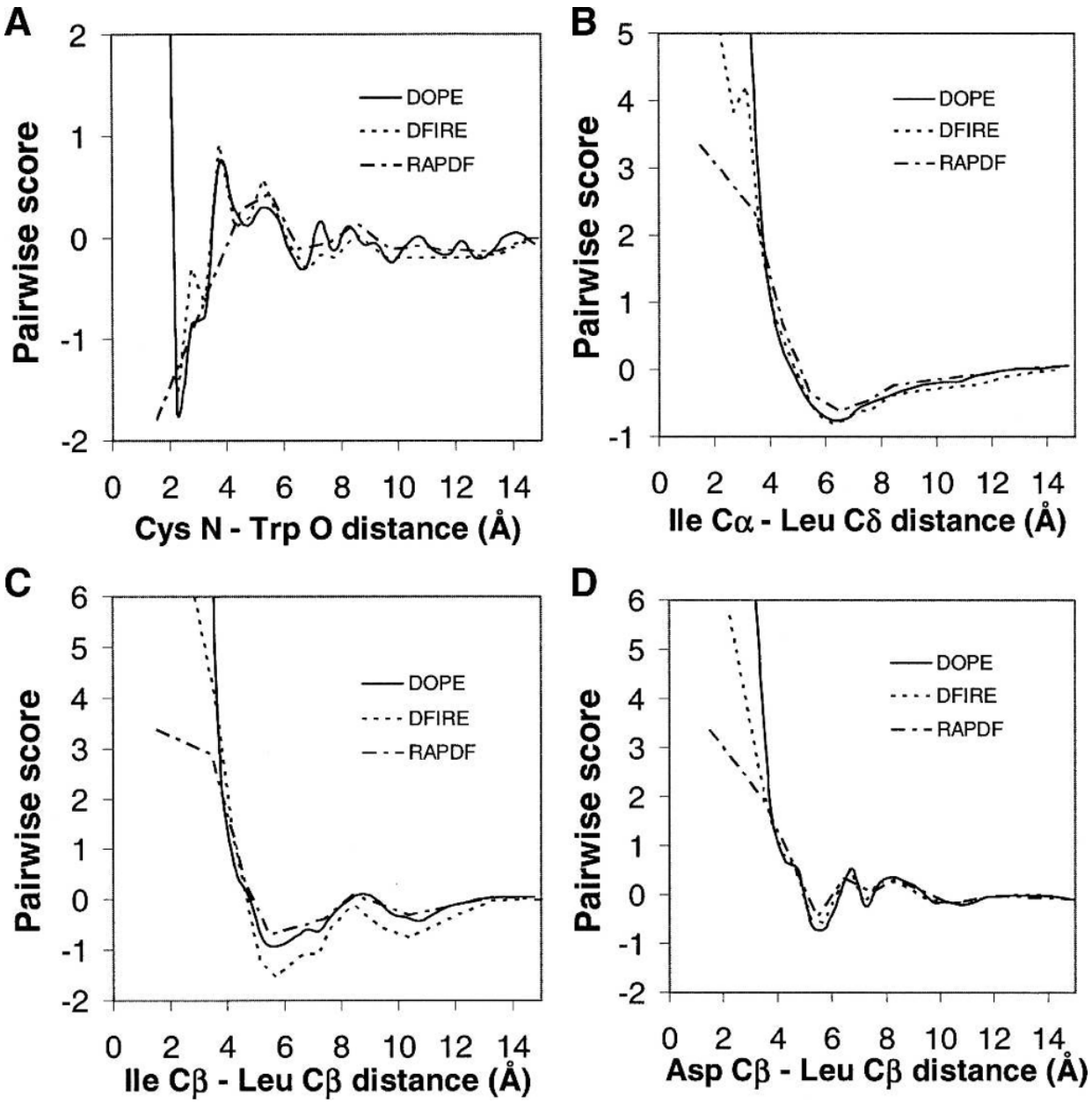
What is our initial level of expectation ?

Problem of defining the "reference state"

→ DOPE uses the "bag of amino acids" state

How do we define conformations ?

→ DOPE uses distances between atoms of different types.




Score for atom pair (*ij*) at distance *r*

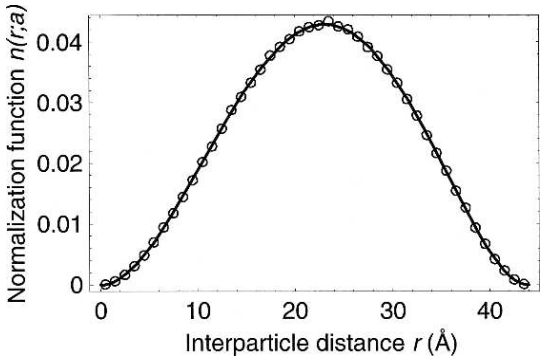
number of atom type pairs (*mn*) at a distance *r* within [*r*, *r*+*dr*] observed in real proteins

$$\mathrm{score}_{ij}(r) = -k_{\mathrm{B}}T \ln \frac{N_{mn}^{\mathrm{OBS}}(r)}{N_{mn}^{\mathrm{REF}}(r)}$$

Boltzmann constant

standard temperature

number in the reference system




Normalization function goes from ***r* = 0** to ***r* = *a***, the size of the protein :

$$a = \sqrt{5/3}\, R_{\mathrm{g}}$$

**Figures from :**

M.-y. Shen & A. Šali. 2006. *Protein Sci.* **15**, 2507-2524.

http://dx.doi.org/10.1110/ps.062416606

# Terminal

| | DOS (on Windows) | Shell (on Linux) |
|---|---|---|
| List the content of current directory : | `C:\> dir` | `> ls` |
| Move into a subdirectory : | `C:\> cd name_of_subdir` | `> cd` |
| Move one directory level down : | `C:\> cd ..` | `> cd ..` |
| Copy a file under a different name : | `C:\> copy name1 name2` | `> cp` |
| Rename a file : | `C:\> ren old_name new_name` | `> mv` |
| View the content of a text file : | `C:\> more name_of_file` | `> more` |
| Edit a text file : | `C:\> edit name_of_file` | `> gedit` |
| Recall the last command typed : | ↑ (“arrow up”) | ↑ (“arrow up”) |