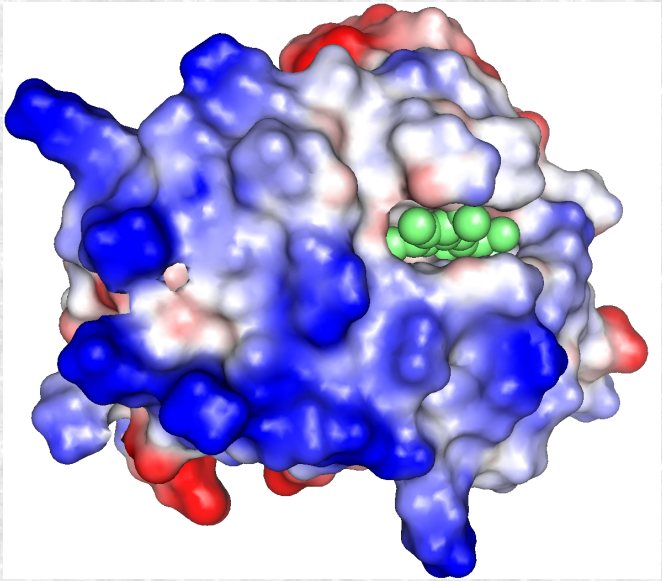
Modeling and predicting protein structure



Mastères *Ingénierie et Chimie des BioMolécules et Biologie et Santé*

Module Bioinformatique Structurale

novembre 2022

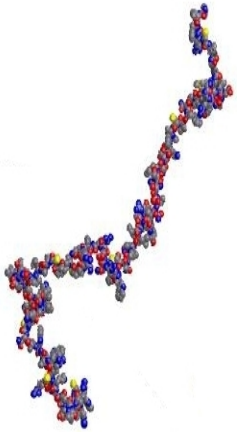
T. Simonson, Ecole Polytechnique

Importance of modeling

For most known proteins, 3D structure is unknown.

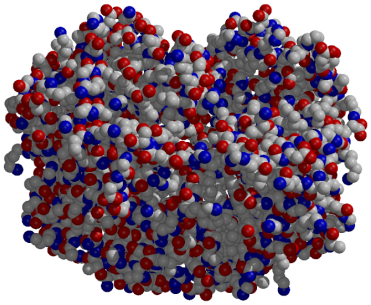
For ~1/3, function is unknown.

AUGCGCUUAUAGCCAAGG:



50% of our dry weight...

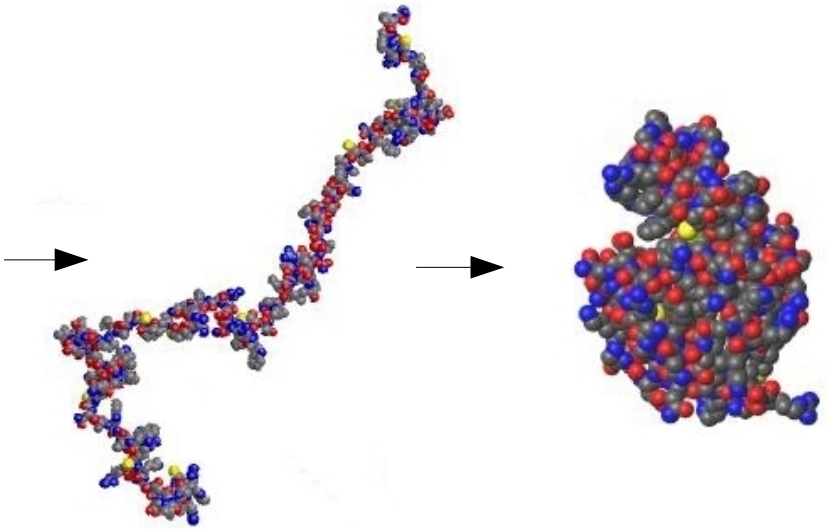
Sequence → **Structure**



Function

The folding problem

K
L
H
G
G
P
M
L
D
S
D
Q
K
F
W
R
T
P
A
A
L
H
Q
N
E
G
F
T



$N_{états} \sim 10^{n}$

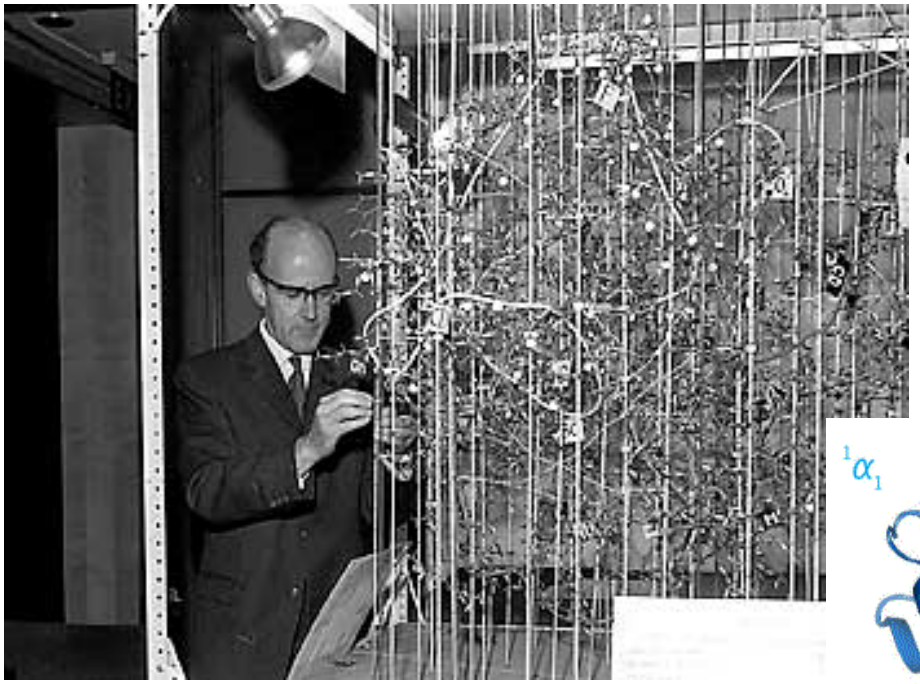
n = 100-300

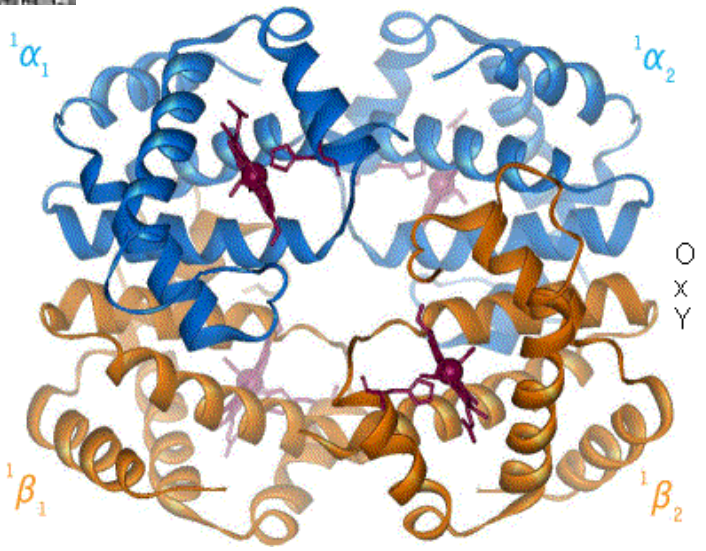
googols of googols….

Levinthal paradox

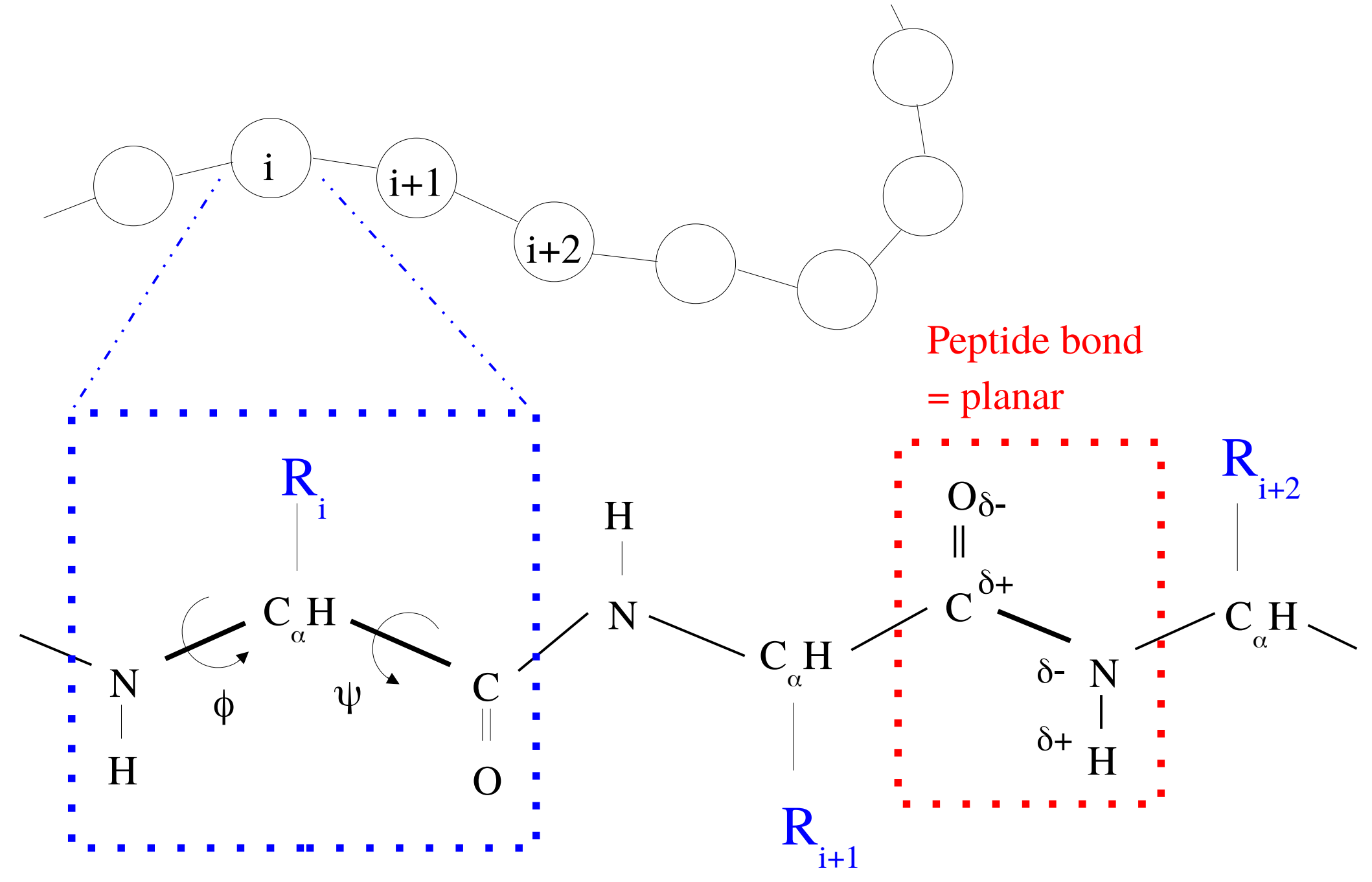
- **Protein structure and stability**
- **Molecular modeling, molecular dynamics**
- ***Predicting secondary structure***
- **Predicting 3D structure: homology modeling**

Protein structure

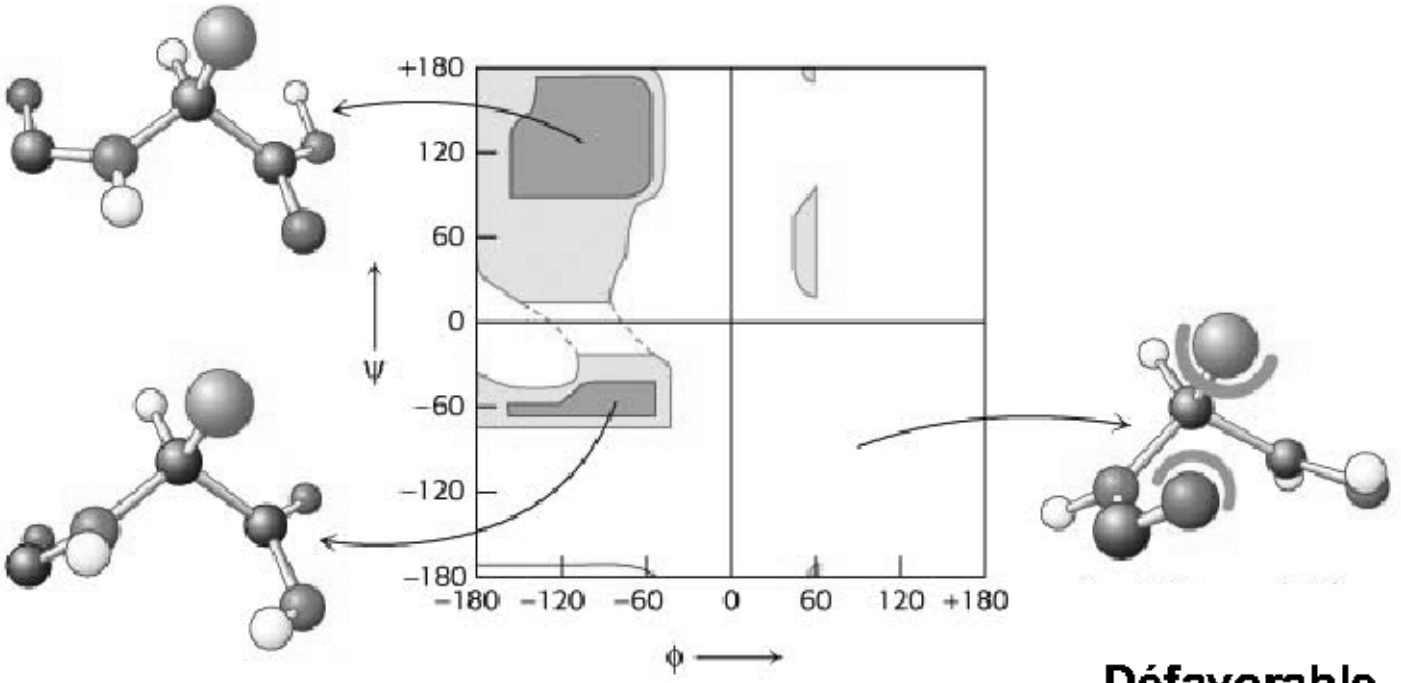


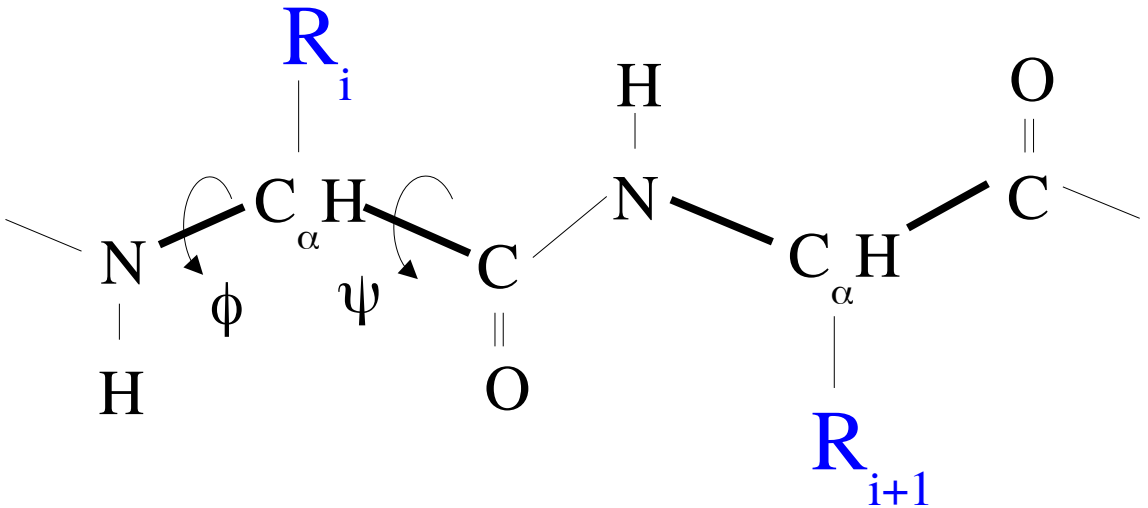


The main chain has two flexible dihedrals per amino acid

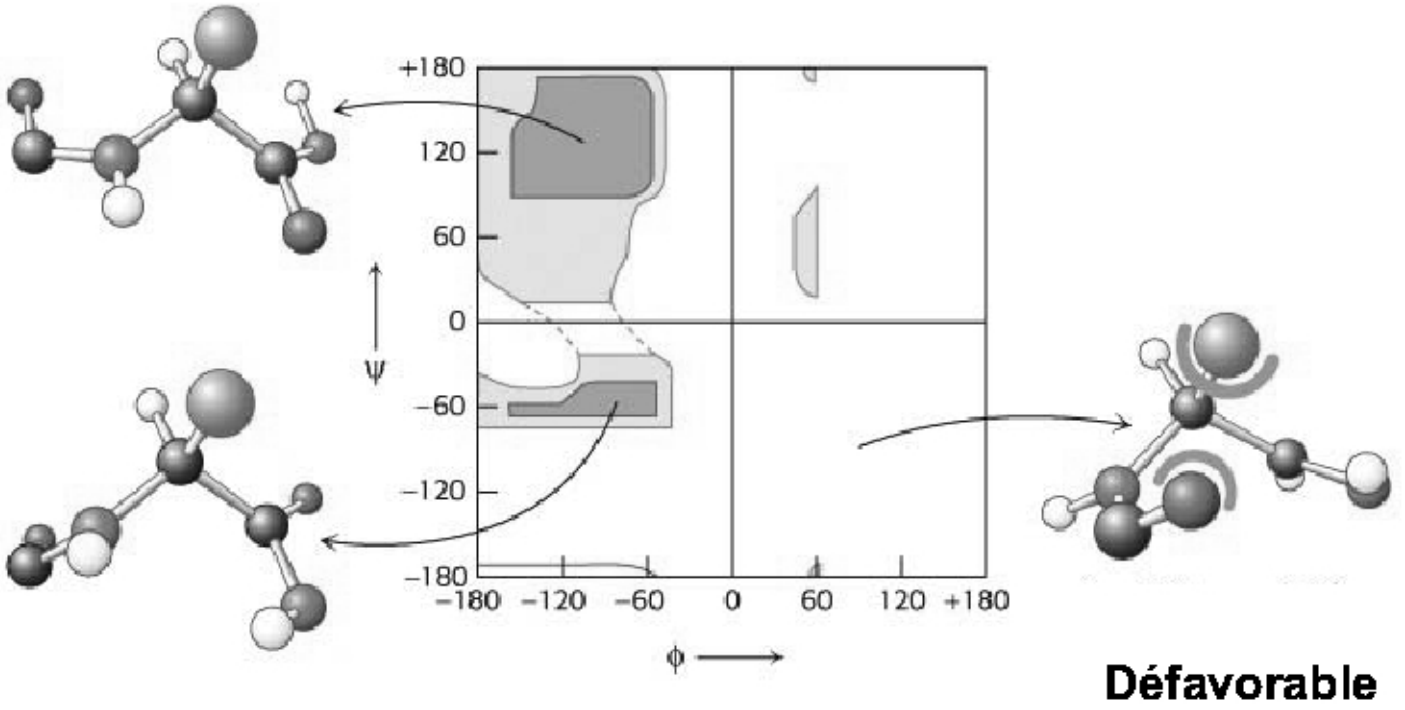


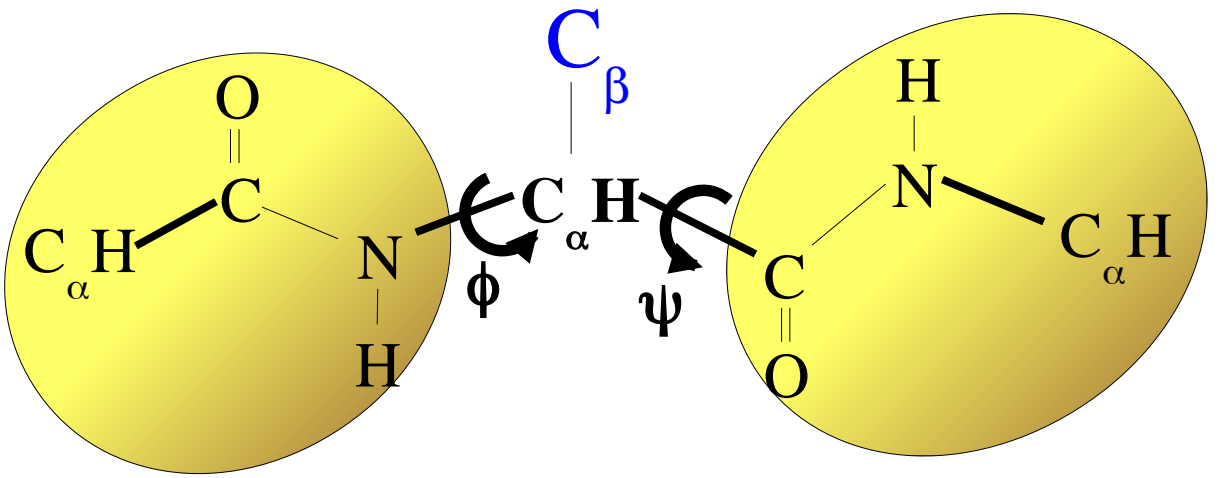
Sterically allowed ϕ, ψ values are those of the Ramachandran plot





Sterically allowed values are those of the Ramachandran plot





Lovell et al, 2003,
Proteins, 50:437

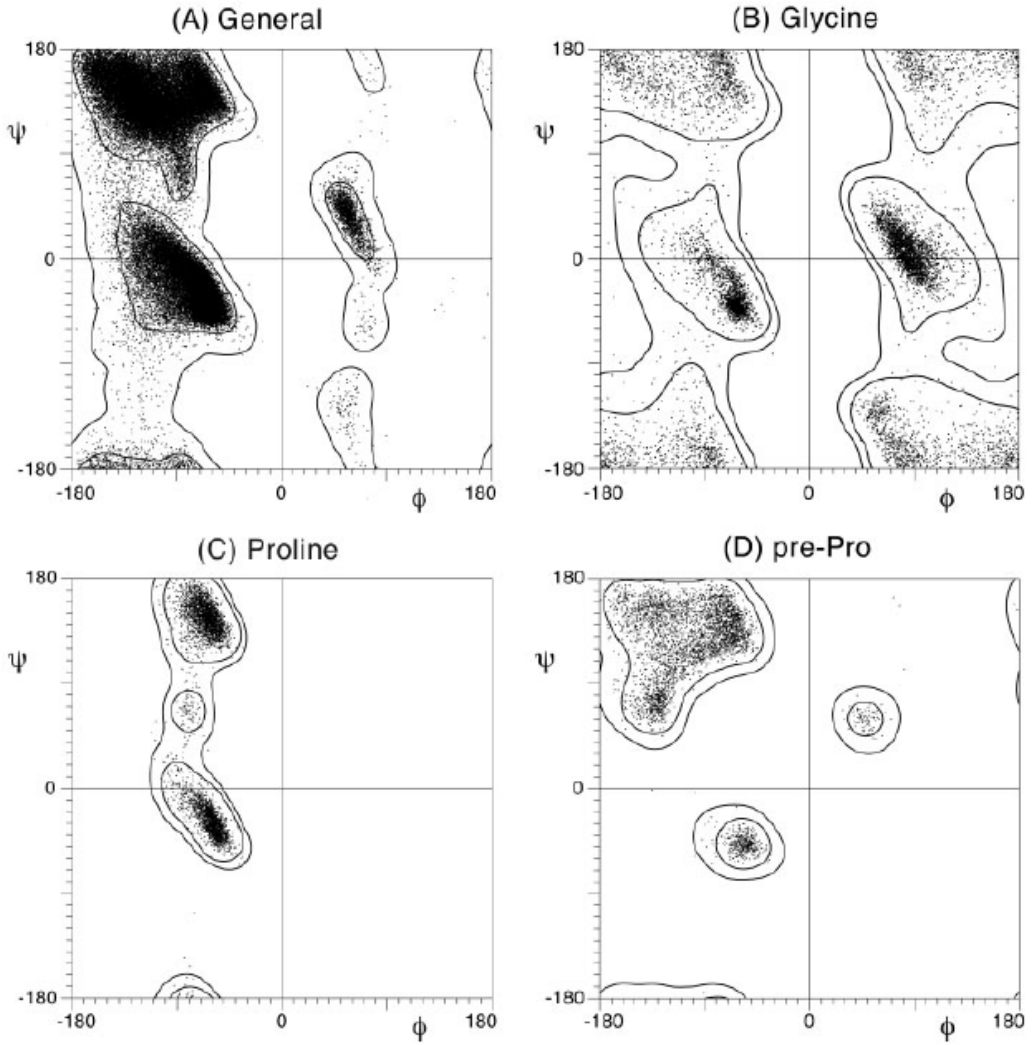
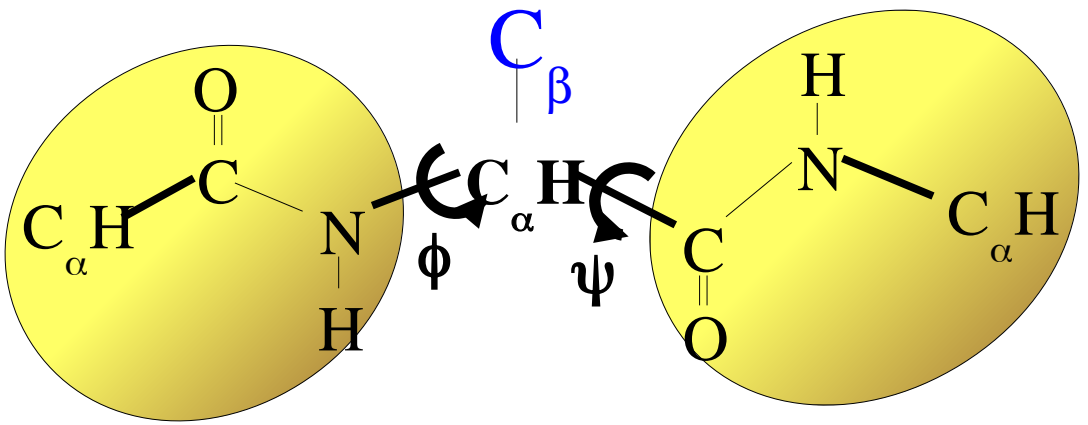
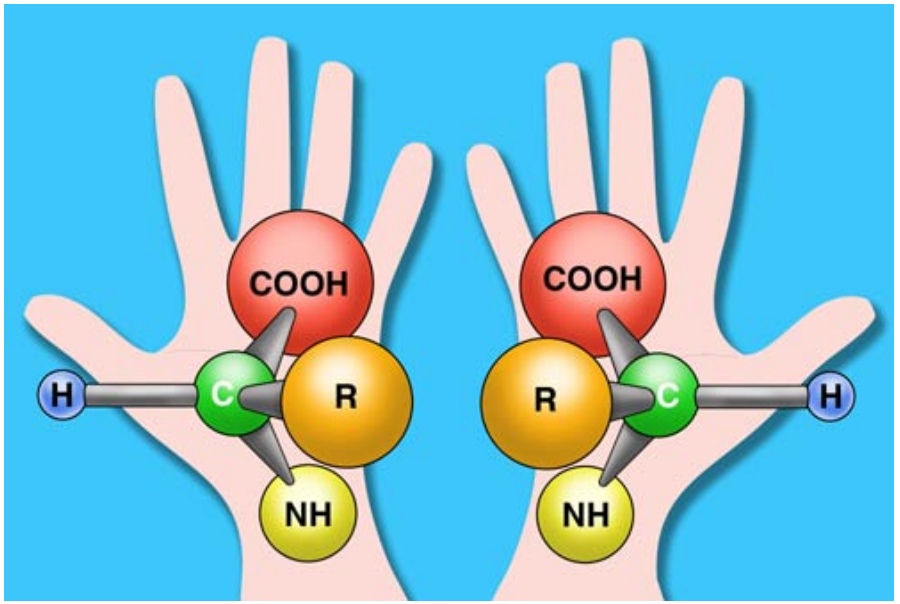


Fig. 4. ϕ,ψ angle distributions for 97,368 residues with backbone *B*-factor < 30 from the 500-structure high-resolution database, along with validation contours for favored and allowed regions. **a:** The general case of 81,234 non-Gly, non-Pro, non-prePro residues. **b:** The 7705 Gly residues, shown with twofold symmetrized contours. **c:** The 4415 Pro residues with contours. **d:** The 4014 pre-Pro residues (excluding those that are Gly or Pro) with contours.



Nature chose L amino acids

L

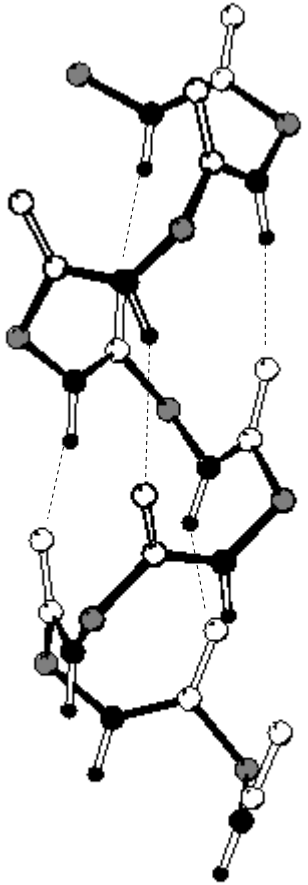


D

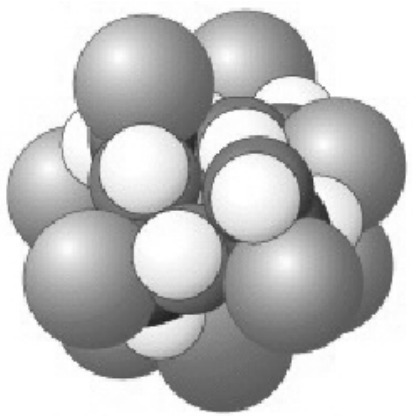
"L" not the same as "levorotary"

In helices and sheets, polar groups form hydrogen bonds

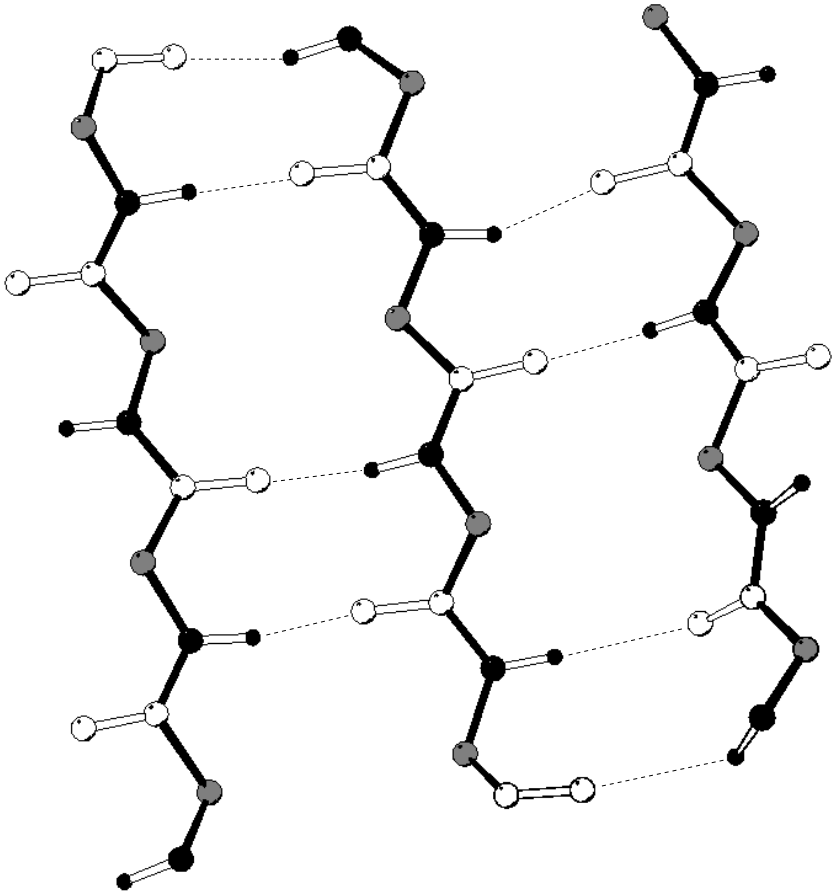
Hélice α



3.6 residues per turn

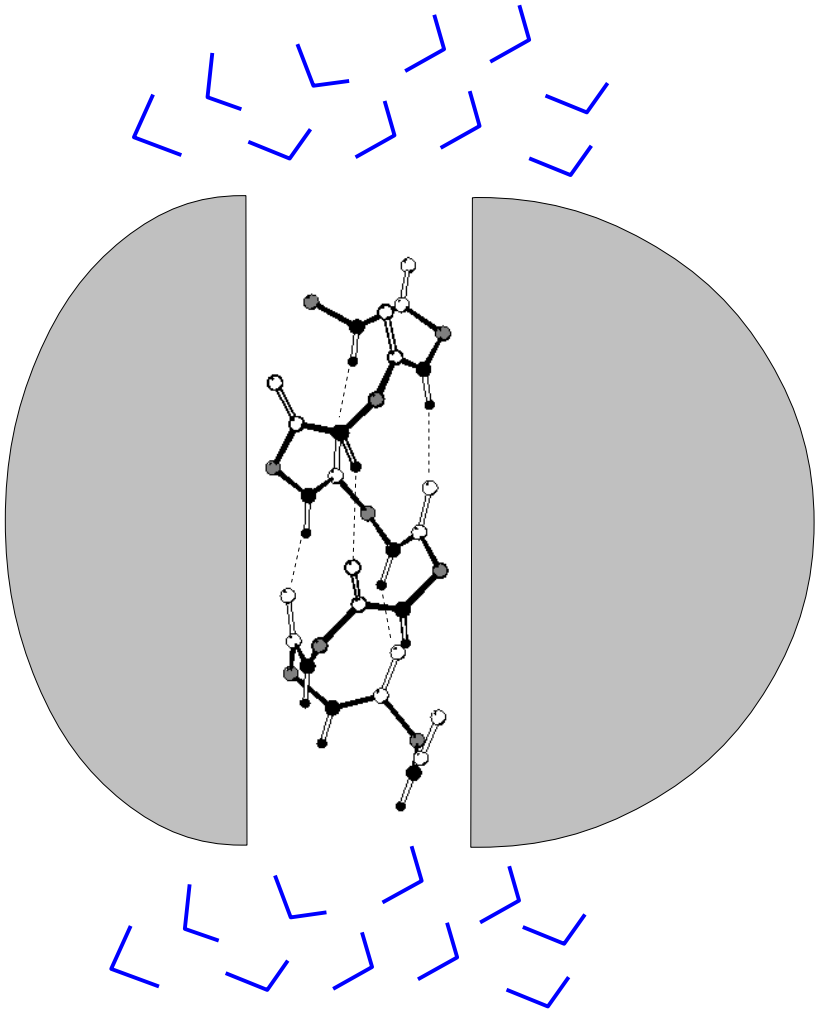


Feuillet β (antiparallel)



Pseudo-period of 2

Helices and sheets are extended structures that can *traverse* the hydrophobic core, while forming all possible hydrogen bonds

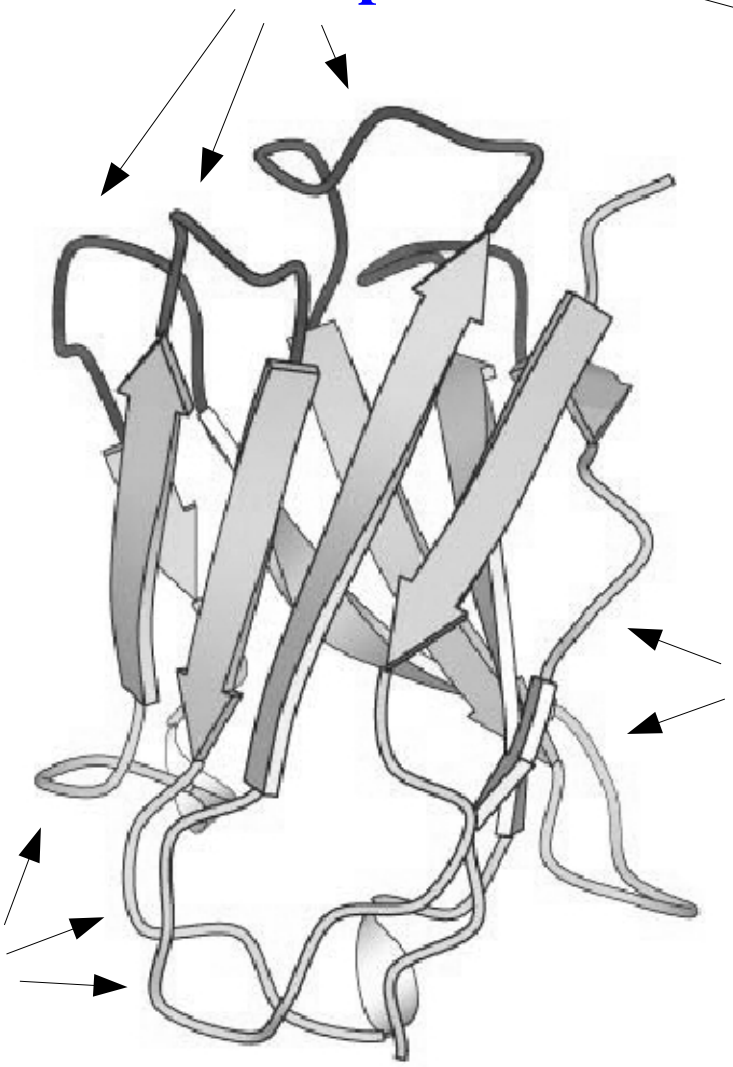


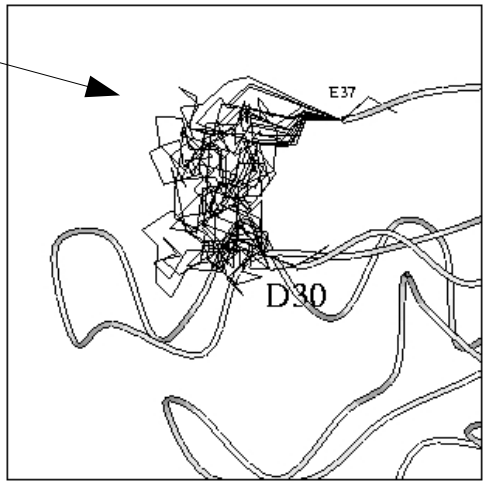
Helices + sheets = over 60% of amino acids

Predicted by Corey & Pauling *before* the experimental structures

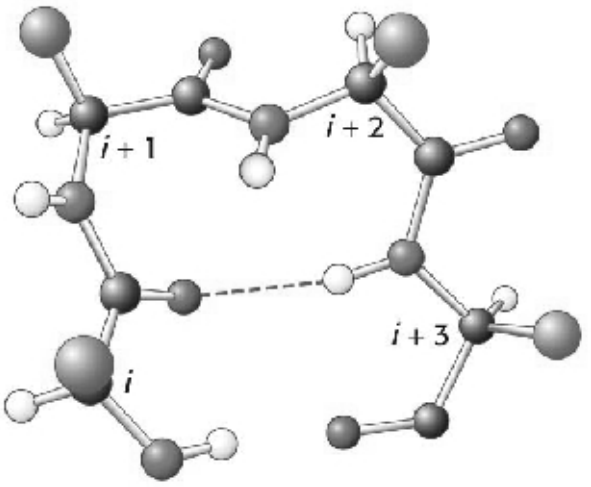
Loops are flexible, less conserved, and harder to predict

Loops



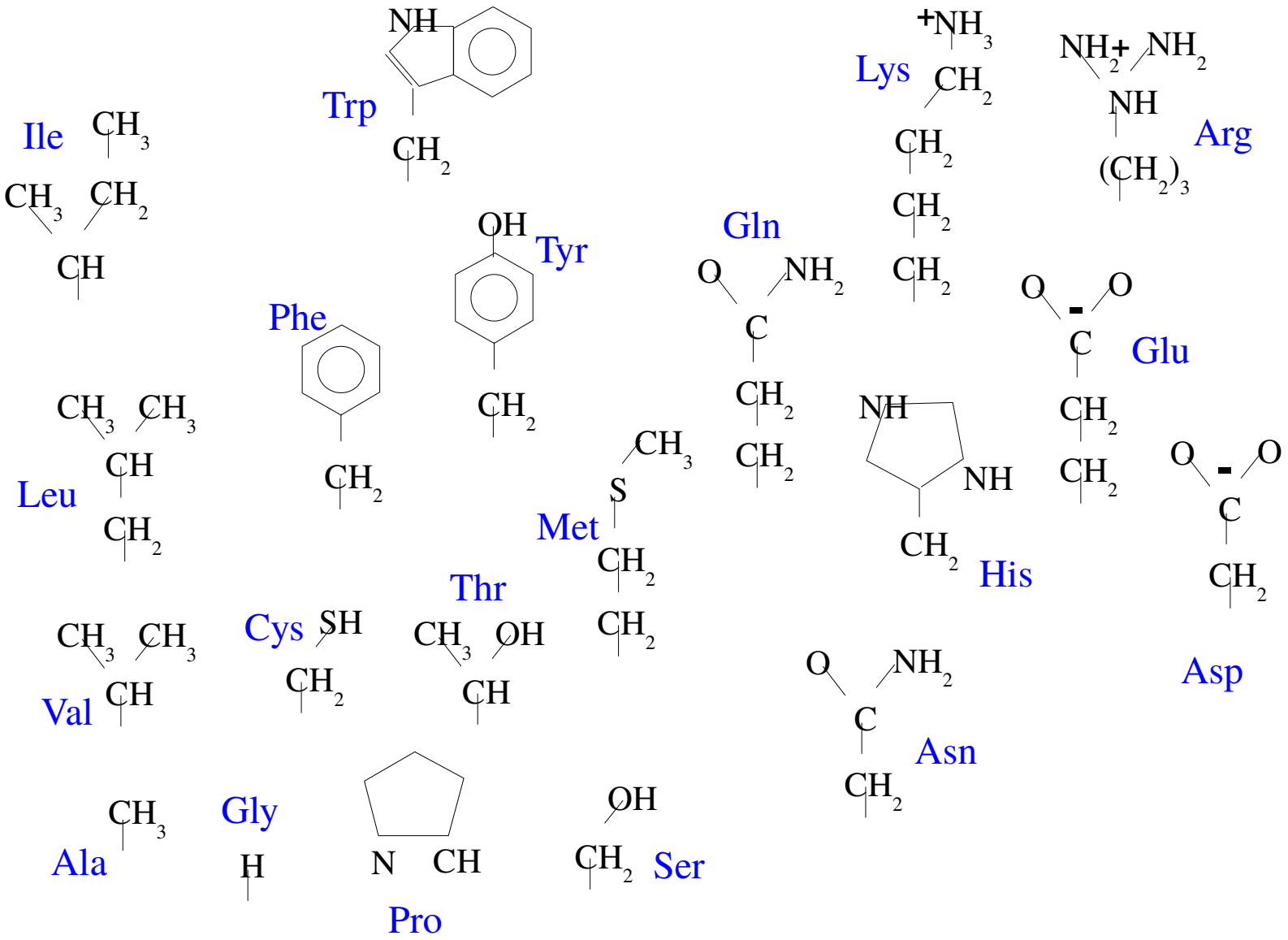


turn

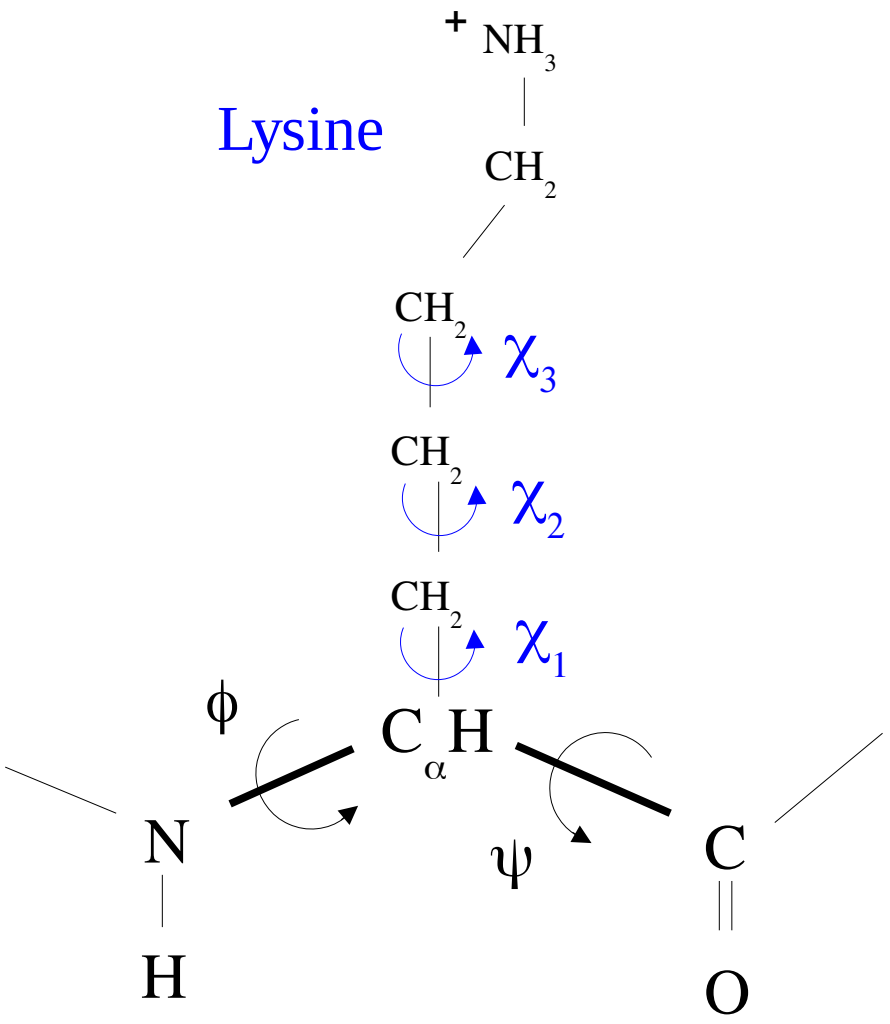


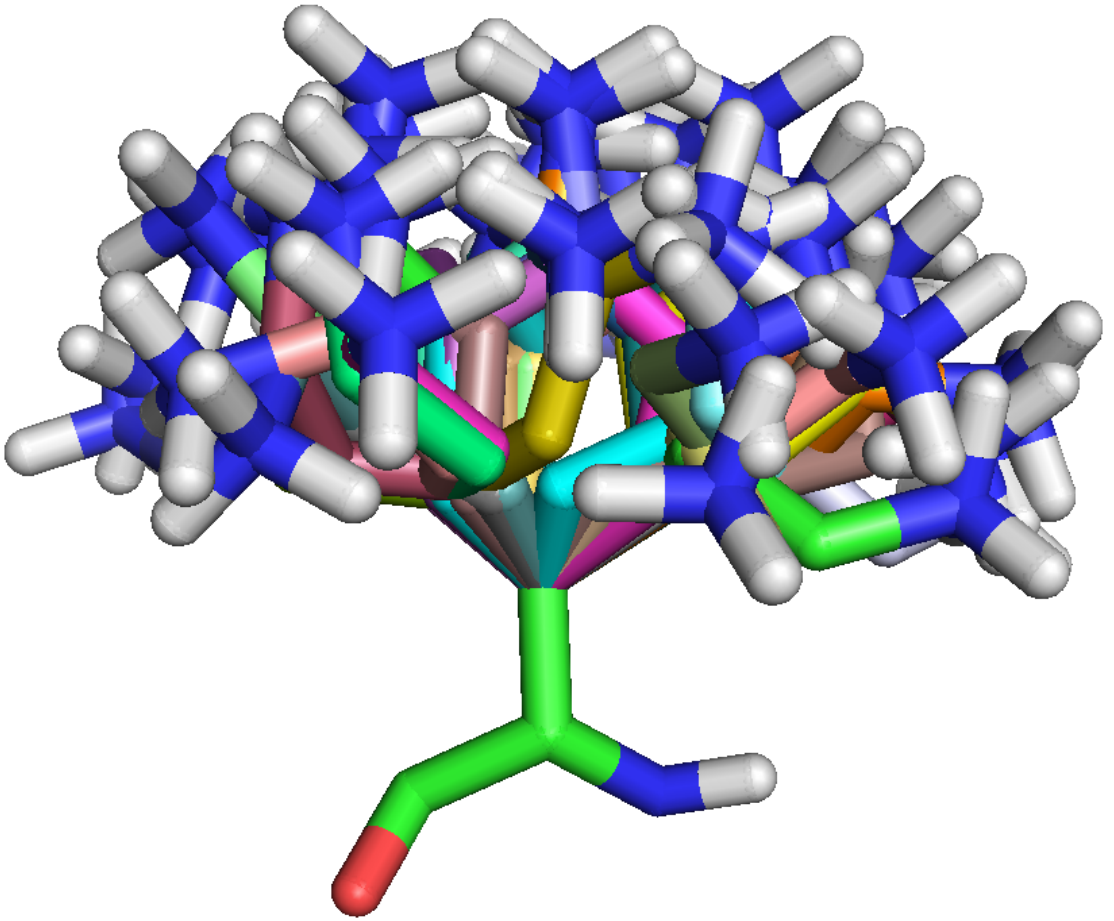
~ 1/3 of amino acids

The amino acids have a range of physico-chimical and structural properties



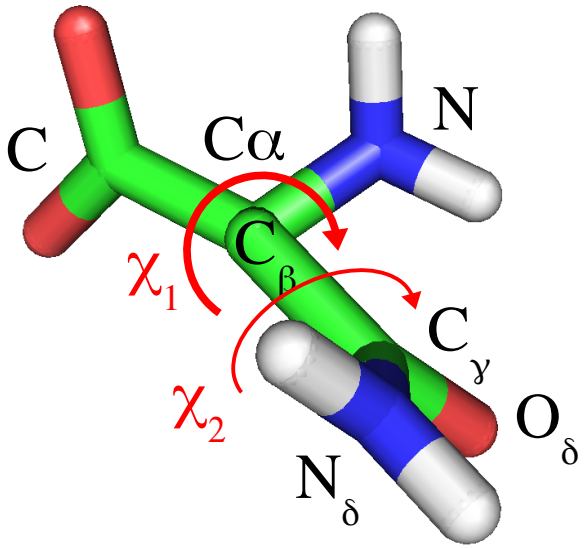
Side chains have flexible torsion angles



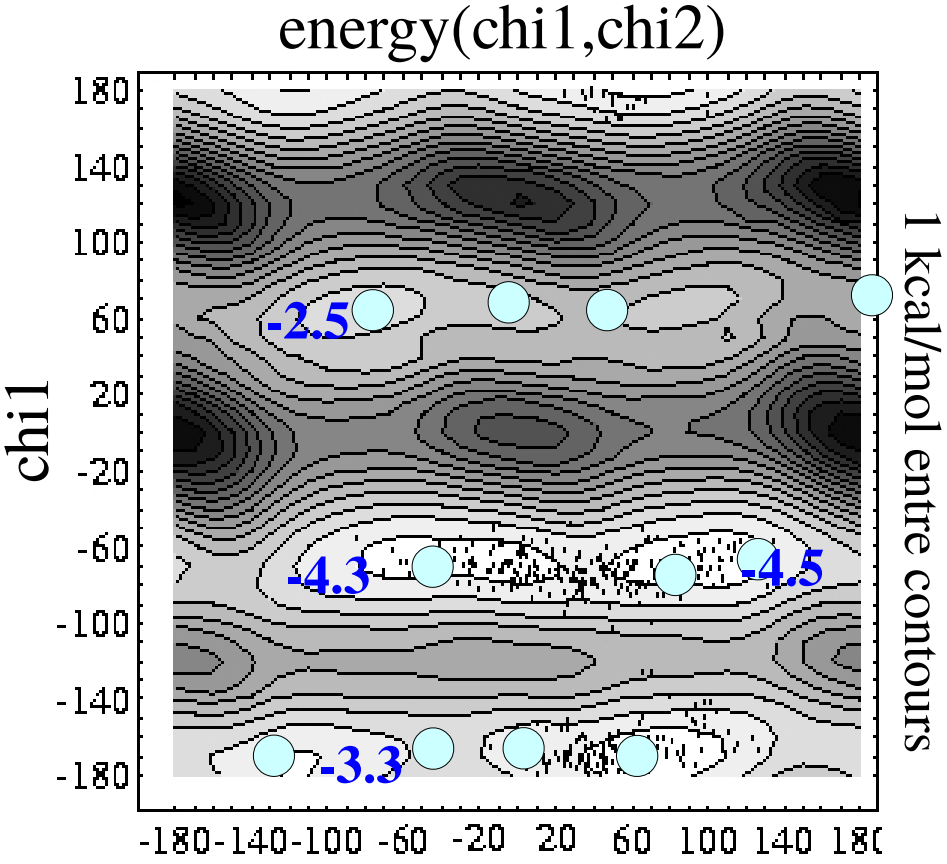


53 conformations of Lys seen in X-ray structures

Side chains have preferred conformations, or “rotamers”

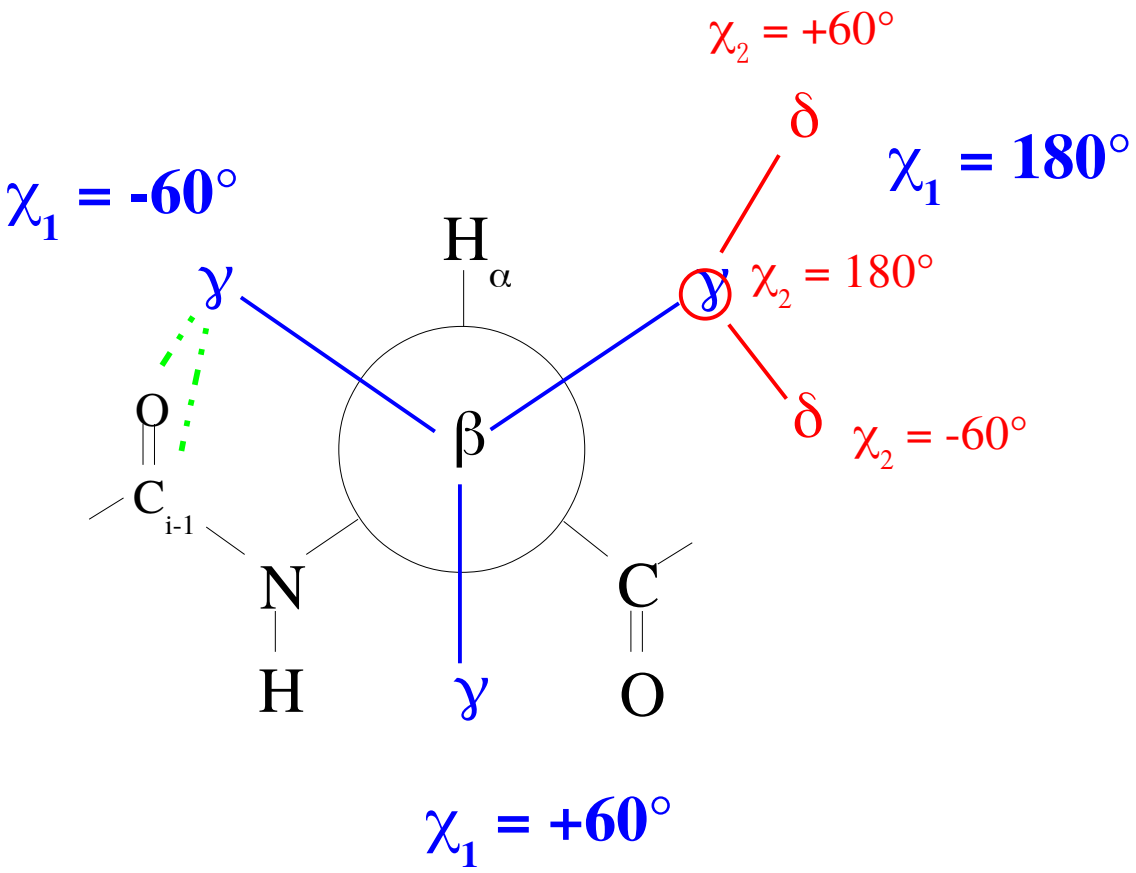


asparagine

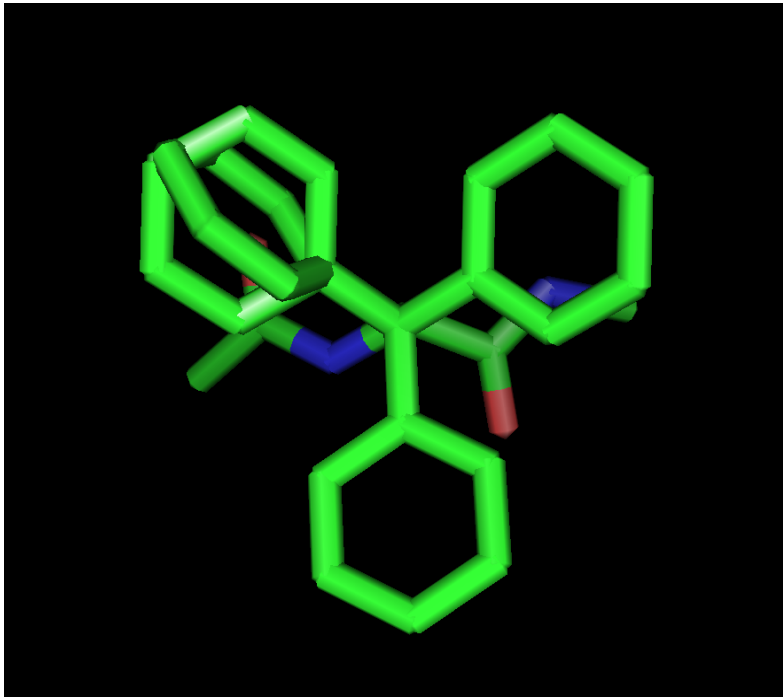


- conformations typical of proteins
- conformations from a simulation

Discrete library of side chain conformations

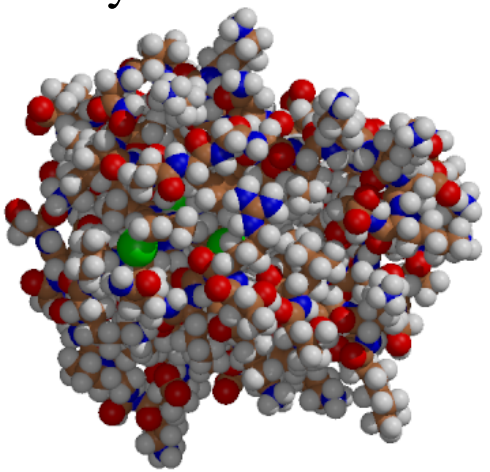


phenylalanine

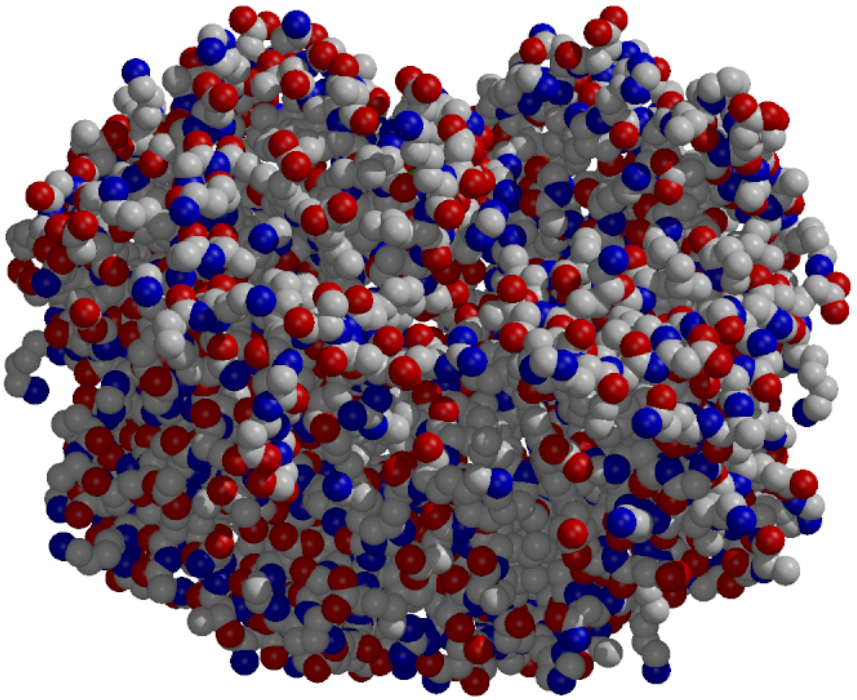


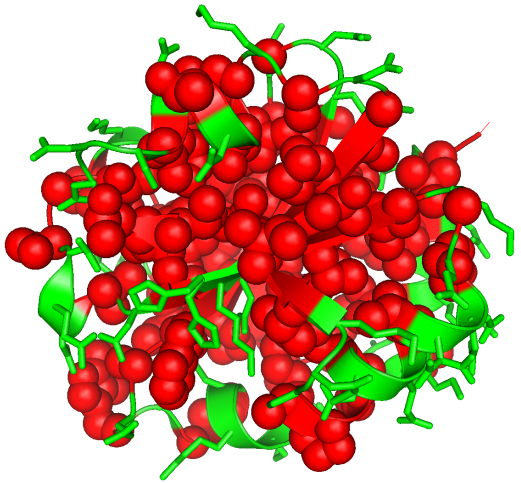
Tertiary structure is globular, with a polar surface and a hydrophobic core

Cytochrome c



Hemoglobin

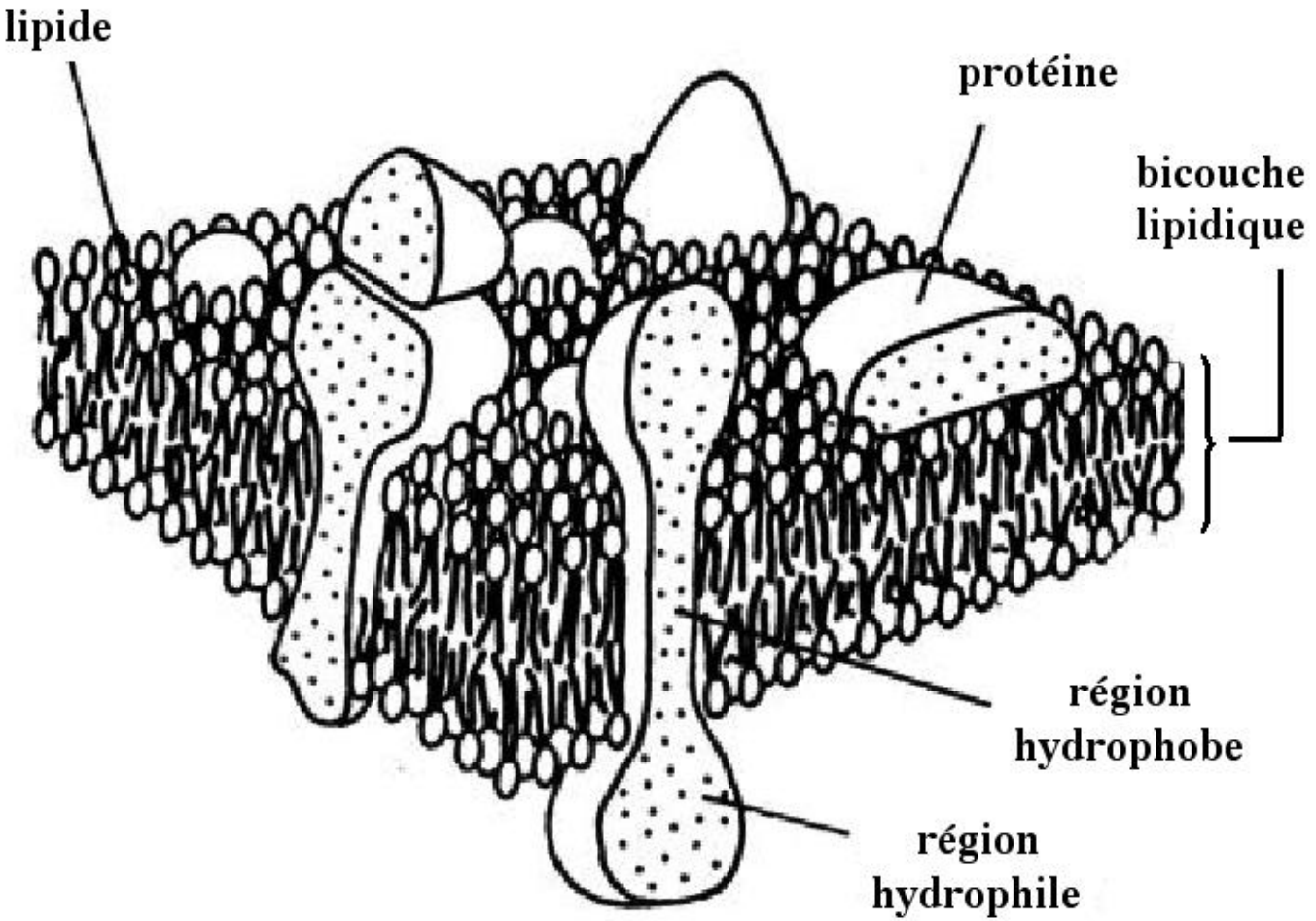


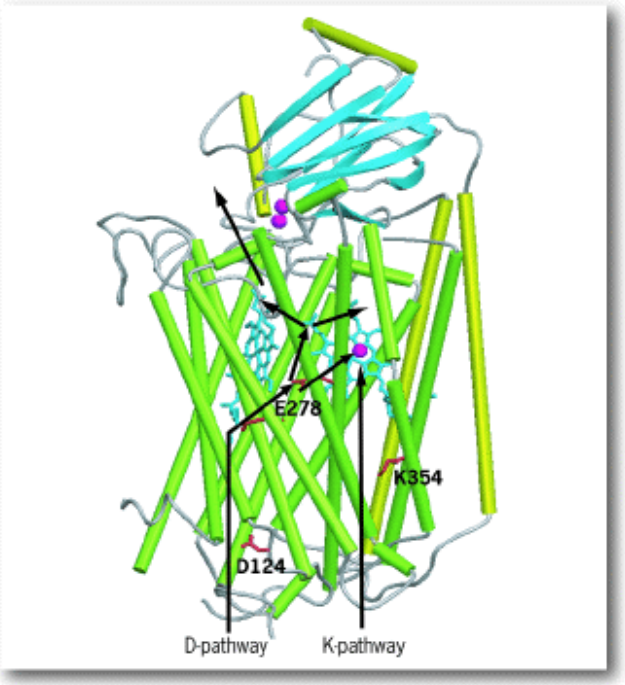


Thioredoxin:
red atoms belong to hydrophobic side chains

water

Membrane proteins are a separate case





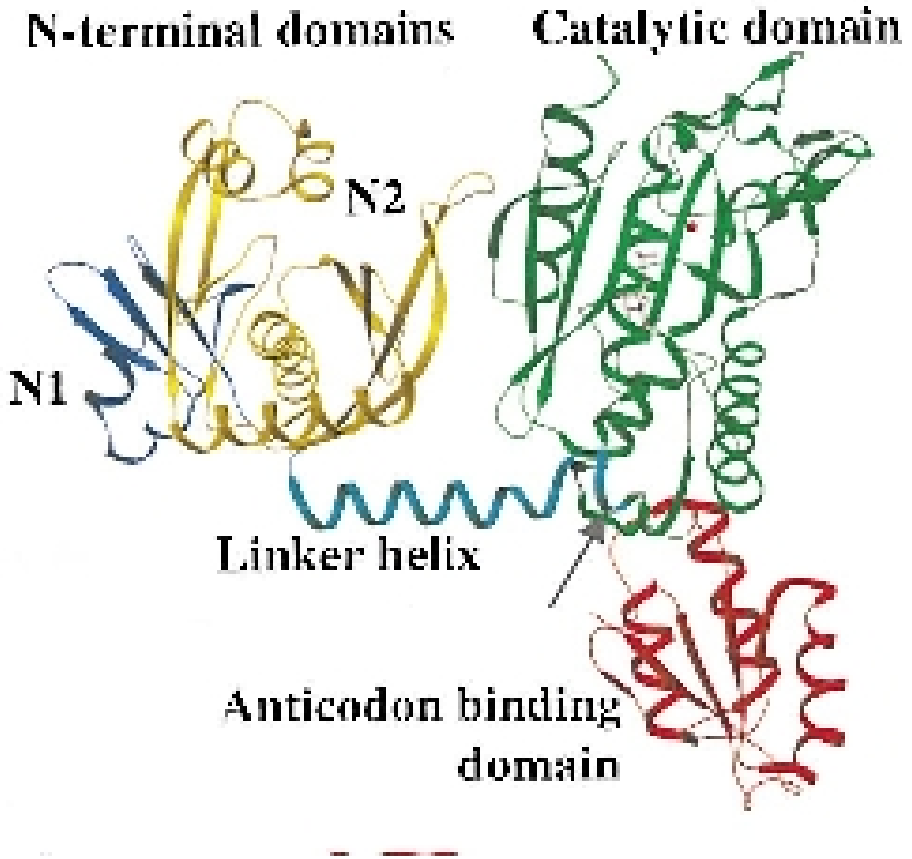
Cytochrome oxidase

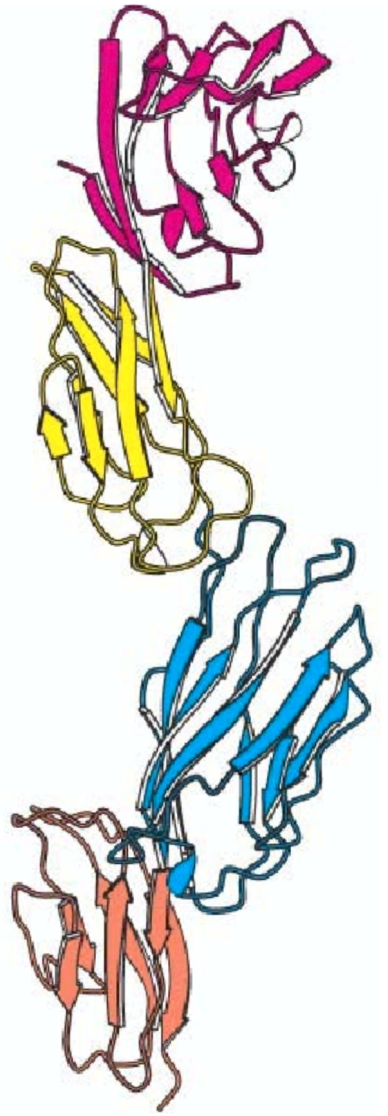
~ 30% of our genome humain, ~ 50% of drugs

Few known structures (~100).

Few architectures.

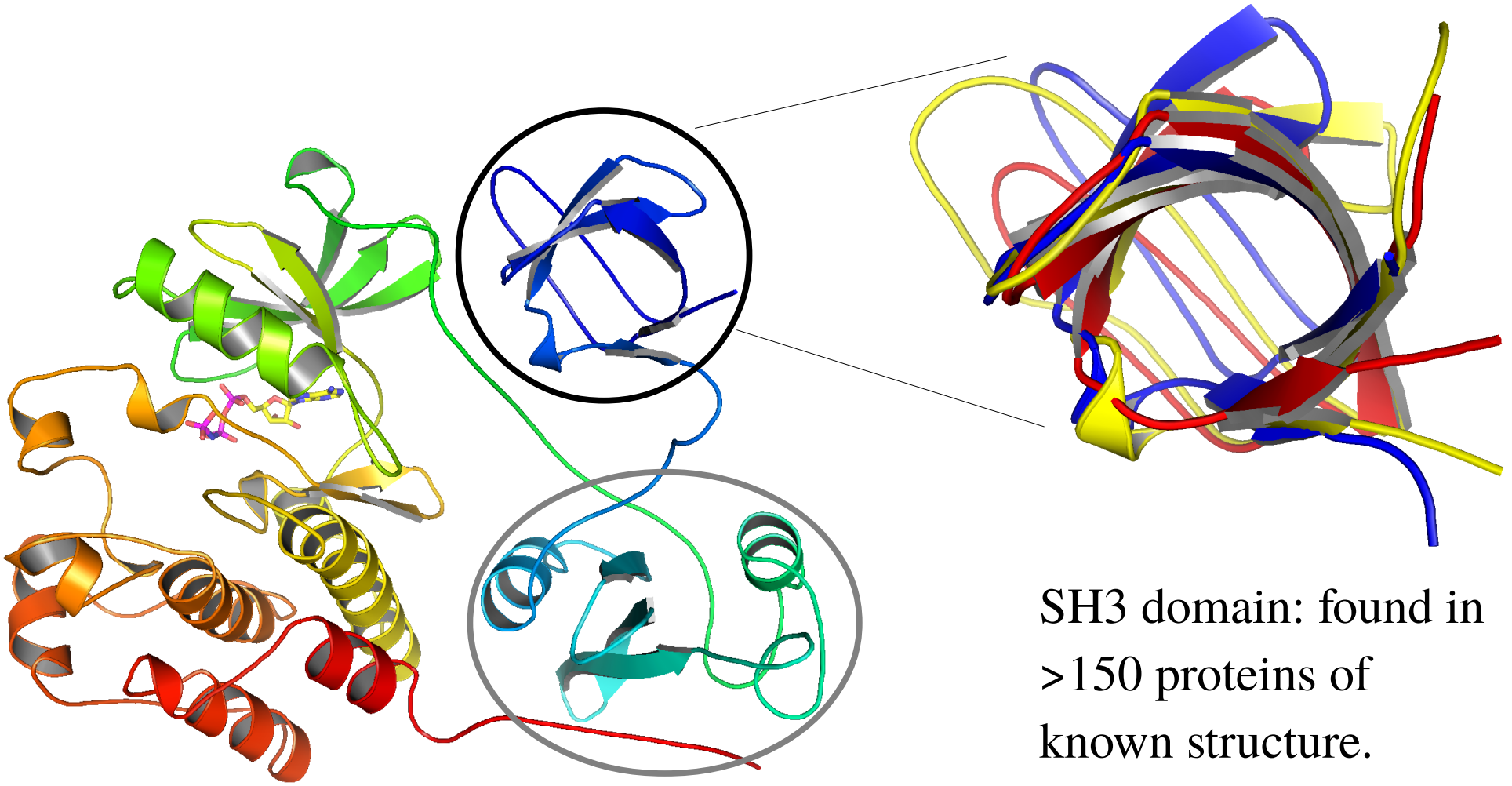
Quaternary structure is a 3D arrangement of structural domains





100-300 amino acids

One domain can appear in many proteins

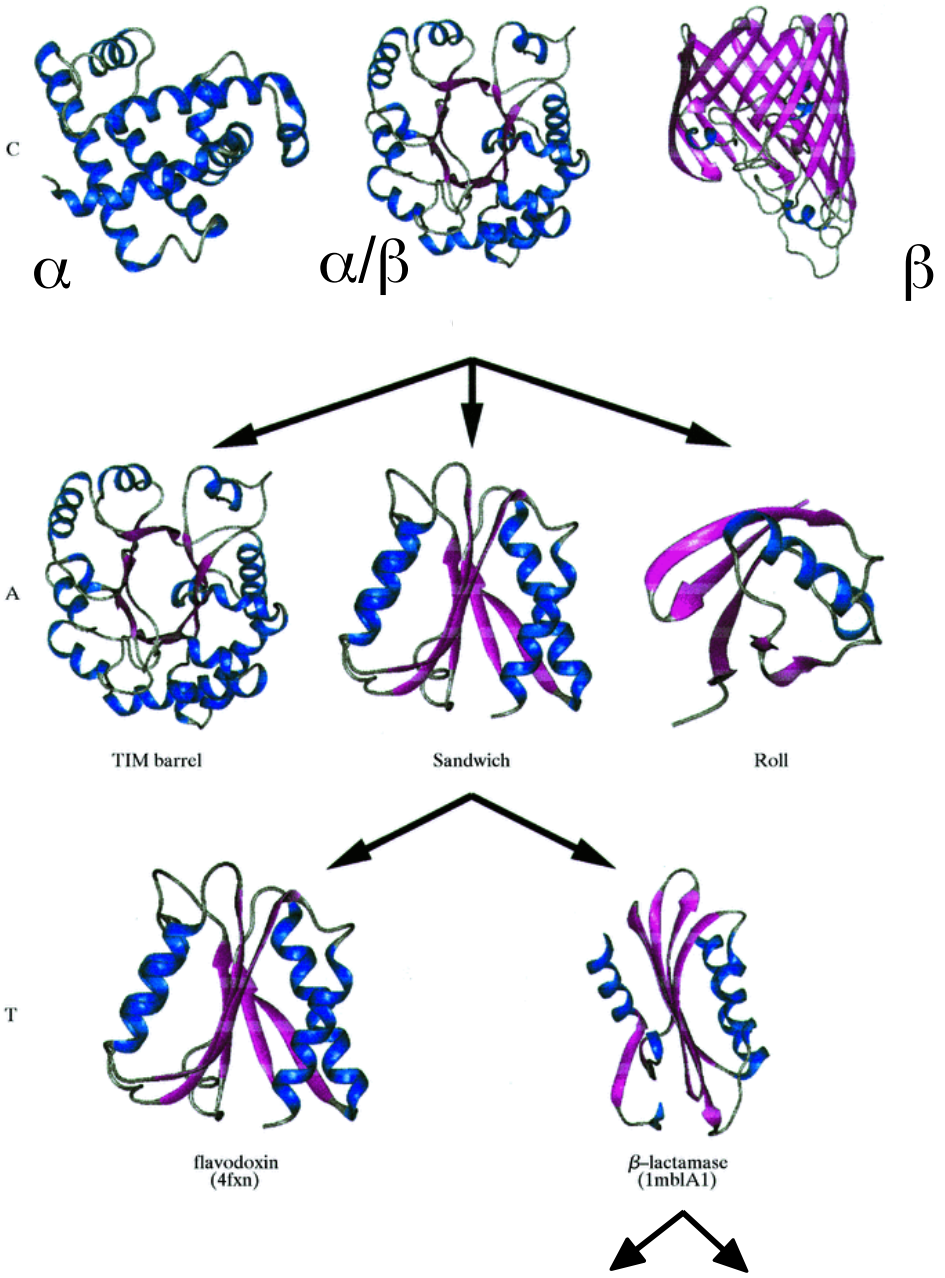


SH3 domain: found in >150 proteins of known structure.

Tyrosine kinase c-Src

23 in yeast; do not cross-react

Structural domains: hierarchical classification



3 classes

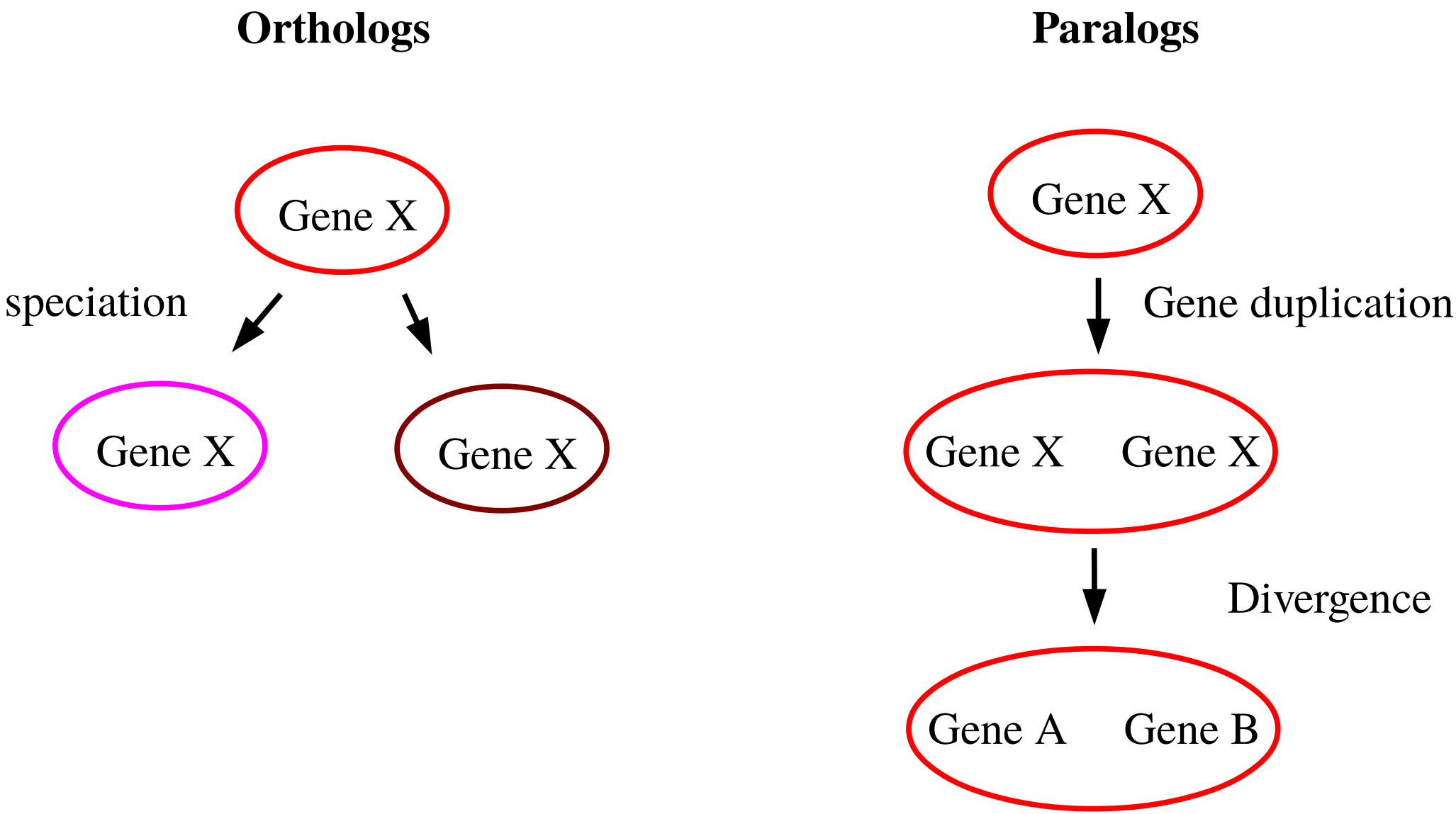
40 architectures

~ 1100 “topologies” or “folds”

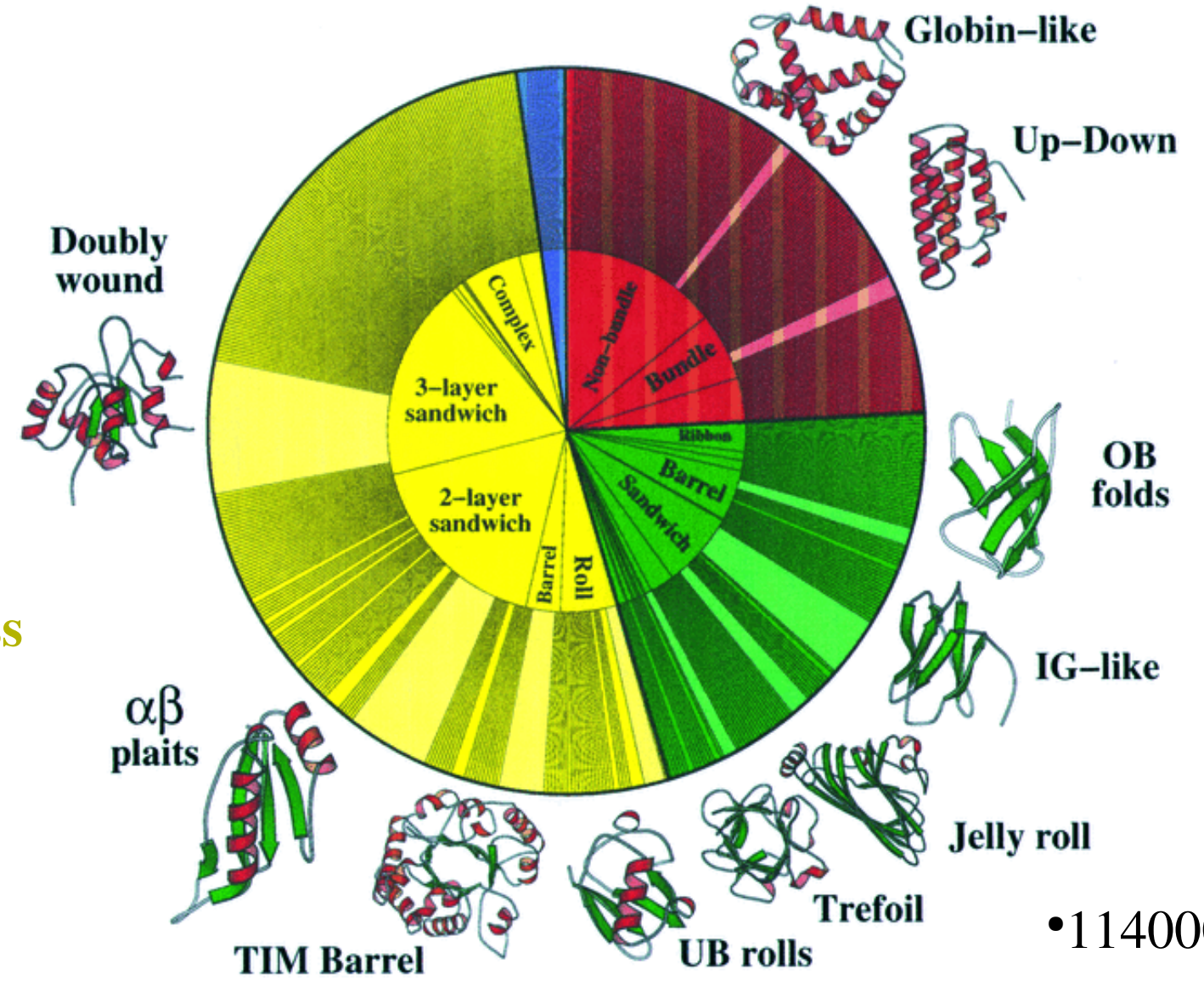
~ 2200 superfamilies

Classifying a Protein in the CATH Database of Domain Structures
Acta Cryst (1998) D54:1155
Orengo, Martin, Hutchinson, Jones, Jones, Michie, Swindells, Thornton

Similarities reflect evolution



Structural classification



α class

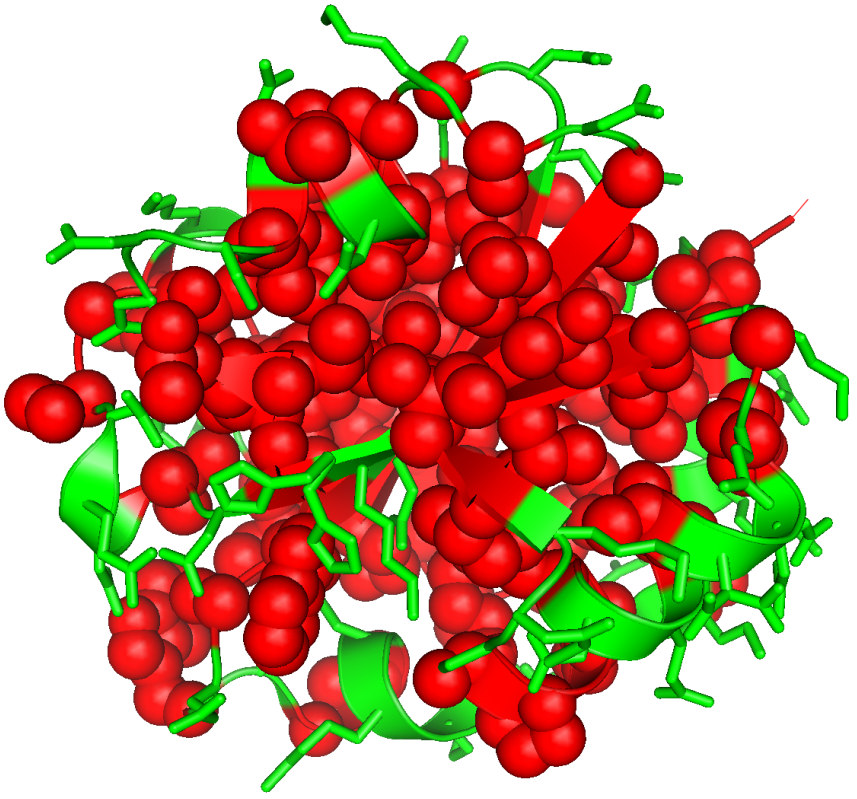
α/β class

β class

- 114000 structures
- **2200 superfamilies**
- 100 contain 60% of all domains

www.cathdb.info

Protein stability



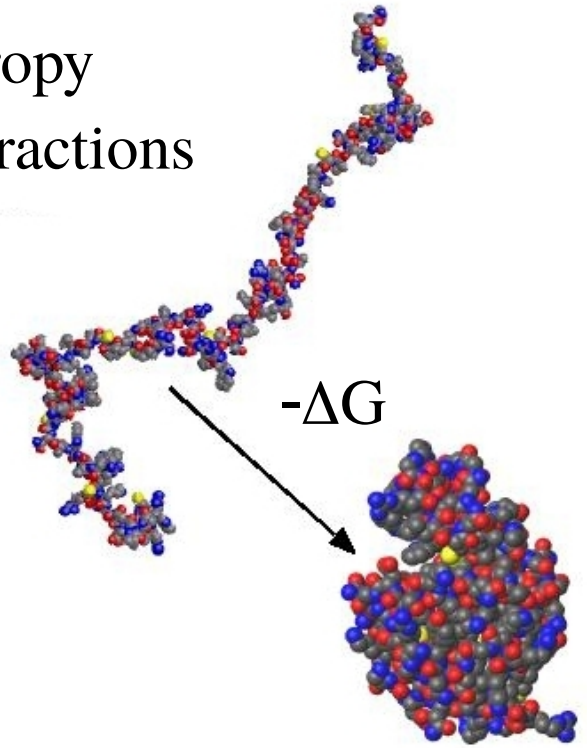
Proteins are marginally stable

ΔG = 5-15 kcal/mol = 10-30 kT

~ 0.1 kT per amino acid

Large conformational entropy

Many protein-solvent interactions



Cancelling effects

Low conformational entropy

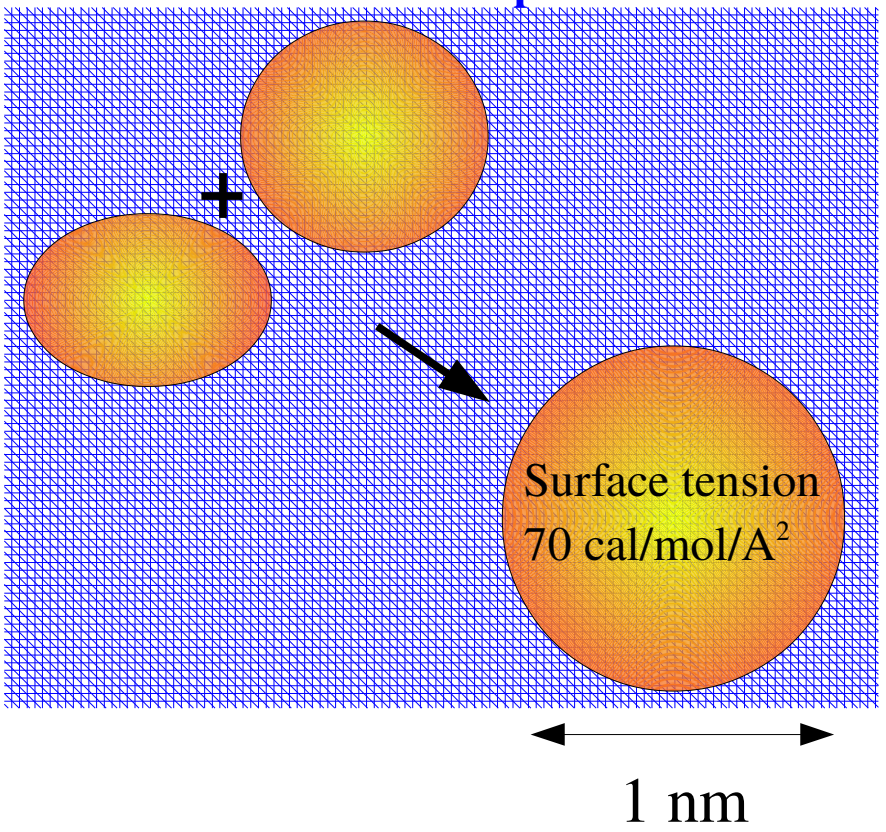
Fewer protein-solvent interactions

Protein-protein interactions

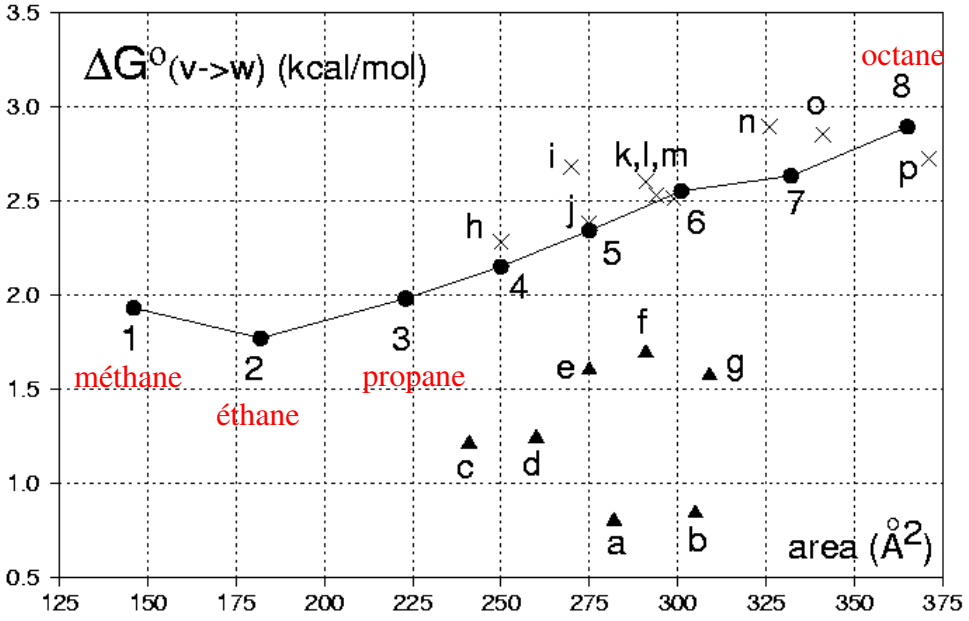
Folding is driven by the hydrophobic effect

Kauzmann, 1959

Alkane/water separation

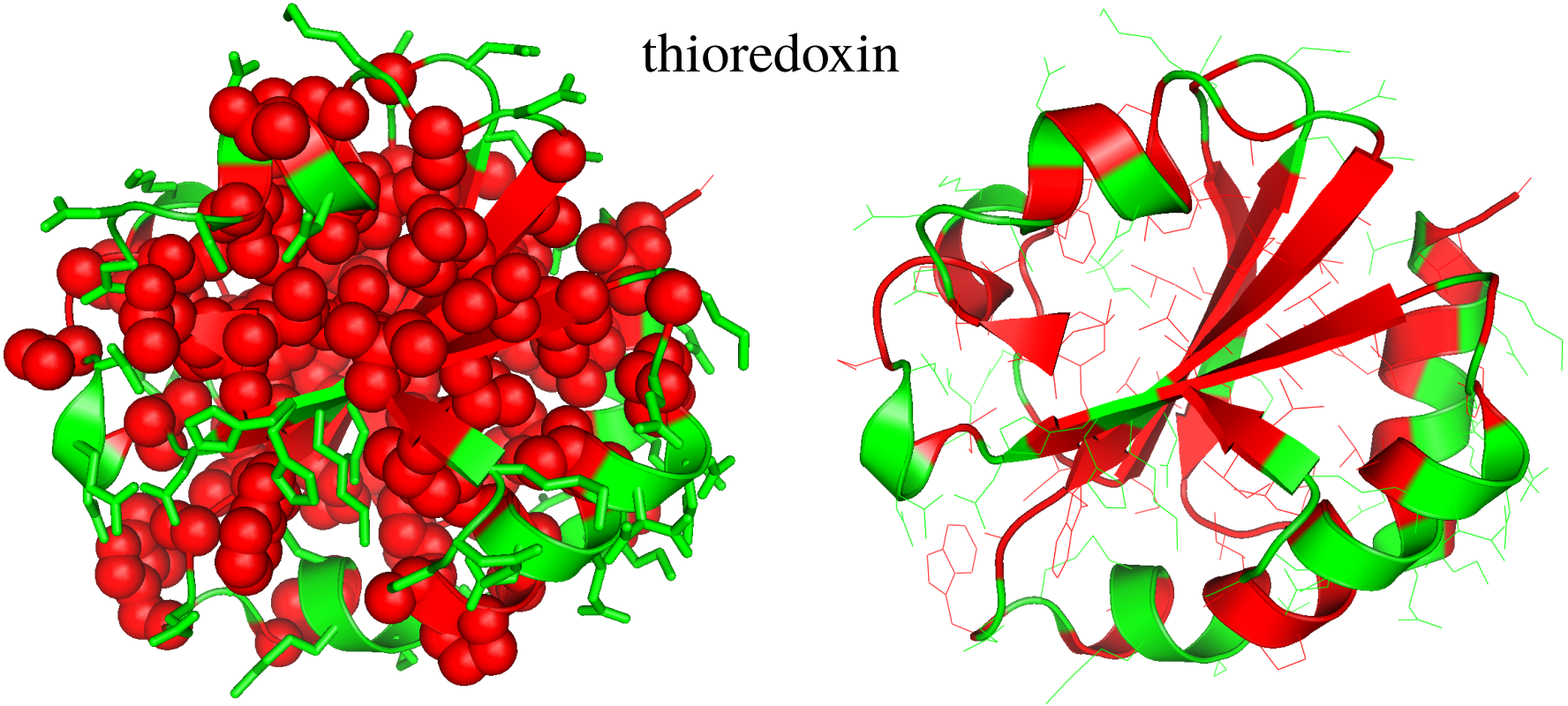


Vapor/water partition



Saturated alkanes don't like water

Proteins tend to bury apolar groupes, with very compact packing

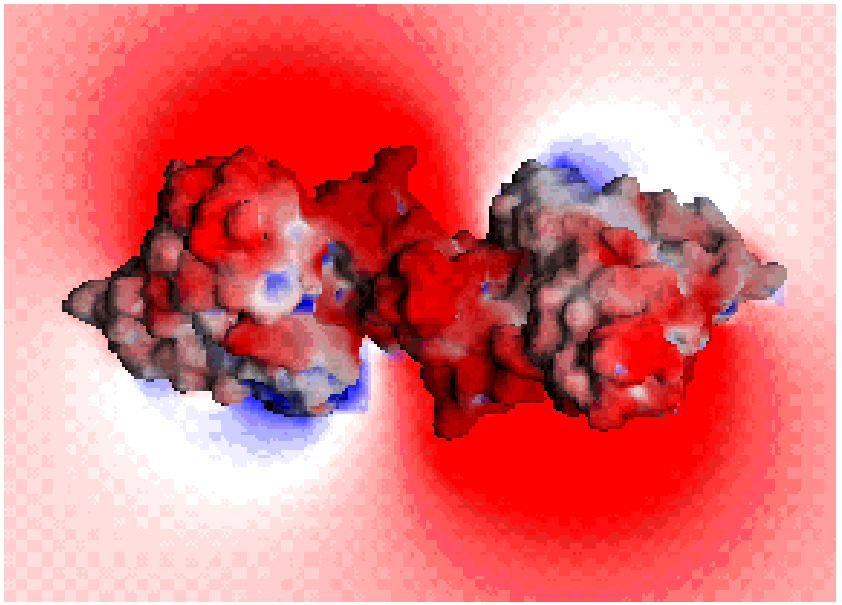


Fraction of occupied volume: ~ 0.74 (average over known structures)

Electrostatic interactions govern molecular recognition

25% of amino acids are ionic

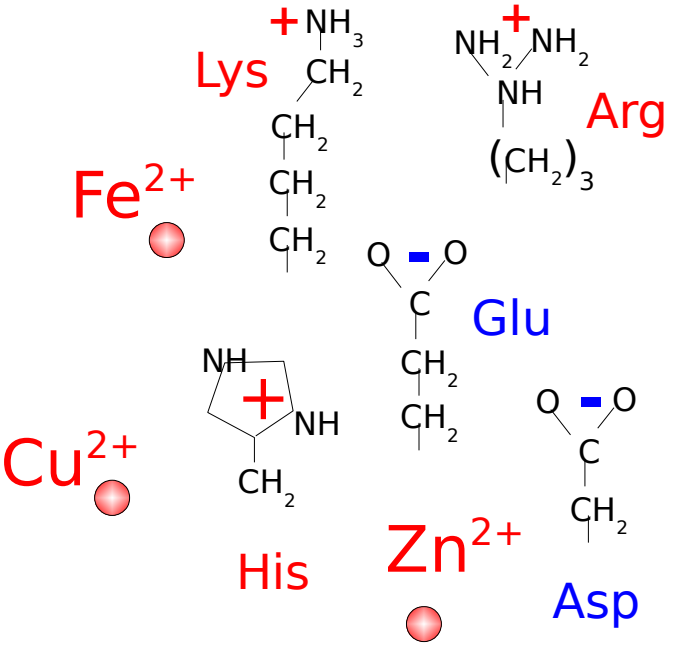
90% of salt bridges are at the surface



Acetylcholine esterase

Red = negative electrostatic potential

Substrate acetylcholine is positive

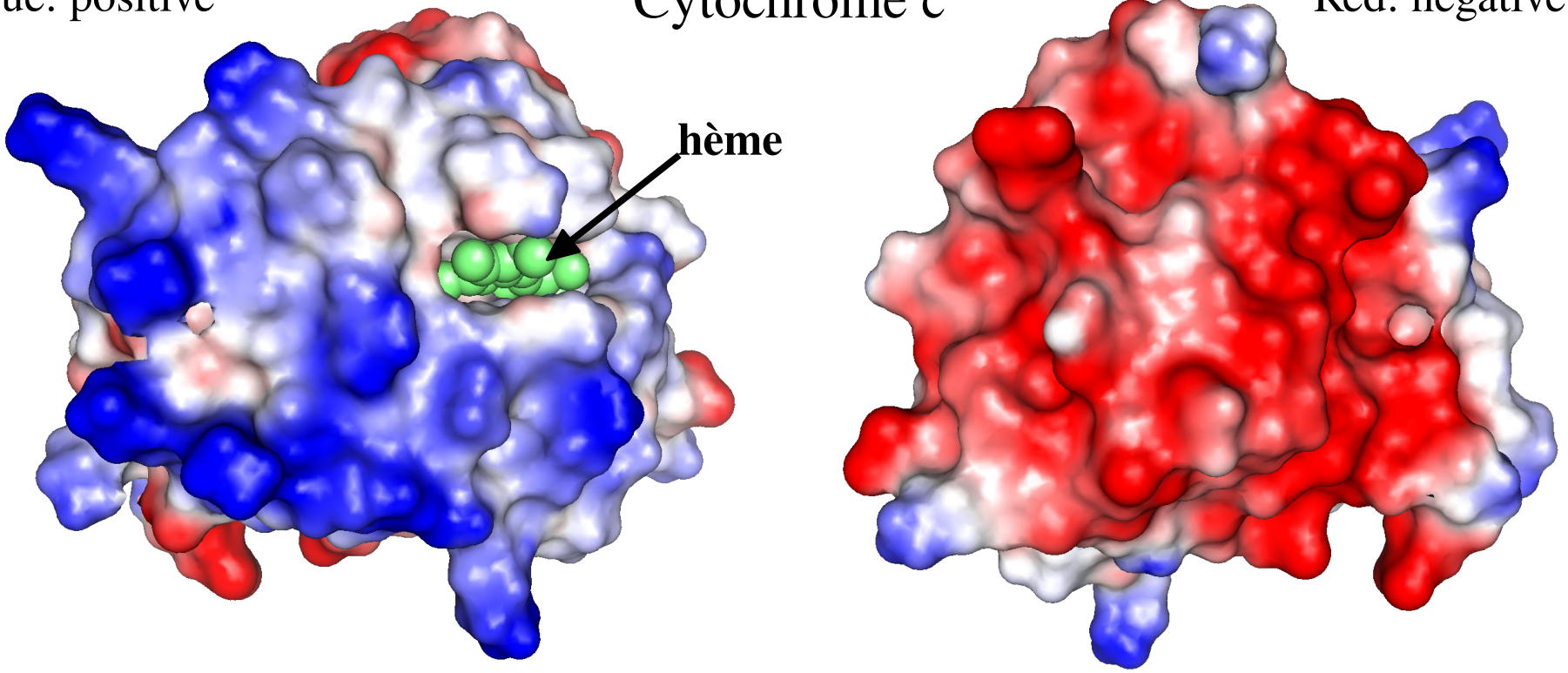


Electrostatic interactions govern molecular recognition

Blue: positive

Cytochrome c

Red: negative

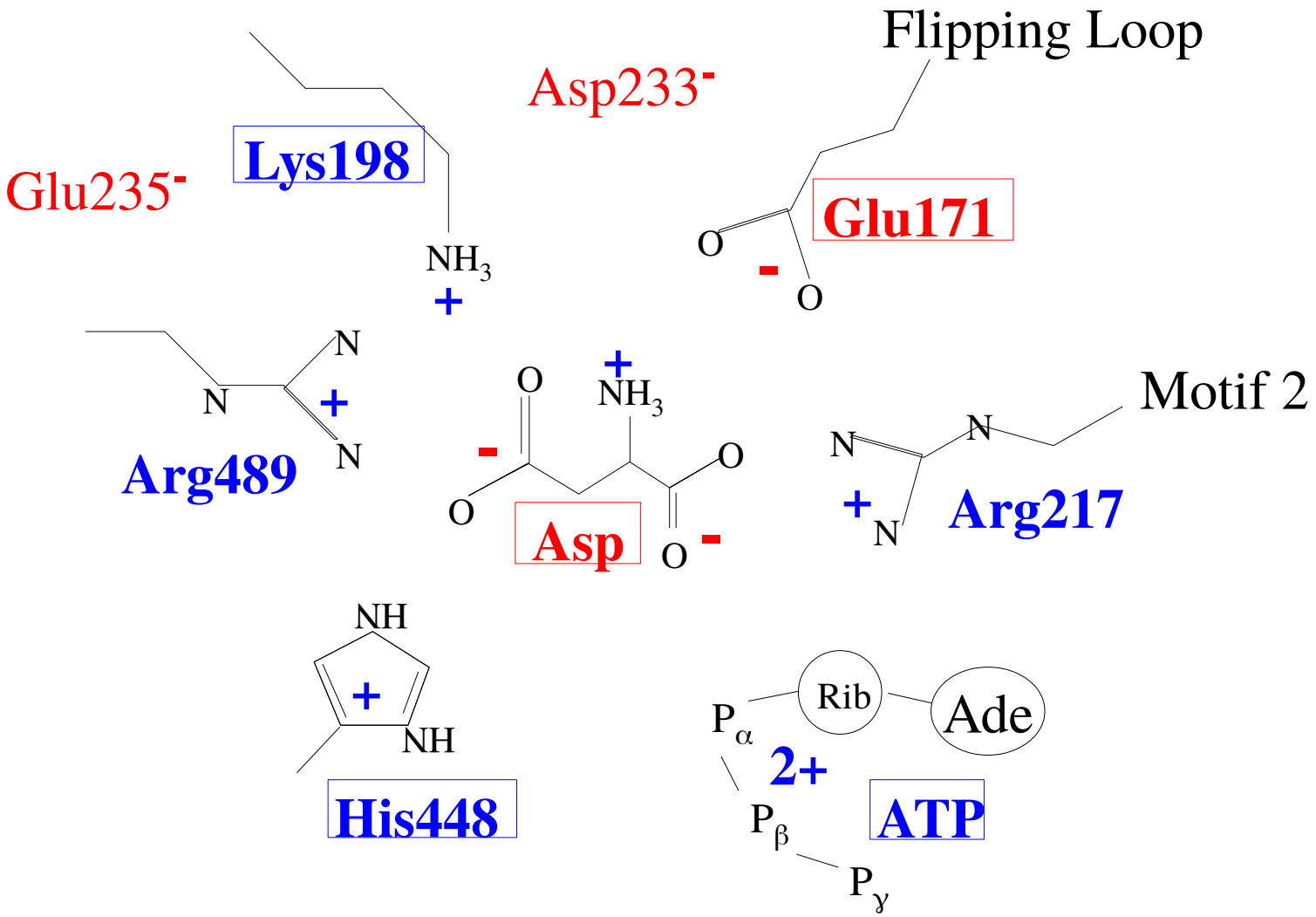


Front

Back

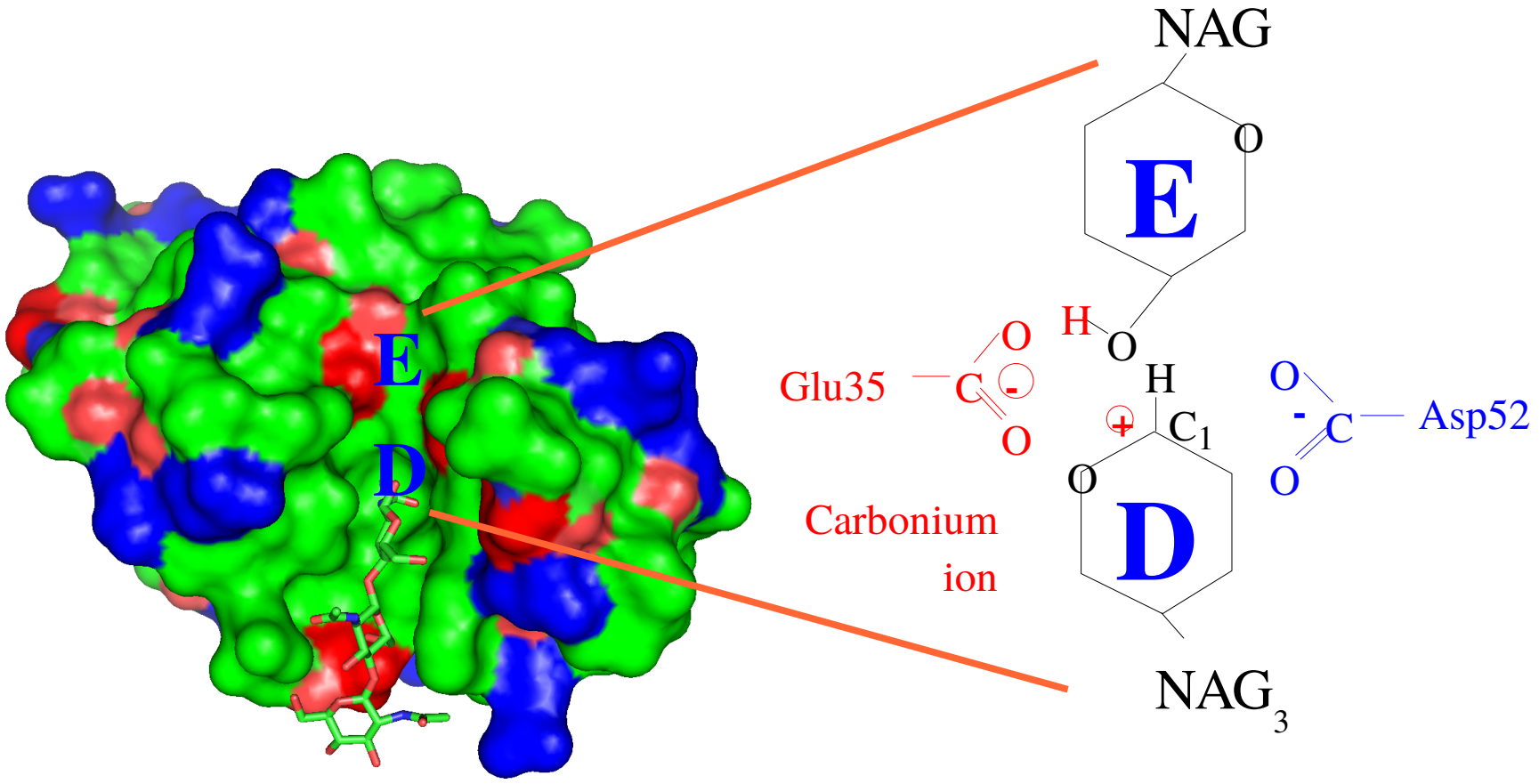
Cytochrome c interacts with the mitochondrial membrane, which is negative, and with negative regionson cytochrome bc1 and cytochrome oxidase.

Electrostatic interactions govern molecular recognition



Aspartyl-tRNA synthetase active site

Electrostatic interactions govern molecular recognition

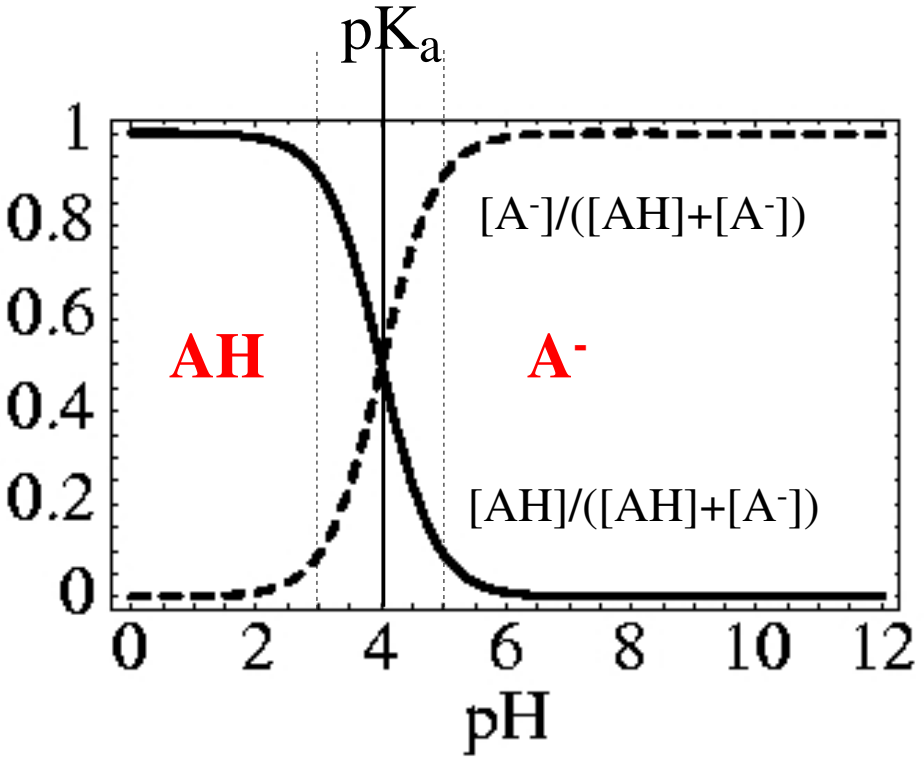


Lysozyme active site

Certain amino acids are ionized at neutral pH

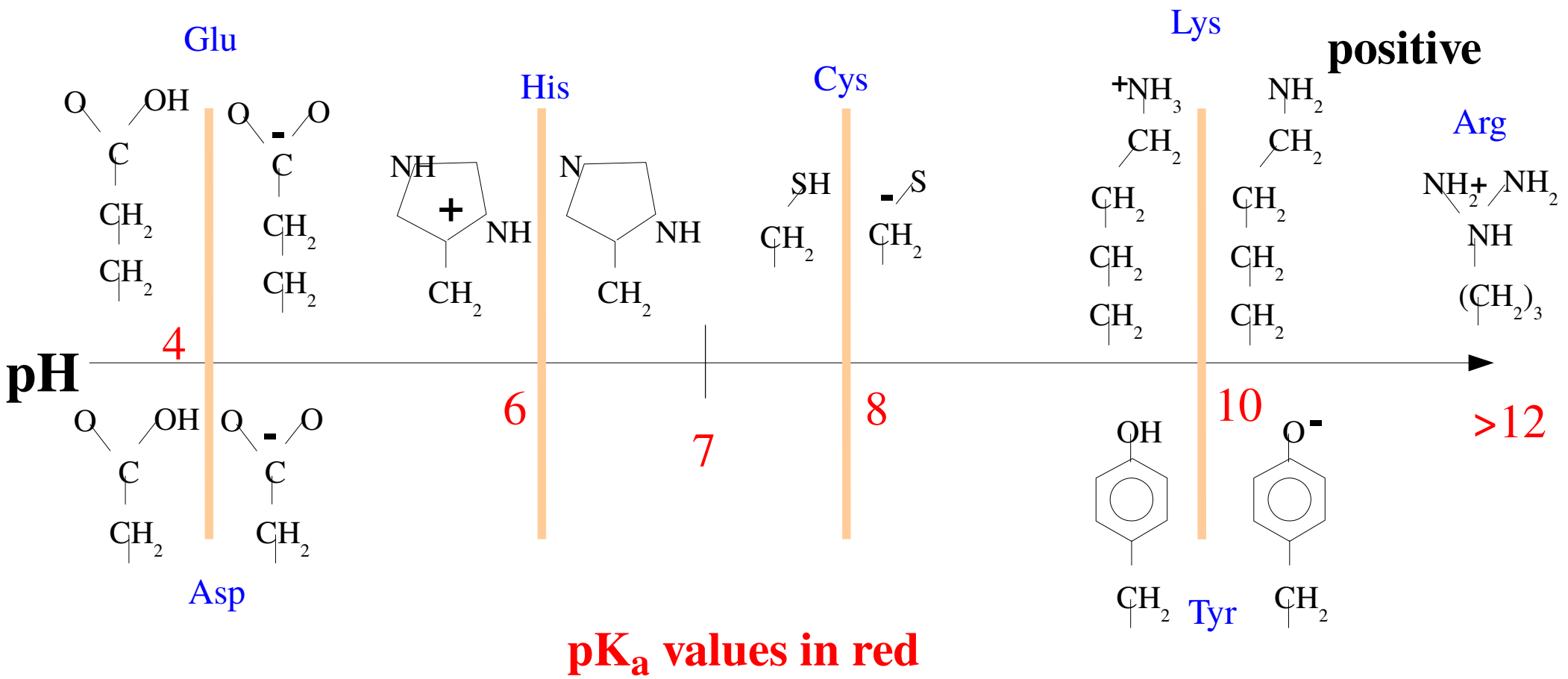
Acid/base equilibrium:

$$AH \rightarrow A^- + H^+$$



Protons abundant → **Protons rare**

Certain amino acids are ionized at neutral pH



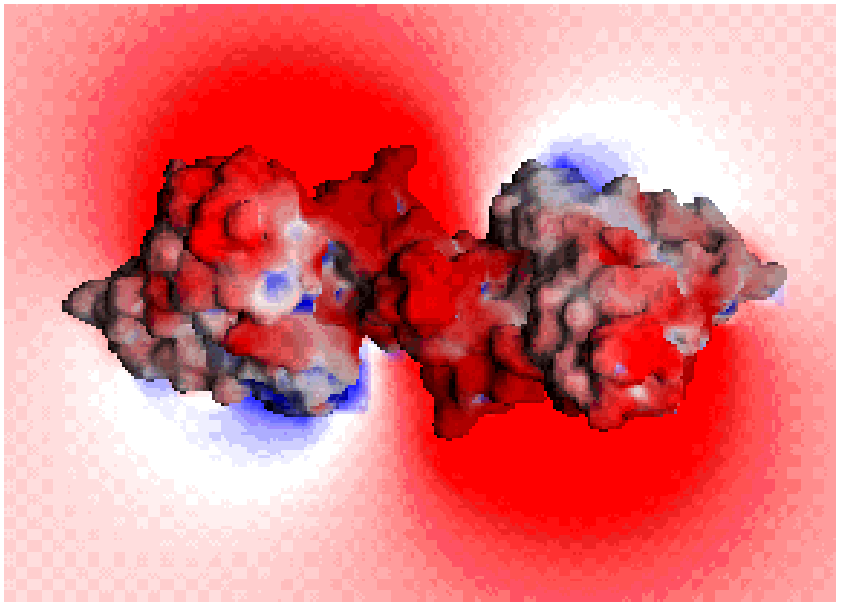
pK_a values in red

Protons abundant → **Protons rare**

Electrostatic interactions govern molecular recognition

25% of amino acids are ionic

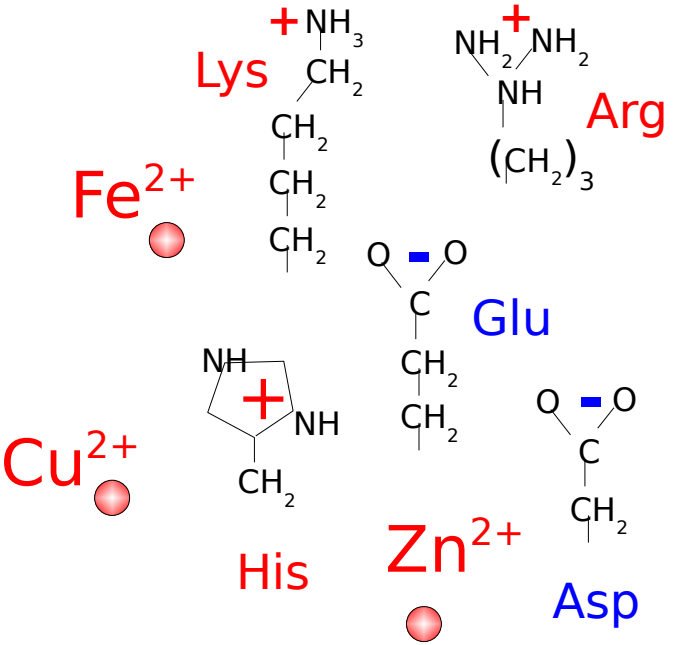
90% of salt bridges are at the surface



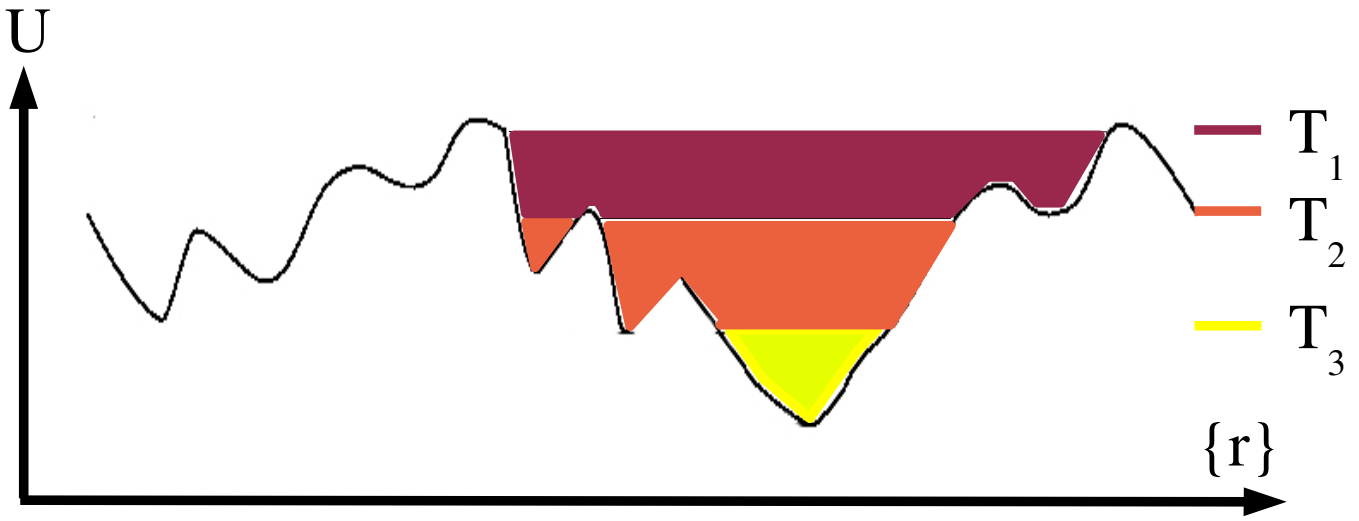
Acetylcholine esterase

Red = negative electrostatic potential

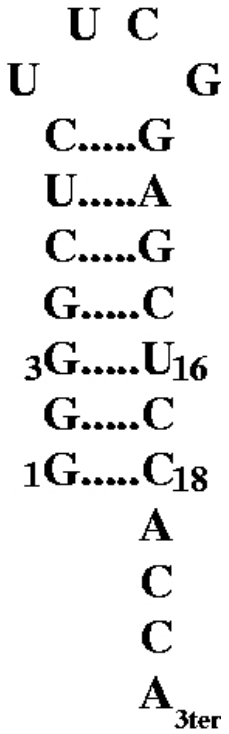
Substrate acetylcholine is positive

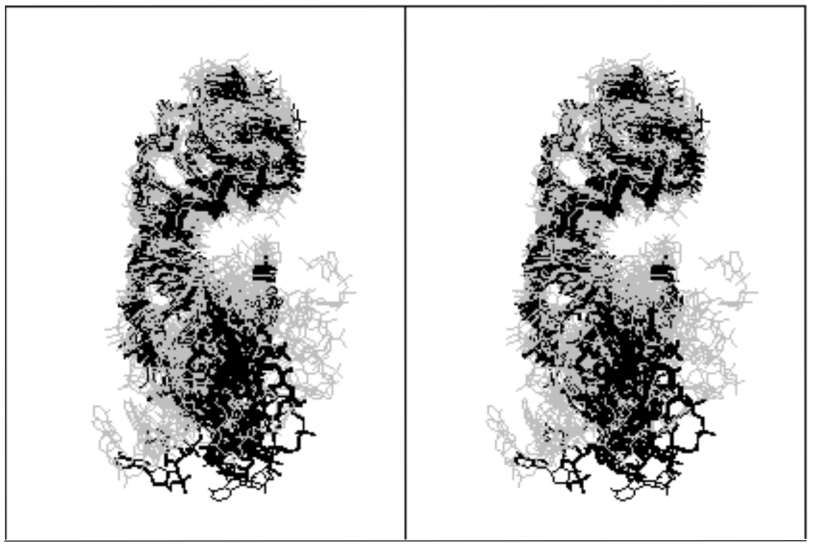


3D structure: molecular mechanics and molecular dynamics



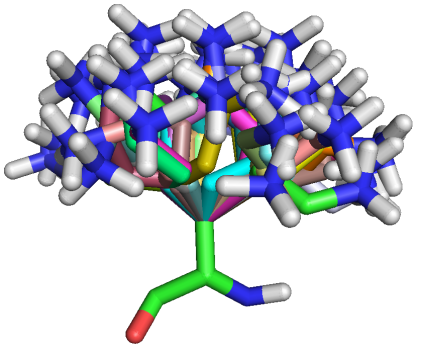
Biomolecular structures are flexible and mobile



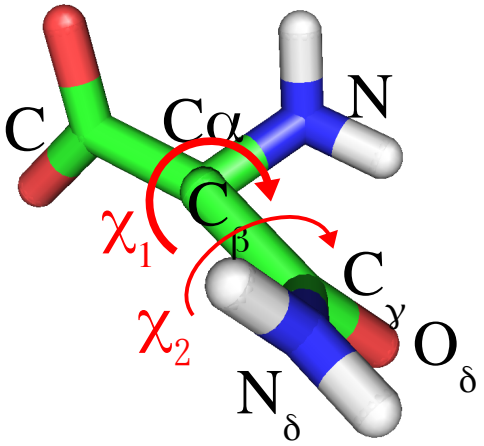


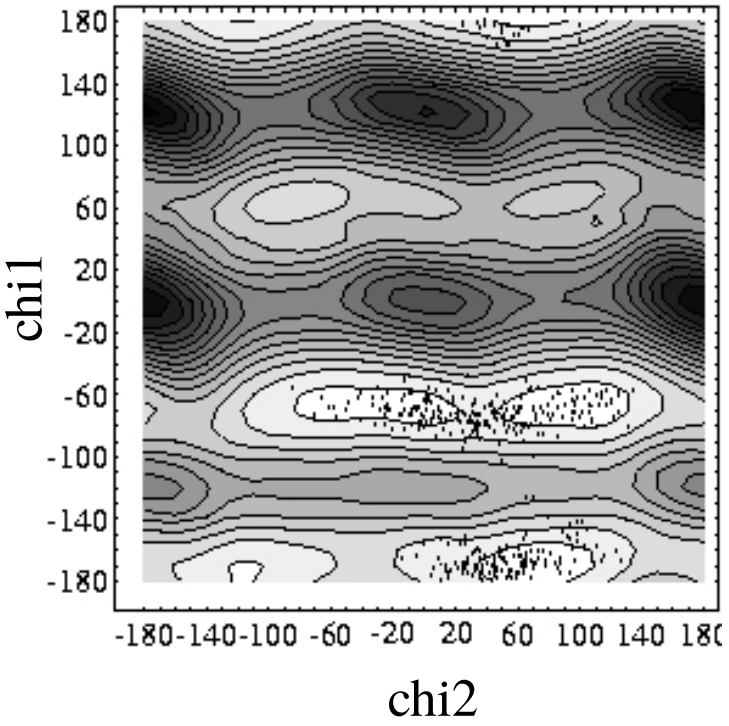
RNA stem-loop
(stereo)

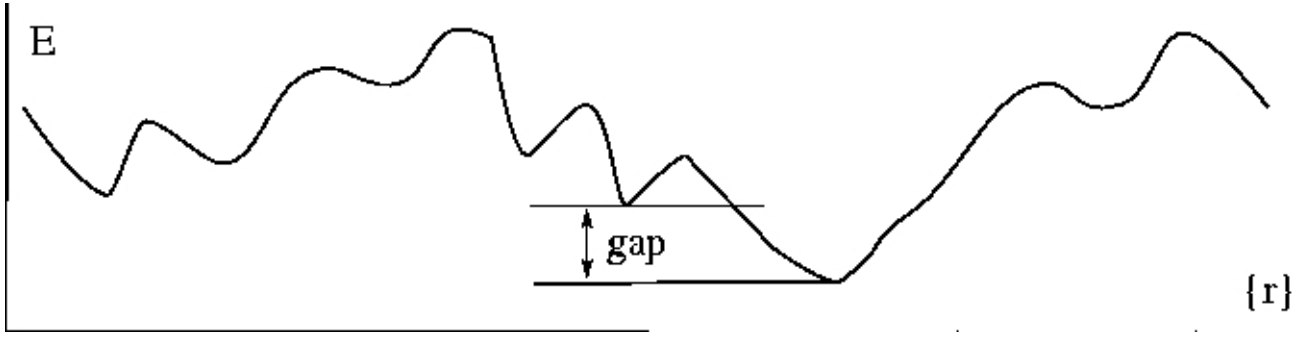
Lysine



We characterize them by an energy surface

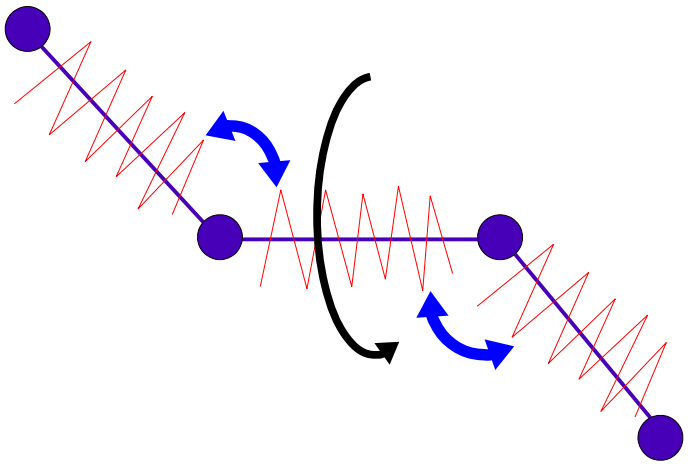






Structure prediction = find low energy structures

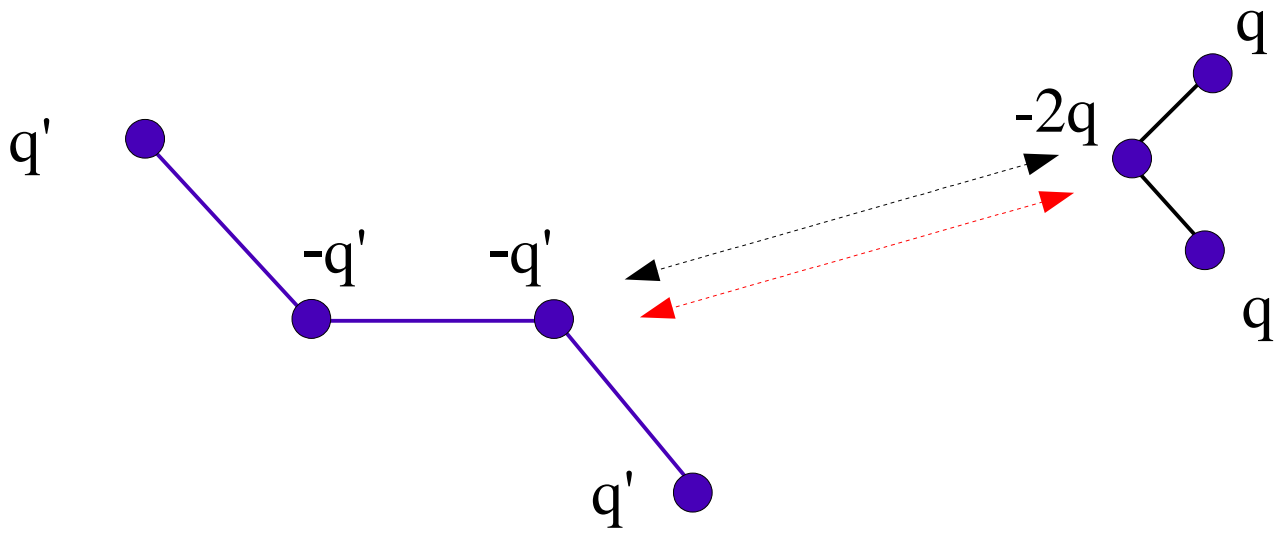
We introduce a semi-empirical energy function



"balls + springs"

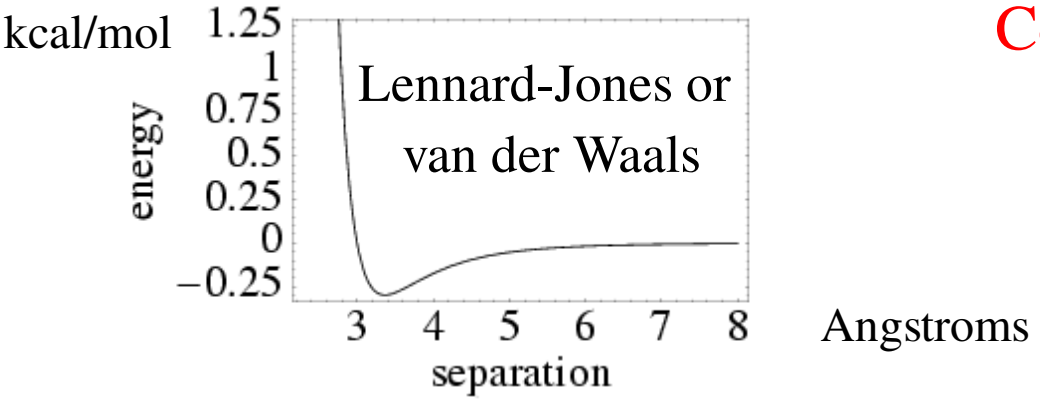
$$U = \Sigma_{bonds}\, k_b\, (b-b_0)^2 + \Sigma_{angles}\, k_a\, (a-a_0)^2 + \Sigma_{torsions}\, k_t\, [1 + \cos(nt-\tau)]$$

We introduce a semi-empirical energy function



$$U = \Sigma_{ij} [A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}] + \Sigma_{ij} q_i q_j / r_{ij}$$

Coulomb



Rôle of van der Waals interactions: Leu → Val mutations

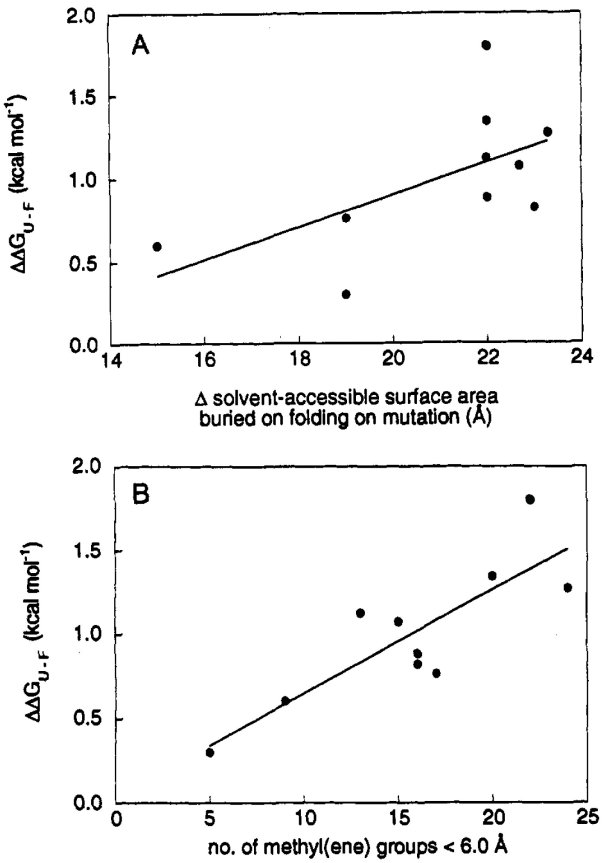
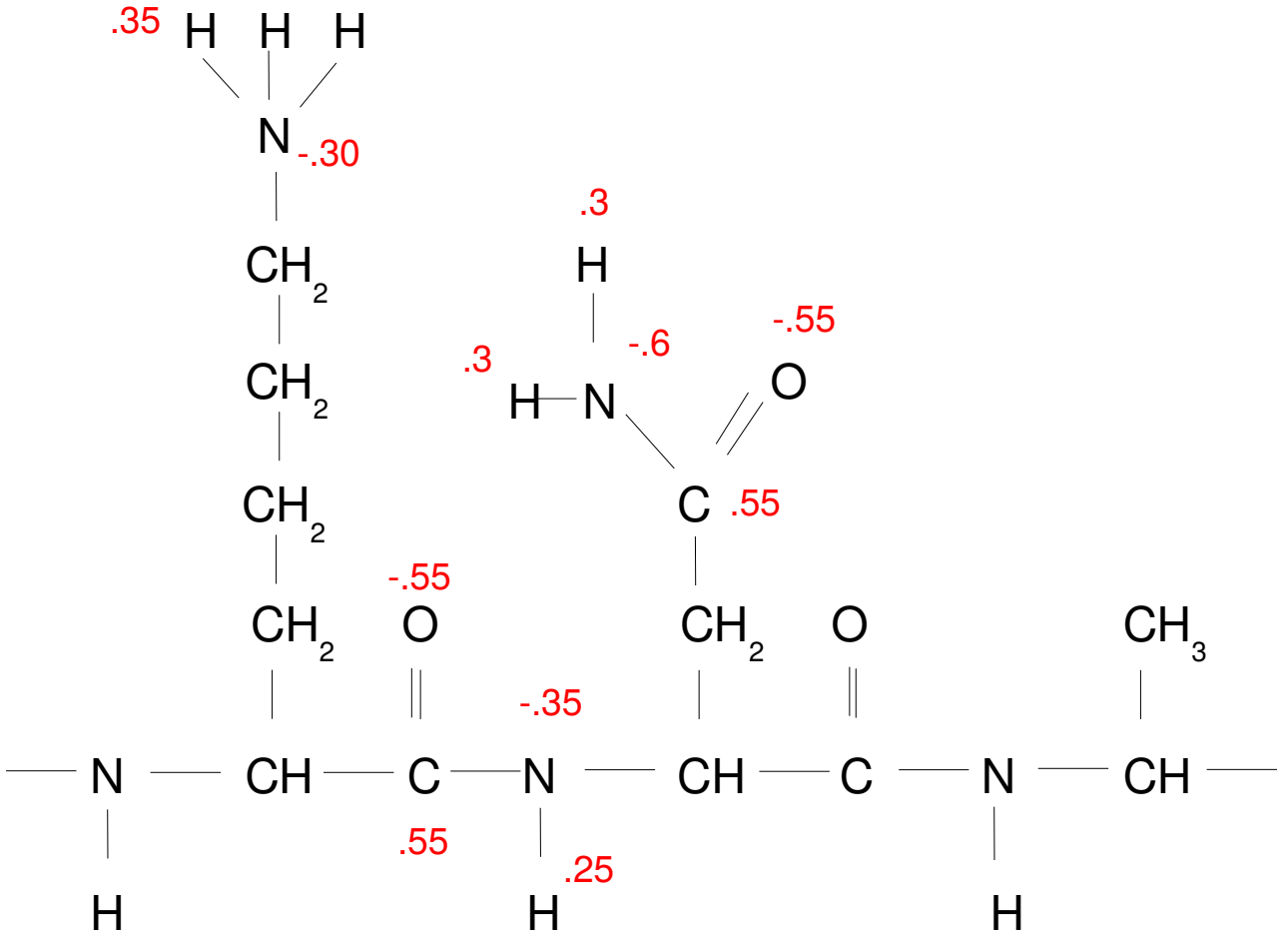


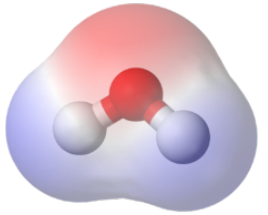
FIGURE 6: (A) Correlation between the difference in the solvent-accessible surface area that is buried on folding between wild-type and mutant side chains and the changes in the free energy of unfolding for Ile → Val mutations only for CI2 and barnase. The solid line shows the best fit of the data to a linear equation (correlation coefficient = 0.59). (B) Correlation between the number of side-chain methylene groups, in a radius of 6 Å of the group deleted from wild-type, and the changes in the free energy of unfolding for Ile → Val mutations only for CI2 and barnase. The solid line shows the best fit of combined CI2 and barnase data to a linear equation (correlation coefficient = 0.83).

Molecular mechanics energy

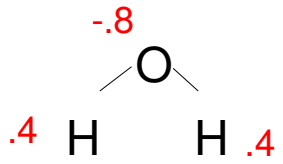
$$U = \Sigma_{bonds}\, k_b\,(b-b_0)^2 + \Sigma_{angles}\, k_a\,(a-a_0)^2 + \Sigma_{torsions}\, k_t\,[1+\cos(nt-\tau)]$$

$$+ \quad \Sigma_{ij}\,[\, A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6} + q_i q_j / r_{ij}\,]$$





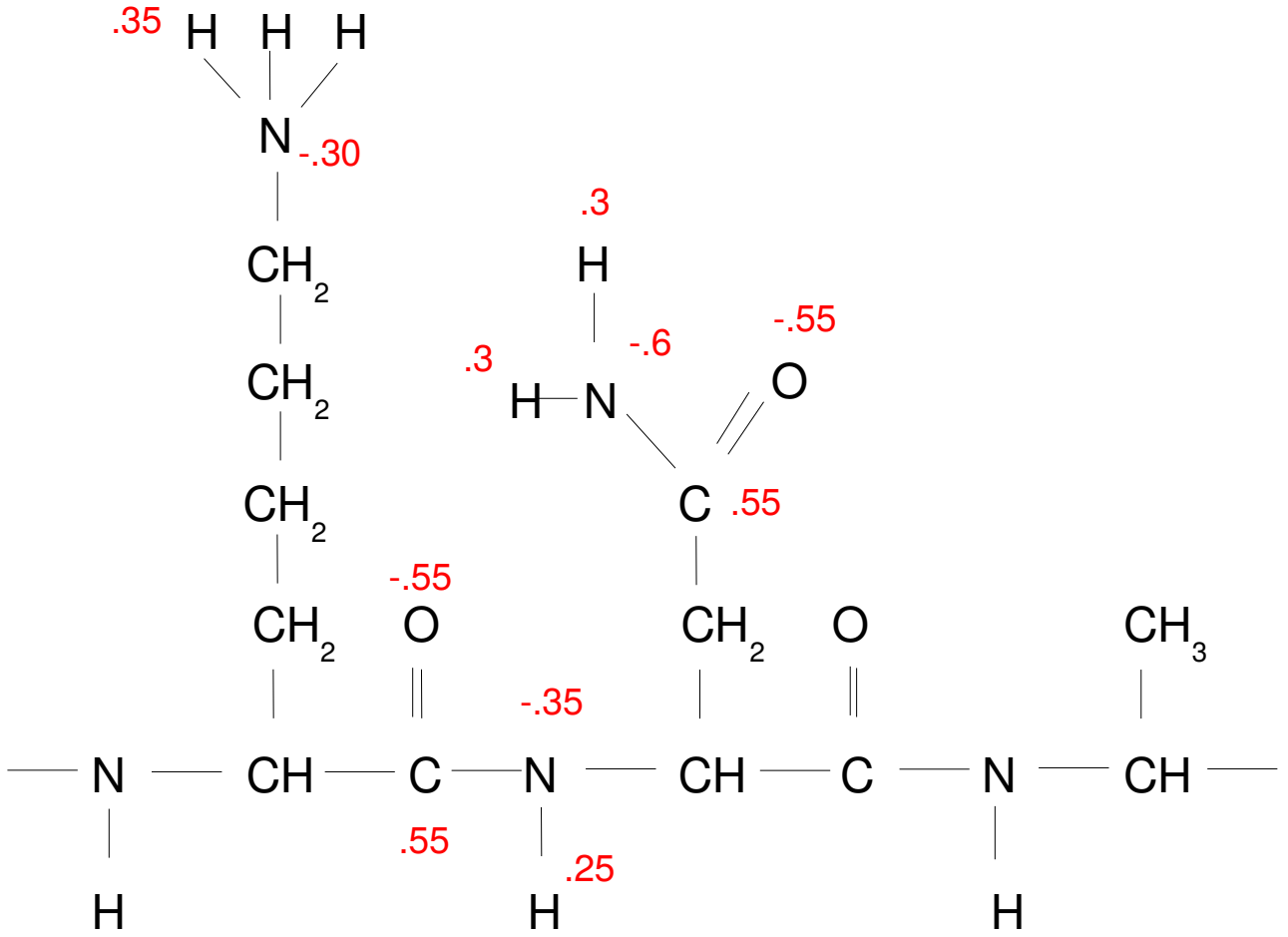
water



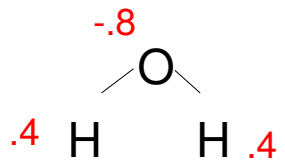
Molecular mechanics energy

$$U = \Sigma_{bonds}\, k_b\,(b-b_0)^2 + \Sigma_{angles}\, k_a\,(a-a_0)^2 + \Sigma_{torsions}\, k_t\,[1+\cos(nt-\tau)]$$

$$+ \quad \Sigma_{ij}\,[\, A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6} + q_i q_j / r_{ij}\,]$$



water



Notion of parameter *transferability*

Fichier de “topologie”, extrait

```
RESIdue ALA
 GROUp
  ATOM N    TYPE=NH1   CHARge=-0.35   END
  ATOM H    TYPE=H     CHARge= 0.25   END
  ATOM CA   TYPE=CH1E  CHARge= 0.10   END
 GROUp
  ATOM CB   TYPE=CH3E  CHARge= 0.00   END
 GROUp
  ATOM C    TYPE=C     CHARge= 0.55   END
  ATOM O    TYPE=O     CHARge=-0.55   END

 BOND N    CA
 BOND CA   C
 BOND C    O
 BOND N    H
 BOND CA   CB

 IMPRoper  CA     N    C  CB  !tetrahedral CA

END {ALA}
```

{Fichier toph19.pro}

Fichier des paramètres d'énergie, extrait

```
bond C    C      450.0  1.38!  B. R. GELIN THESIS AMIDE AND DIPEPTIDES
bond C    CH1E   405.0  1.52!  EXCEPT WHERE NOTED.  CH1E,CH2E,CH3E, AND CT
bond C    CH2E   405.0  1.52!  ALL TREATED THE SAME. UREY BRADLEY TERMS ADDED
bond C    CH3E   405.0  1.52
bond C    CR1E   450.0  1.38

:
:

angle C    C    C       70.0 106.5!  FROM B. R. GELIN THESIS WITH HARMONIC
angle C    C    CH2E    65.0 126.5!  PART OF F TERMS INCORPORATED. ATOMS
angle C    C    CH3E    65.0 126.5!  WITH EXTENDED H COMPENSATED FOR LACK
angle C    C    CR1E    70.0 122.5!  OF H ANGLES.

:
:

NONBonded  H      0.0498   1.4254   0.0498  1.4254  0. 1.
NONBonded  HA     0.0450   2.6157   0.0450  2.6157  0. 1      ! charged group.

NONBonded  C      0.1200   3.7418   0.1000  3.3854  7.116  1. ! carbonyl carbon
NONBonded  CH1E   0.0486   4.2140   0.1000  3.3854  9.944  1. !
NONBonded  CH2E   0.1142   3.9823   0.1000  3.3854  17.626 1. ! extended carbons
```

{fichier param19.inp}

Molecular mechanics energy

- No explicit electrons
- Partial charges on atoms
- No chemical reactions
- big systems: >100.000 atoms
- Conformational energies
- Explicit or implicit solvent representation
- explicit force calculation
- Molecular dynamics

Parametrization: 1970-present

CHARMM M. Karplus, A. Mackerell and coll. (Harvard)
AMBER P. Kollman, D. Case and coll. (UCSF, Scripps)
OPLS W. Jorgensen and coll. (Yale)
GROMOS H. Berendsen, W. van Gunsteren and coll. (Groningen, Zürich)
Plus récent: AMOEBA Jay Ponder (St Louis)

Atomic charges : crystals of small molecules
simple liquids
quantum chemistry calculations

Lennard-Jones: crystals of small molecules
simple liquids

bonds, angles, torsions: quantum chemistry calculations
small molecule spectroscopy

Software

CHARMM19 (M Karplus et al, Harvard)

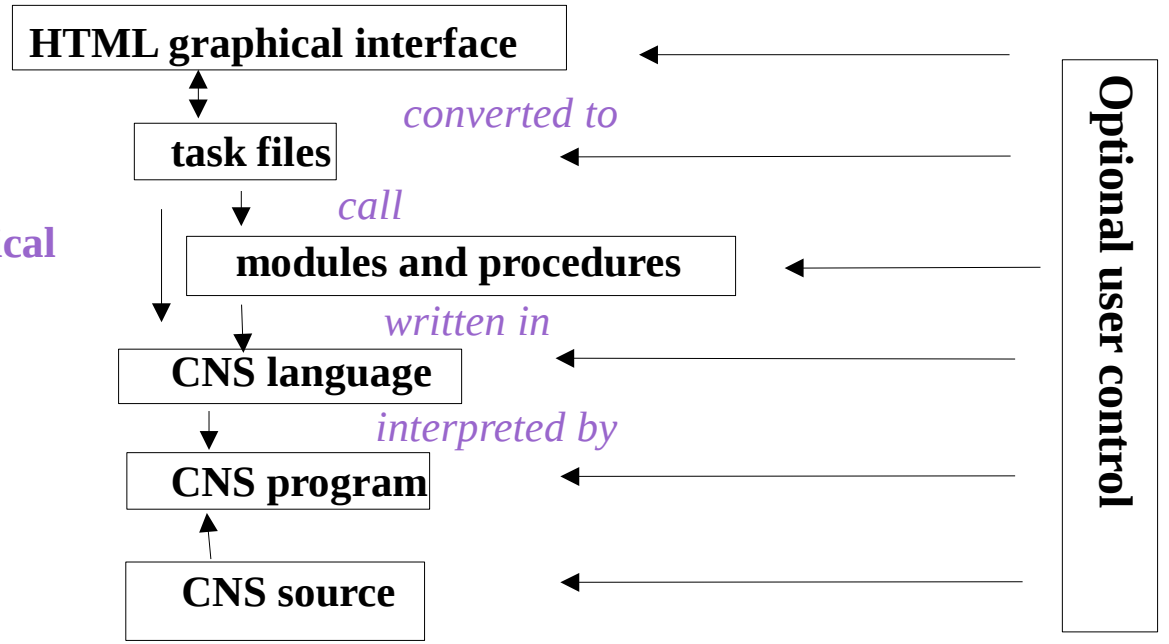
XPLOR (A Brunger, Yale)

NIH-XPLOR
(M Clore,
C Schweiters,
J Kuszewski)

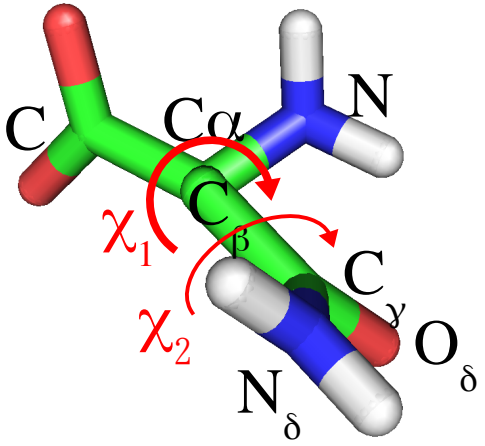
CNS ('Crystallography and NMR System') ~220,000 lines

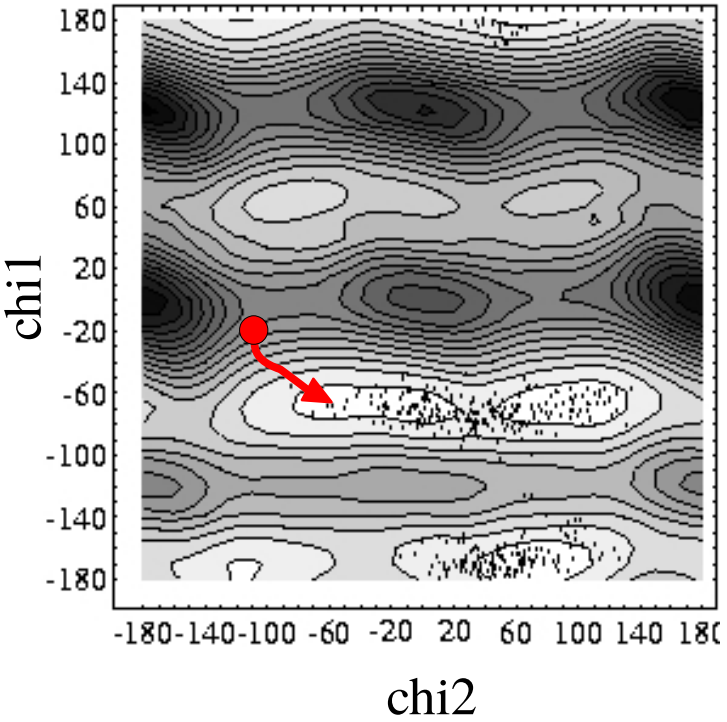
A Brunger, P Adams, G Clore, W Delano, P Gros, R Grosse-Kunstleve, J Jiang, J Kuszewski, M Nilges, N Pannu, R Read, L Rice, T Simonson, G Warren (1998) Acta Cryst D54, 905.

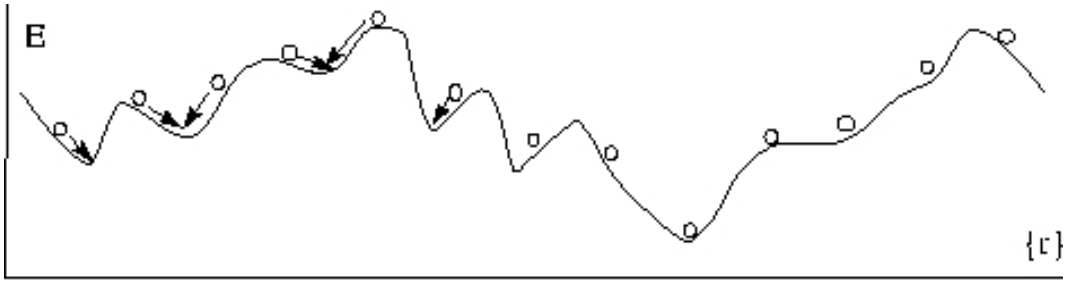
Hierarchical stucture:



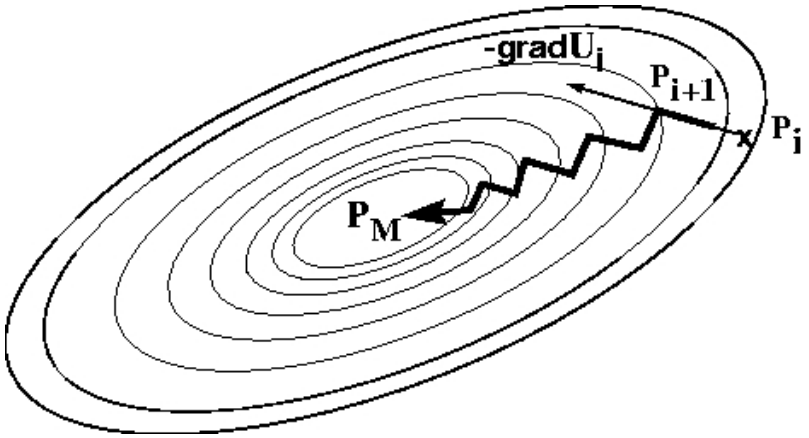
Finding low energy structures: energy minimization







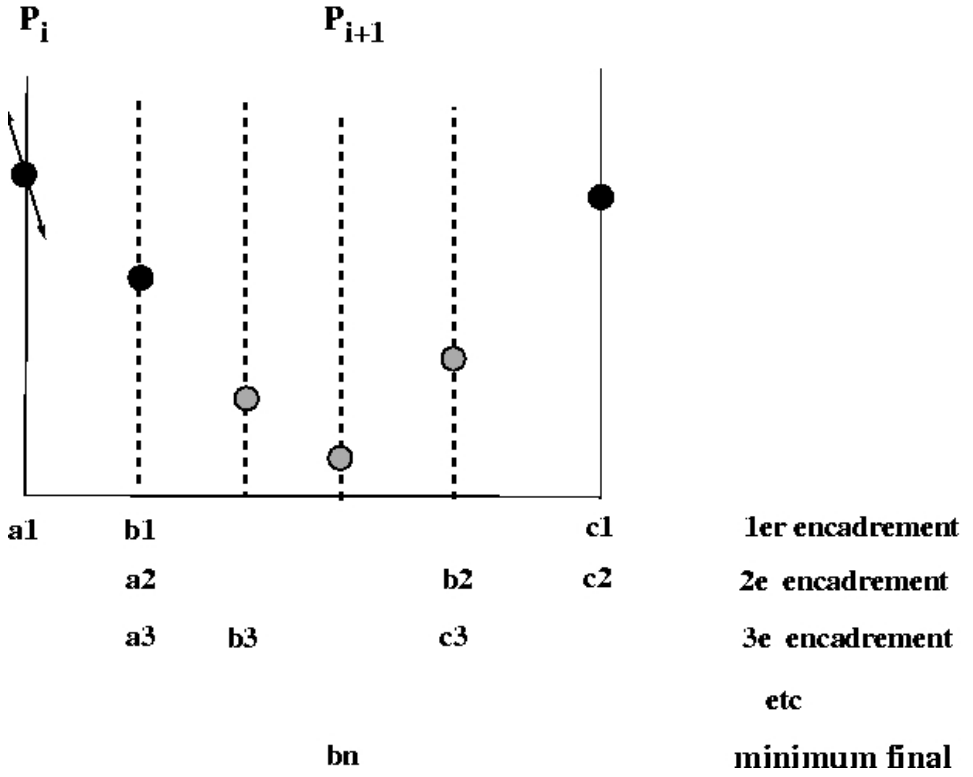
Steepest descent minimization



(1) Method: starting from P_i, move along the energy gradient: -grad $U(P_i)$ until a minimum, P_{i+1}. Repeat.

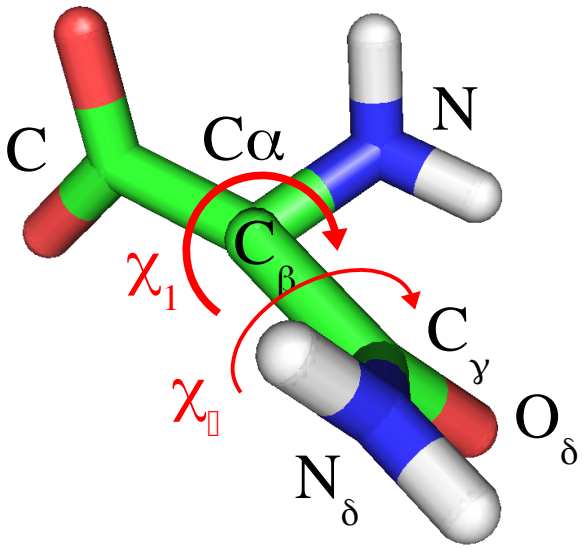
(2) Finding P_{i+1}:

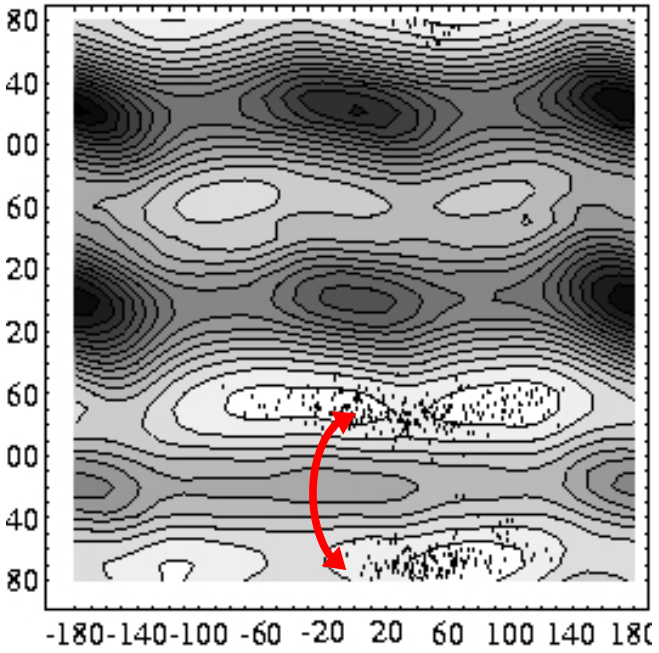
successive interpolations



Cf Numerical Recipes; Press et al.

Explorating conformations with molecular dynamics





chi2

Can cross energy barriers (not too high)

Solve equations of motion numerically:

$$m_i \gamma_i = F_i = - \text{grad } U_i$$

3 equations per atom

Verlet algorithm

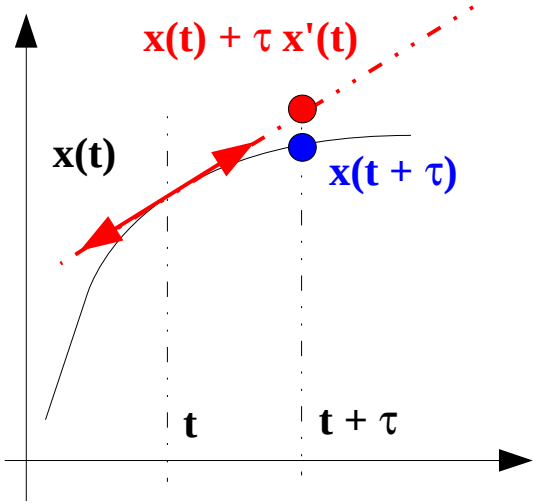
Consider each coordinate of each atom i as a function of time t

The change between two instants t and t+τ can be approximated:

$$x_i(t+\tau) \cong x_i(t) + \tau\, x'_i(t)$$

$$y_i(t+\tau) \cong \ldots$$

$$z_i(t+\tau) \cong \ldots$$



Verlet algorithm

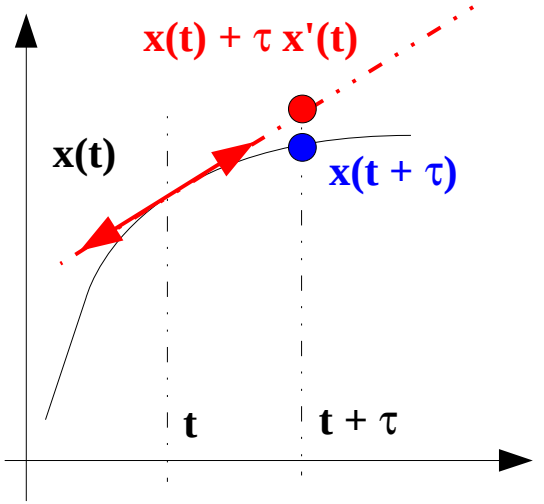
Consider each coordinate of each atom i as a function of time t

The change between two instants t and t+τ can be approximated:

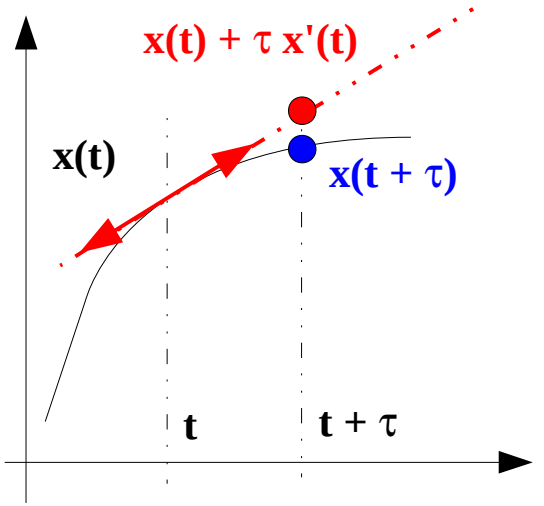
$$x_i(t+\tau) \cong x_i(t) + \tau\, x'_i(t) + (\tau^2/2)\, x''_i(t)$$

$$y_i(t+\tau) \cong \ldots$$

$$z_i(t+\tau) \cong \ldots$$



Verlet algorithm



On soustrait les valeurs pour t+τ et t-τ:

$$x_i(t+\tau) \cong x_i(t) + \tau\, x'_i(t) + (\tau^2/2)\, x''_i(t)$$

$$x_i(t-\tau) \cong x_i(t) - \tau\, x'_i(t) + (\tau^2/2)\, x''_i(t)$$

$$x_i(t+\tau) + x_i(t-\tau) \cong 2\, x_i(t) + \tau^2\, x''_i(t)$$

$$x_i(t+\tau) \cong 2\, x_i(t) - x_i(t-\tau) + (\tau^2/m_i)\, F_{xi}(t)$$

← $x''_i(t) = F_{xi}(t) / m$

- τ is the **integration step**. It should be smaller than the system's "fastest natural vibrations: ~ **10^{-15} seconds**
- At each step we perform one force calculation

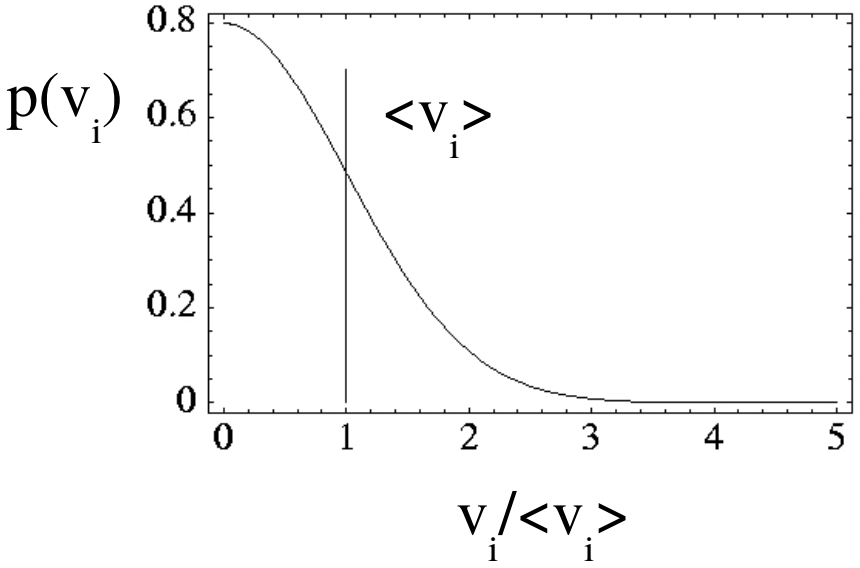
Link between temperature and atomic velocities

$p(v_i) = A \exp(-m_i v_i^2 / 2kT)$

T temperature

k Boltzmann constant

Maxwell-Boltzmann distribution



- Assign random initial velocities
- Rescale them from time to time

Recap: to simulate the dynamics of a biomolecule

- **"Molecular mechanics" "ball and stick" energy function**

- **Adjustable parameters (force constants, charges, etc) from small molecules, considered "transferable"**

- "Parabolic" approximation of $x_i(t)$

- Recursive Verlet algorithm to solve equations of motion

- (Simple) relation between atomic velocities and temperature

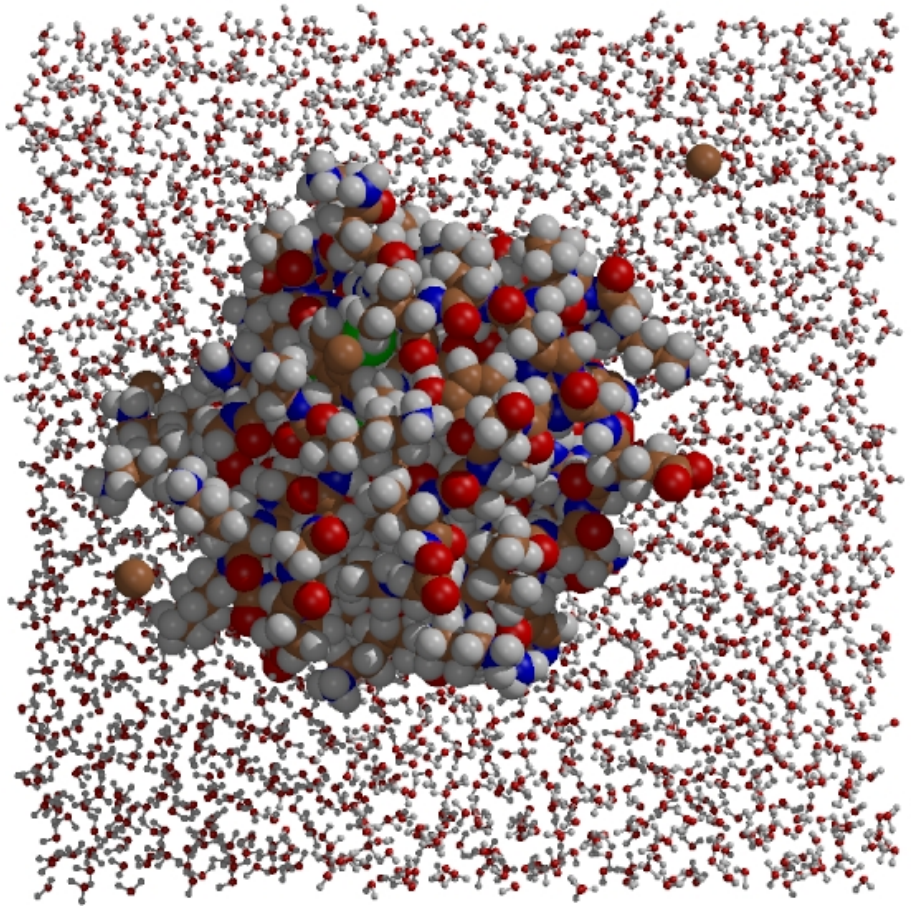
- **Strategy to model solvent....**

Simulated annealing:

Optimization on a rough energy surface

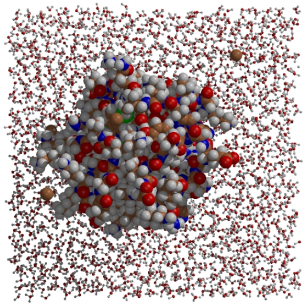
- Explore with molecular dynamics
- High temperature; lower gradually
- Refine structures from NMR or crystallography

Explicit solvent

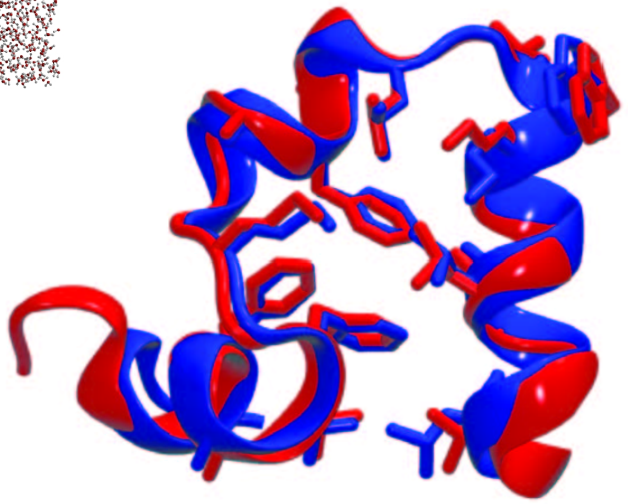


- Water box
- Periodic boundary conditions

In silico protein folding



vilin



FiP35

Domaine WW
35 acides aminés
Se replie expérimentalement en 14 µs
Par simulation en 10 µs
4000 molécules d'eau

Bleu: crystallography
Red: simulations

Shaw et al (2010)
Science, 330:341

Nobel chimie 2013
Martin Karplus

Structural Bioinformatics; editors PE Bourne, H Weissig; Wiley, 2003

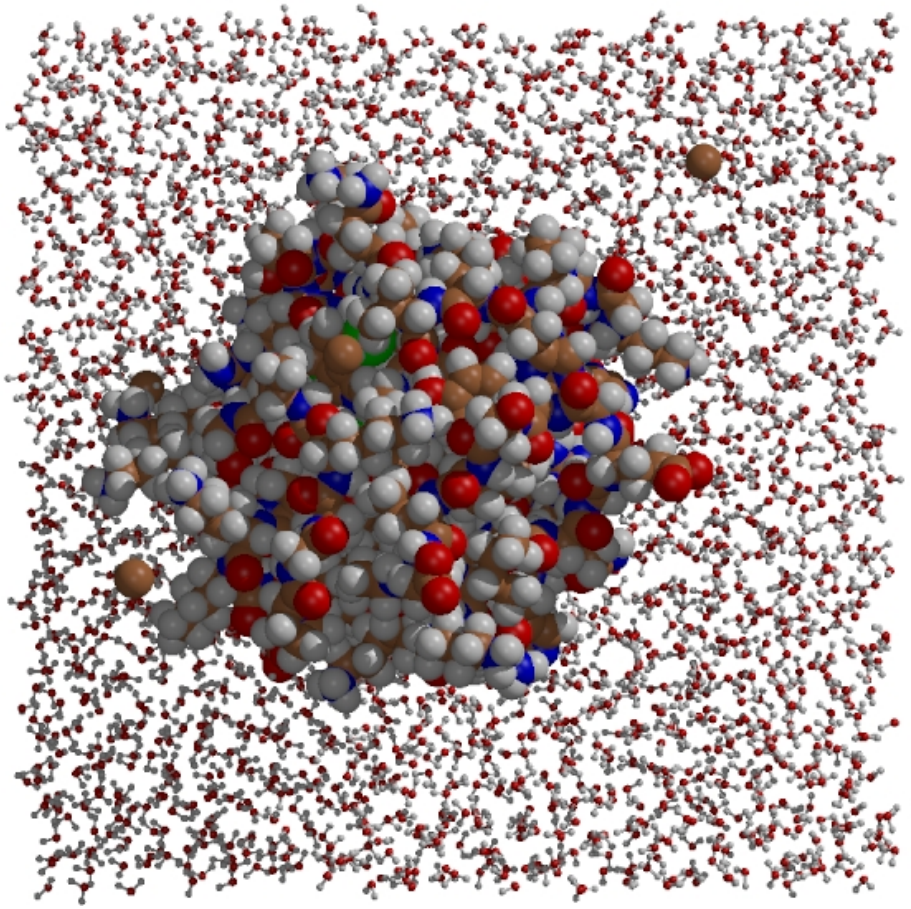
Computational Biochemistry & Biophysics, editors Becker, Mackerell, Roux, Watanabe; Marcel Dekker, 2001
Bioinformatics: from genomes to drugs; editor T Lengauer; Wiley, 2002
Molecular modelling: principles and applications; A Leach; Prentice Hall, 2001
Introduction to Genomics; A Lesk, Oxford University Press, 2007
Bioinformatique: génomique et post-génomique; Dardel & Képès, Ecole Polytechnique, 2002

X-PLOR version 3.8, A System for X-ray crystallography and NMR; Brunger, Yale University Press, 1992

Simonson (2005) Médecine Sciences 21:609-612; Peut-on prédire la structure des protéines?

CNS; Brunger, Adams, Clore, Delano, Gros, Grosse-Kunstleve, Jiang, Kuszewski, Nilges, Pannu, Read, Rice, Simonson, Warren (1998) Acta Cryst D54, 905.
Brunger, Adams, DeLano, Gros, Grosse-Kunstleve, Jiang, Pannu, Read, Rice, Simonson (2001)
The structure determination language of the Crystallography and NMR System.
International Tables of Crystallography, Volume F. Editors: M.G. Rossmann
and E. Arnold; Dordrecht: Kluwer Academic Publishers, the Netherlands.

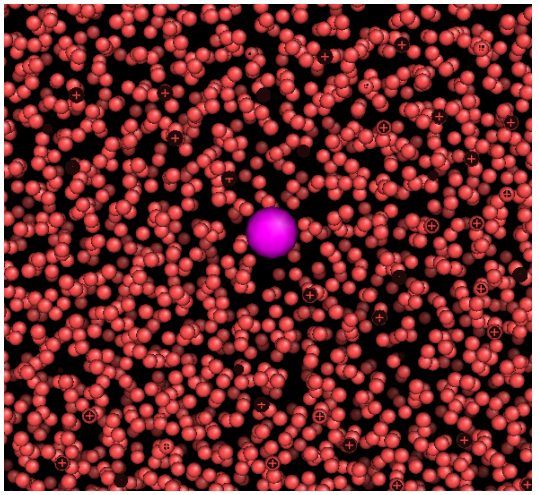
Explicit solvent



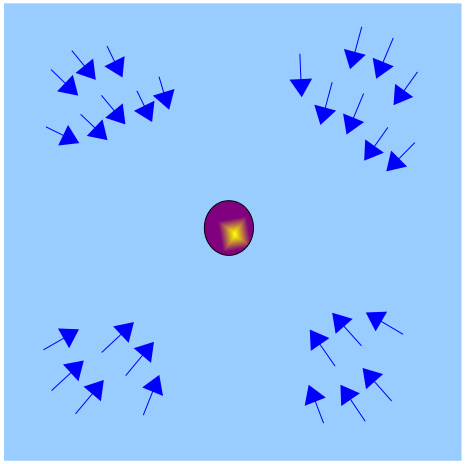
- Water box
- Periodic boundary conditions

Implicit solvent

“explicit”



“implicit”

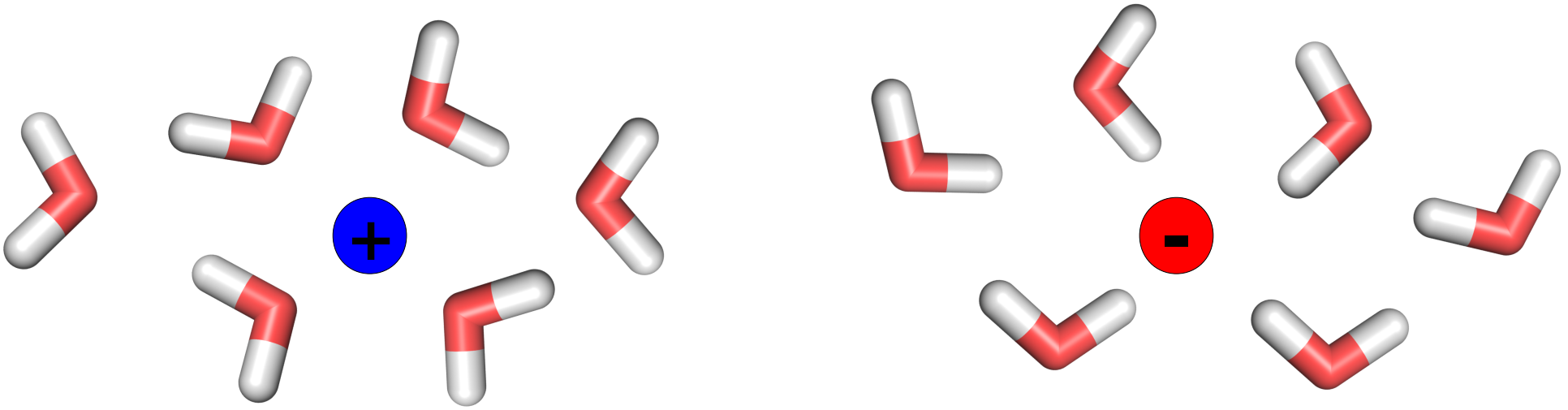


Dielectric continuum

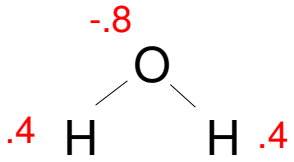
Electrostatic energy of a pair of ions: $U = q\,q' / \varepsilon\, r$

$\varepsilon = 80 =$ dielectric constant of water

Ion-ion interactions are greatly reduced in water



From a distance (a nanometer), each ion's charge is partly compensated by nearby water oxygens and hydrogens.



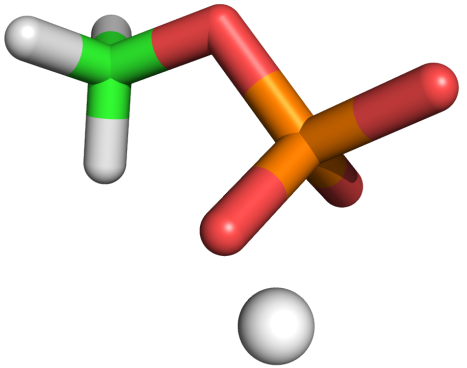
$$U = q\,q' / \varepsilon\, r$$

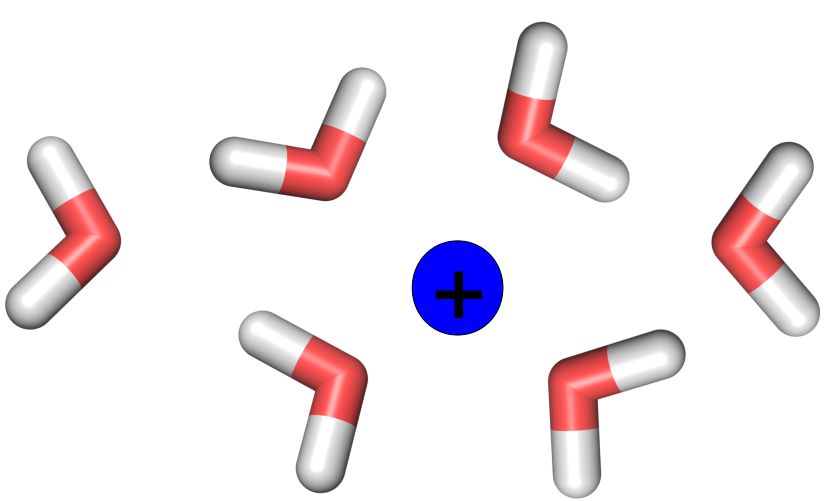
Importance of solvent: Mg^{2+} – HPO_4^{2-} association

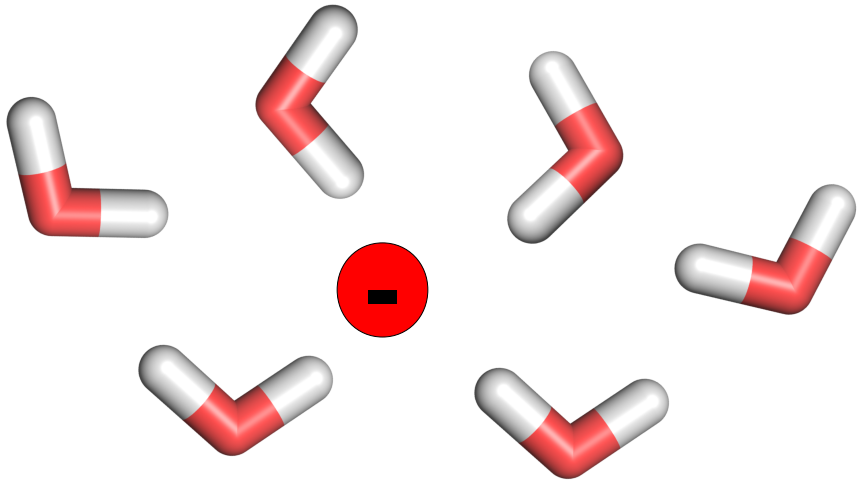
Association energy in vacuum: **-515 kcal/mol**

In water: **-3.7 kcal/mol**

[standard state conditions]





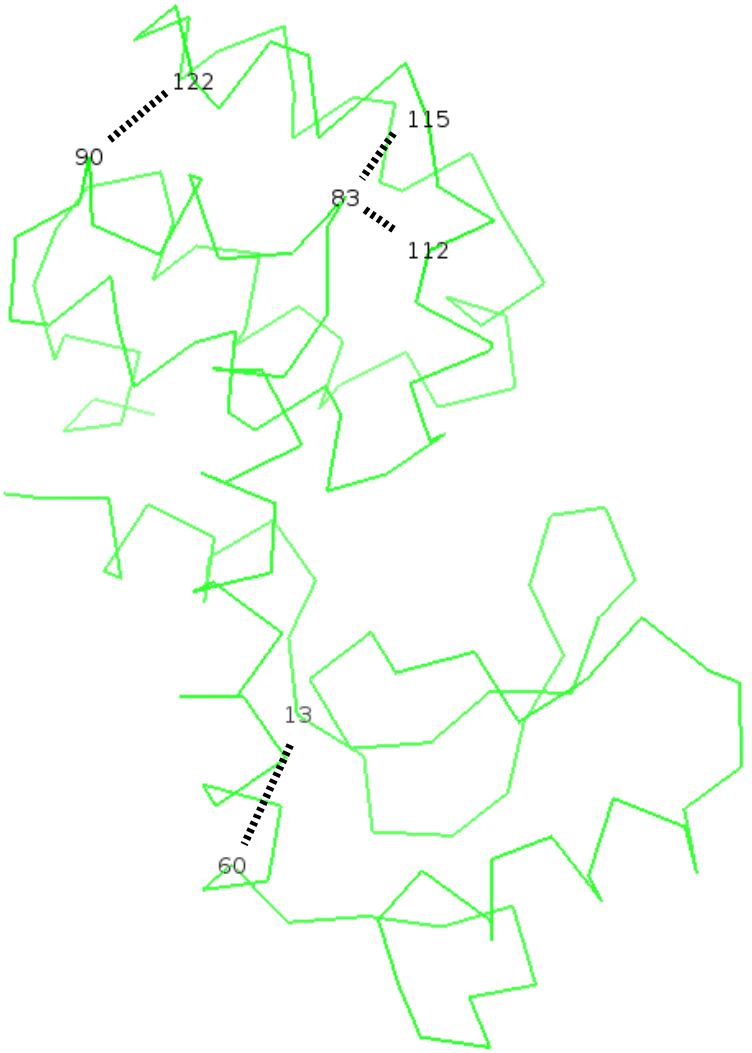


Water has a "screening" effect, reduction by almost 100.

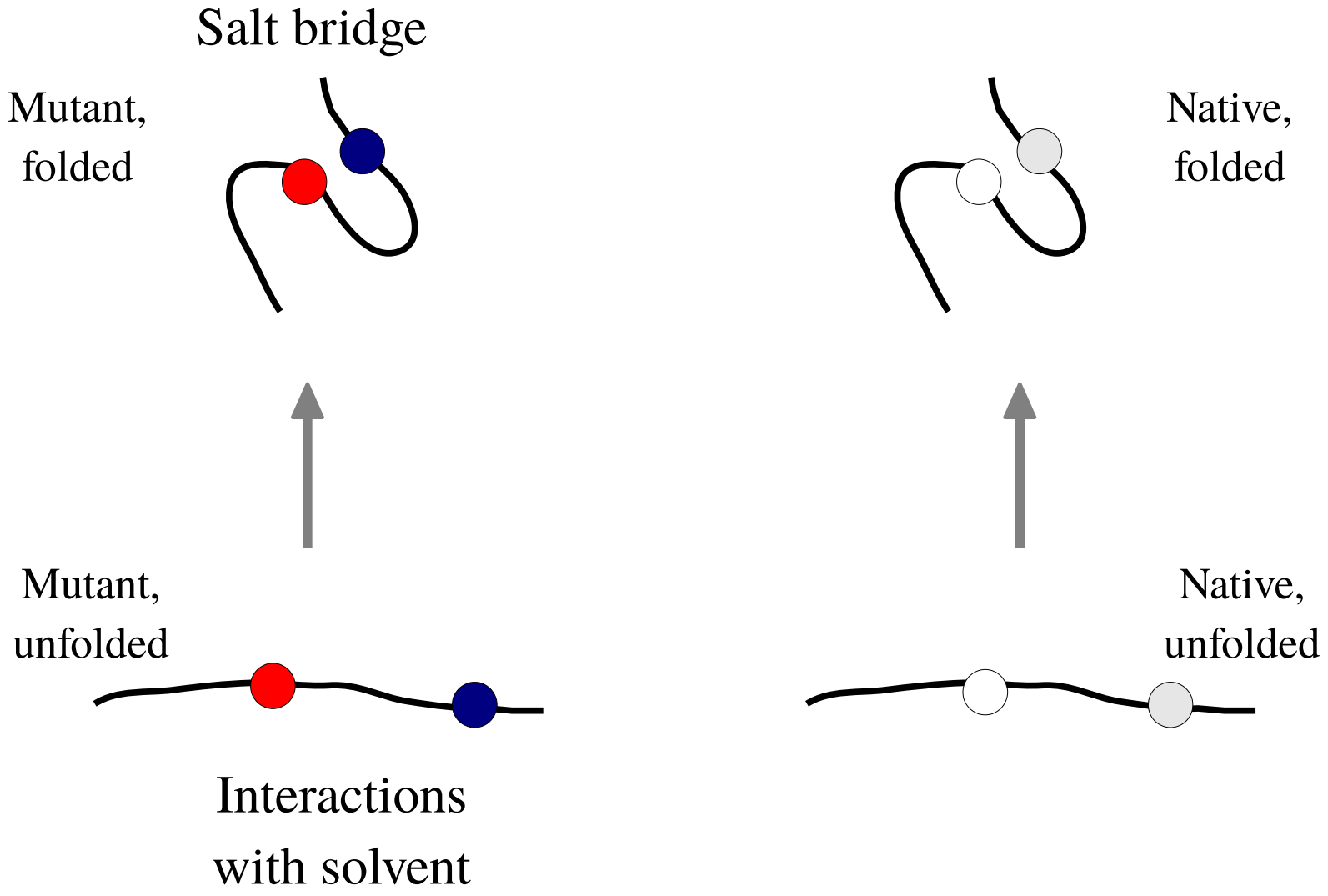
Effect of salt bridges on stability of T4 lysozyme

Dao-Pin, Nicholson, Matthews (1991) Biochemistry, 30:7142

Mutation	ΔΔG(kcal/mol)
K60H/L13D	-2.8
K83H/A112D	-1.5
S90H/Q122D	-2.2
T115E (...K83)	+0.3
S90H	-1.1
K60H	-0.1
K83H	-0.5



Small contribution, compared to enormous vacuum interaction energies

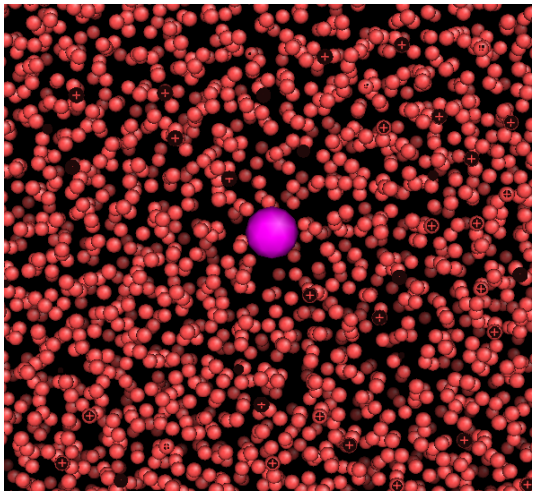


Compensating effects

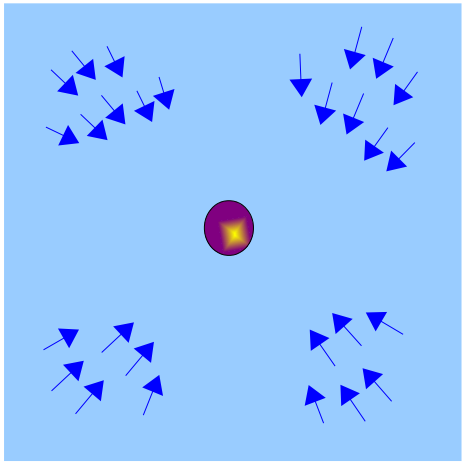
90% of salt bridges are at the surface

Implicit solvent

“explicit”



“implicit”



Dielectric continuum

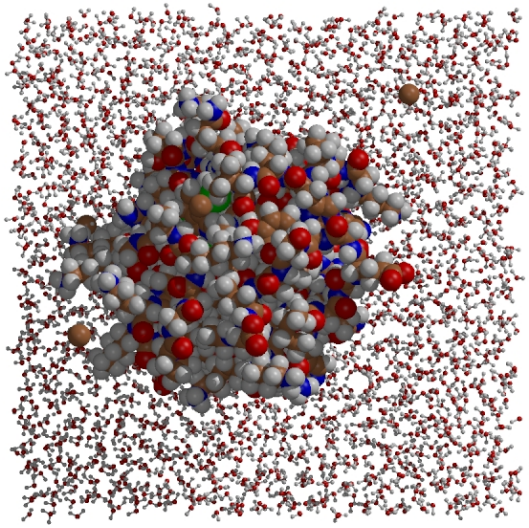
Electrostatic energy of a pair of ions: $U = q\,q' / \varepsilon\, r$

$\varepsilon = 80 =$ dielectric constant of water

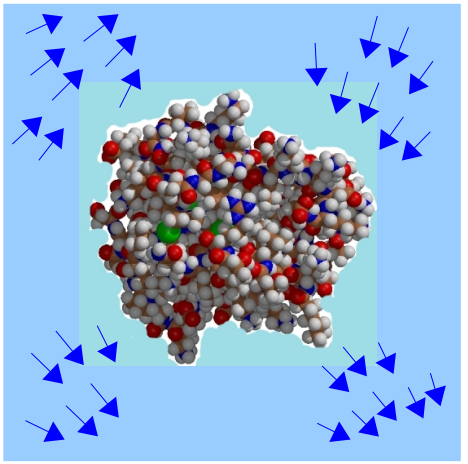
Fine for a few ions in water.....

Implicit solvent

“explicit”



“implicit”



Dielectric continuum

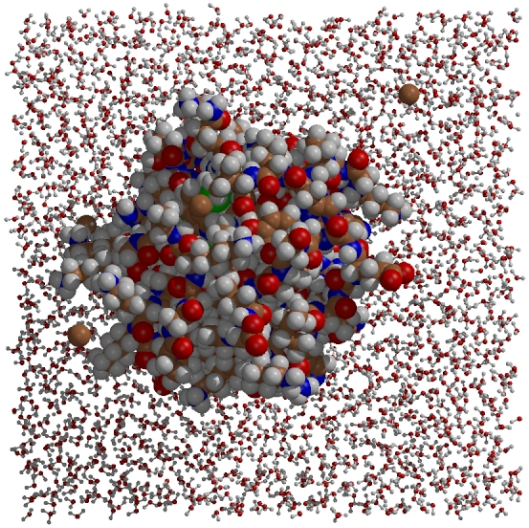
A biomolecule is more complex: heterogeneous system

Initially, we can ignore this complexity: $U = q\,q' / \varepsilon\, r$

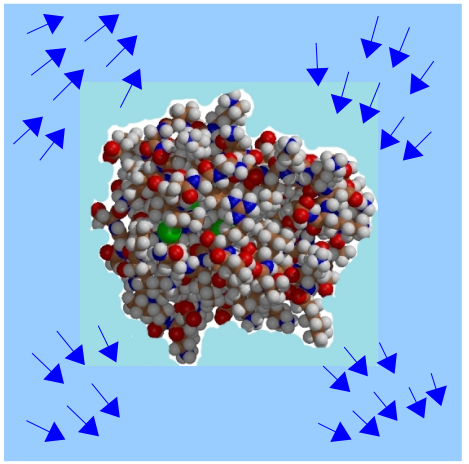
Going further: more sophisticated model

Implicit solvent: "generalized Born" model (GB)

"explicit"



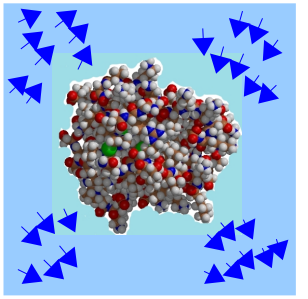
"implicit"



- Buried charges: $E = q\,q' / r$
- Exposed charges: $E = q\,q' / \varepsilon\, r$

  $\varepsilon = 80$
- Intermediate cases: **interpolation**

“Generalized Born” solvent model (GB)



$$U_{elec} = U_{Coul} + U_{GB}$$

$$U_{Coul} = \Sigma_{i,j}\, q_i\, q_j / r_{ij}$$ interaction with no solvent

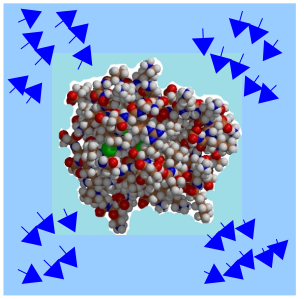
$$U_{GB} = (-1 + 1/\varepsilon)\ \Sigma_{i,j}\, q_i\, q_j\, F_{GB}(i,j)$$ term due to solvent

$$F_{GB}(i,j) = 1/(r_{ij}^2 + b^2_{ij} \exp[-r_{ij}^2/4b^2_{ij}])^{1/2}$$

$$b_{ij} = (\, b_i\, b_j\,)^{1/2}$$

b_i , b_j = distance au solvant des atomes i, j

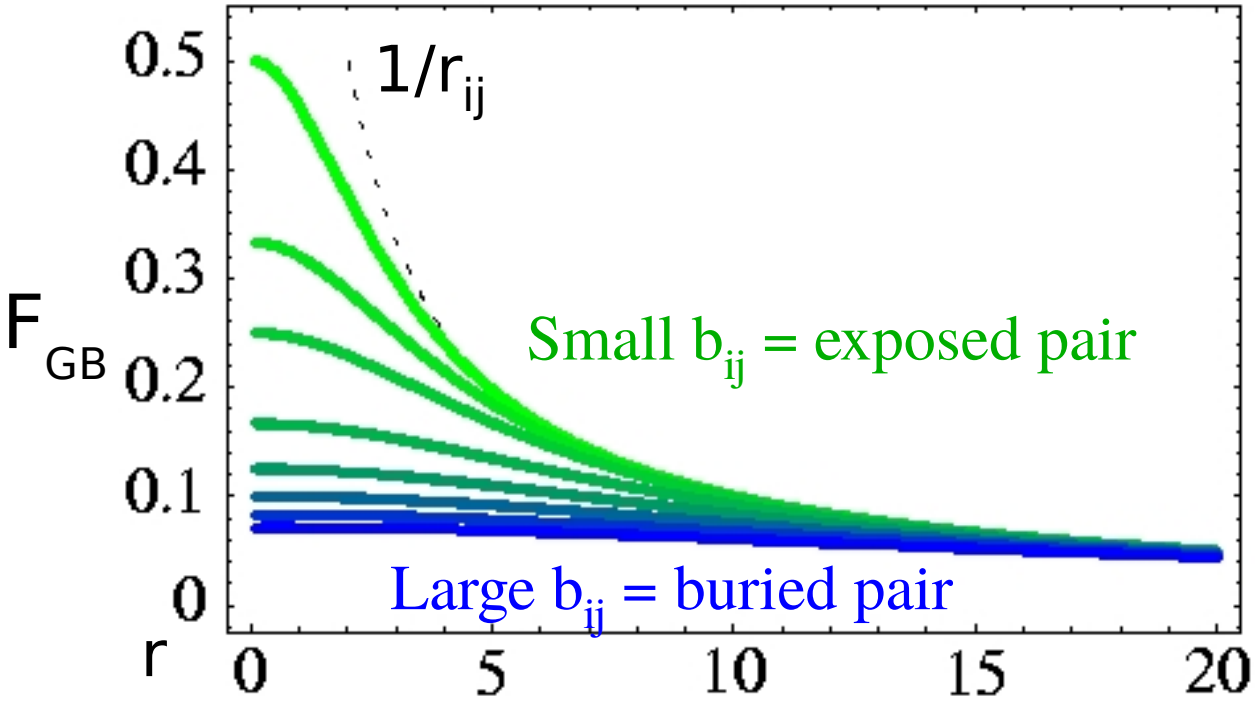
“Generalized Born” solvent model (GB)



$$U_{elec} = U_{Coul} + U_{GB}$$

$$U_{GB} = (-1 + 1/\varepsilon)\ \Sigma_{i,j}\ q_i\ q_j\ F_{GB}(i,j)$$

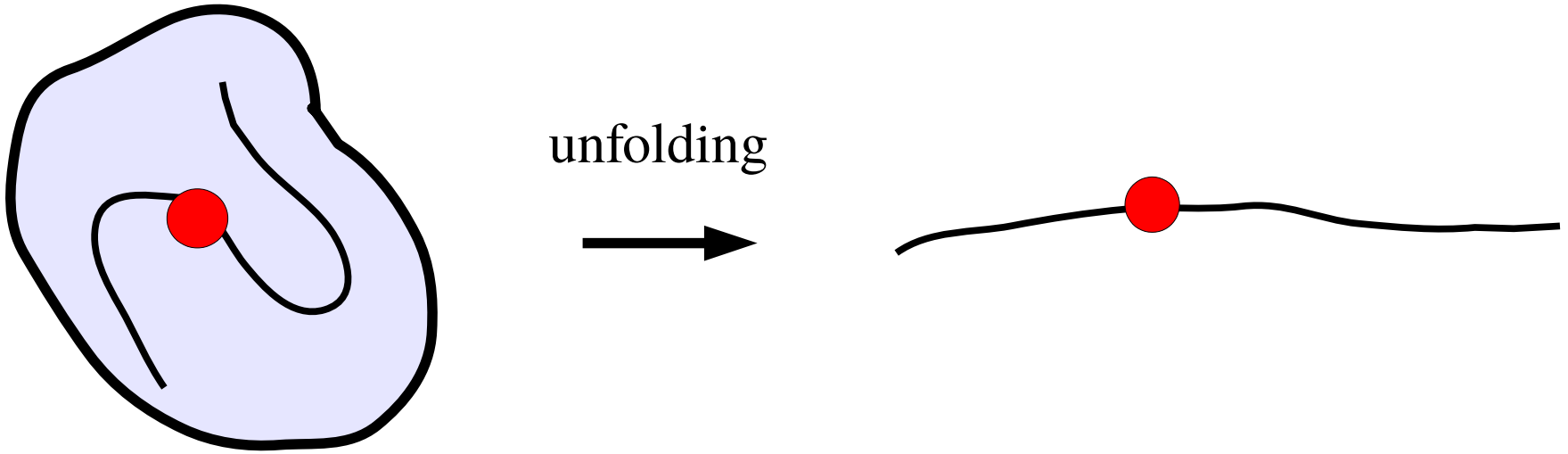
$$F_{GB}(i,j) = 1/(r_{ij}^2 + b^2_{ij}\ \exp[-r_{ij}^2/4b^2_{ij}])^{1/2}$$



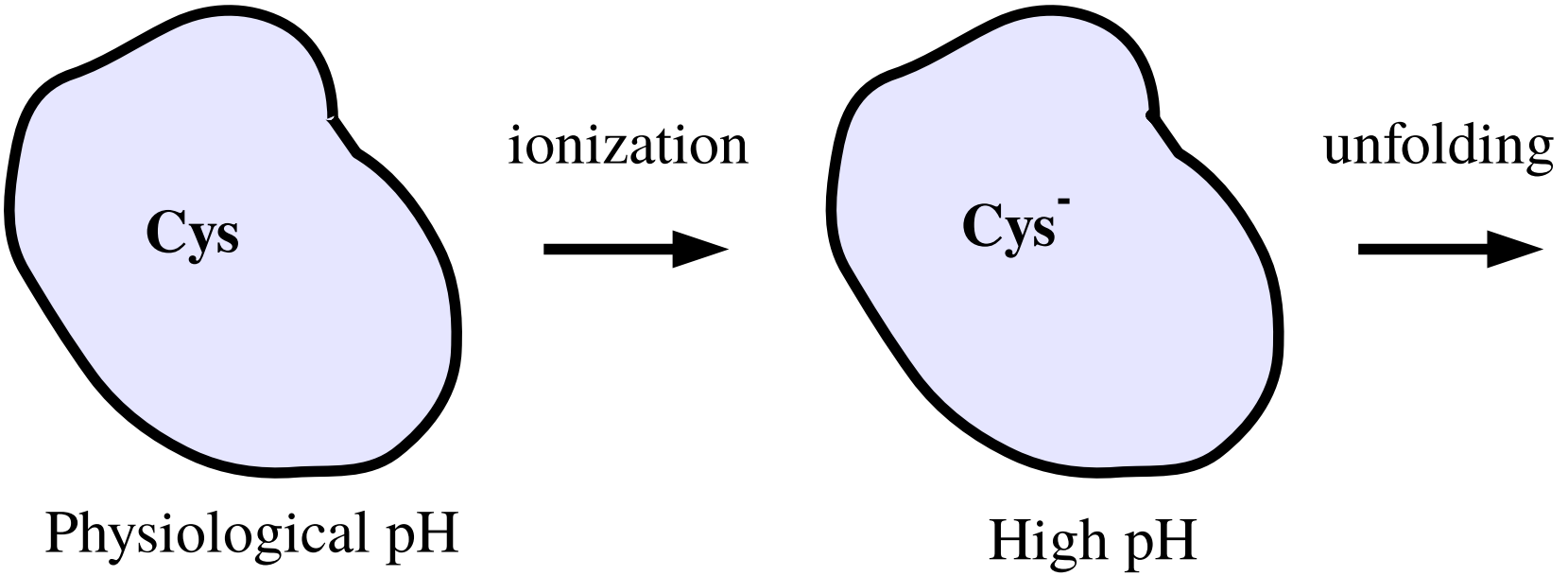
Small b_{ij} : compensation between U_{Coul} et U_{GB}, U_{elec} close to U_{Coul}/ε

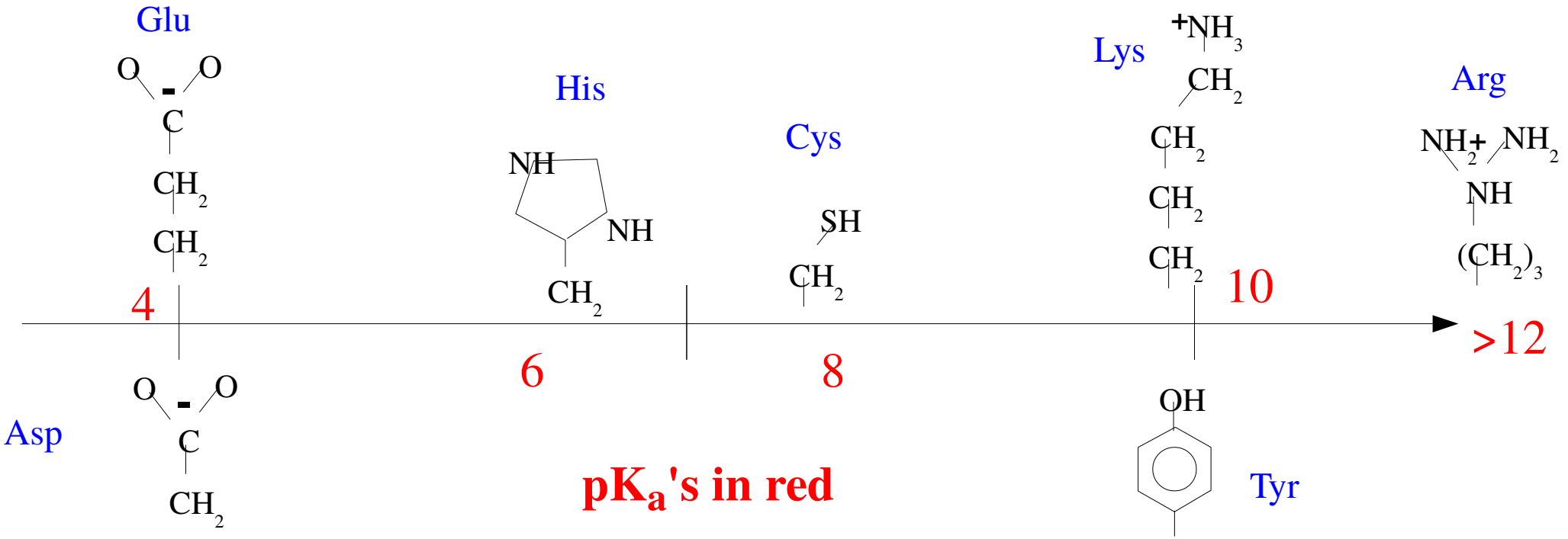
Large b_{ij}: little compensation, U_{elec} close to U_{Coul}

A buried charge would destabilize a protein: why?



A buried charge would destabilize a protein: why?

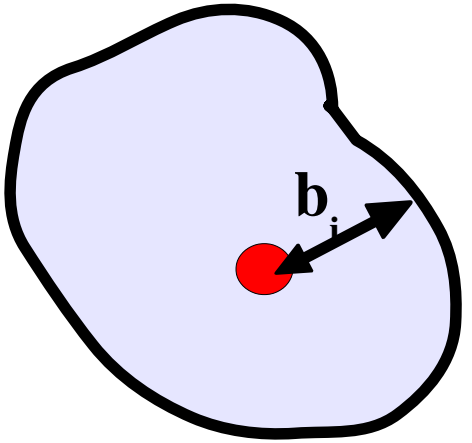




How does GB handle this?

U_{GB} contains "ii" terms

A single charge:



$$U_{GB} = (-1 + 1/\varepsilon)\, q_i^2 / b_i$$

$$\cong - q_i^2 / b_i$$

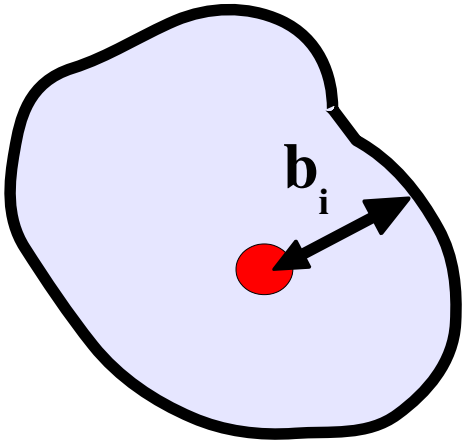
$1/\varepsilon$ is small, so $U_{GB} = -q_i\, q_i / b_i$:

$q_i \leftrightarrow$ solvent interaction = as if q_i sees a charge $-q_i$ at a distance of b_i

"Self" energy term GBSE in XPLOR

U_{GB} contains "ii" terms

To get ii term, let $r_{ij} \rightarrow 0$:



$$U_{GB} = (1/\varepsilon - 1) \Sigma_{ij} q_i q_j / \underbrace{(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j])^{1/2}}_{F_{GB}(r_{ij})}$$

$$i = j \; ; \quad r \rightarrow 0 \; ; \quad F_{GB} \rightarrow 1 / b_i \; ; \quad U_{GB} \rightarrow (1/\varepsilon - 1) \, q_i^2 / b_i$$

Polarity of side chain analogs with GB

Some ions

ion	radius[a] (Å)	ΔG_{exp}[b] (kcal/mol)
Li^+	0.78	−122.1
Na^+	0.98	−98.2
K^+	1.33	−80.6
Rb^+	1.49	−75.5
Cs^+	1.65	−67.8
Mg^{2+}	0.78	−455.5
Mn^{2+}	0.91	−437.8
Ca^{2+}	1.06	−380.8
Sr^{2+}	1.27	−345.9
Ba^{2+}	1.43	−315.5

Cf also Fersht; Structure & mechanism in protein science; Freeman.

	solute	area	Exp.	ACE
1	acetamide	216	-9.70	-13.6
2	acetic acid	212	-6.73	-10.2
3	methane thiol	192	-1.24	-6.6
4	propionamide	247	-9.40	-13.3
5	propionic acid	245	-6.47	-9.3
6	4-methyl imidazole	259	-10.25	-9.7
7	3-methyl indole	331	-5.90	-4.9
8	p-cresol	303	-6.13	-5.3
9	ethyl methyl sulfide	246	-1.49	-3.0
10	methanol	174	-5.10	-8.9
11	ethanol	206	-4.95	-7.5
12	toluene	297	-0.76	-0.9
13	butyl amine	272	-4.38	-1.1
14	ethane thiol	220	-1.30	-5.9
15	dimethyl sulfide	222	-1.54	-3.3
16	N-methyl acetamide	254	-10.07	-9.4
17	acetone	233	-3.80	-3.5
18	butanoic acid	274	-6.35	-9.3
19	1-propanol	236	-4.84	-6.7
20	2-propanol	235	-4.75	-6.3
21	2-methyl-2-propanol	260	-4.52	-5.2
22	methyl acetate	246	-3.31	-2.3
23	ethyl acetate	280	-3.09	-2.7
24	methyl amine	178	-4.57	-4.0
25	ethyl amine	209	-4.56	-3.3
26	propyl amine	241	-4.45	-3.0
27	benzene	275	-0.88	-2.4
28	phenol	253	-6.50	-5.4
29	dimethyl acetamide	272	-8.54	-5.1
		rms error		1.33

Solvation free energies (transfer vapor →water) in kcal/mol.

J Comp Chem, 1999, 20:322

Exc: expliquer le dépliement des protéines à fort et faible pH

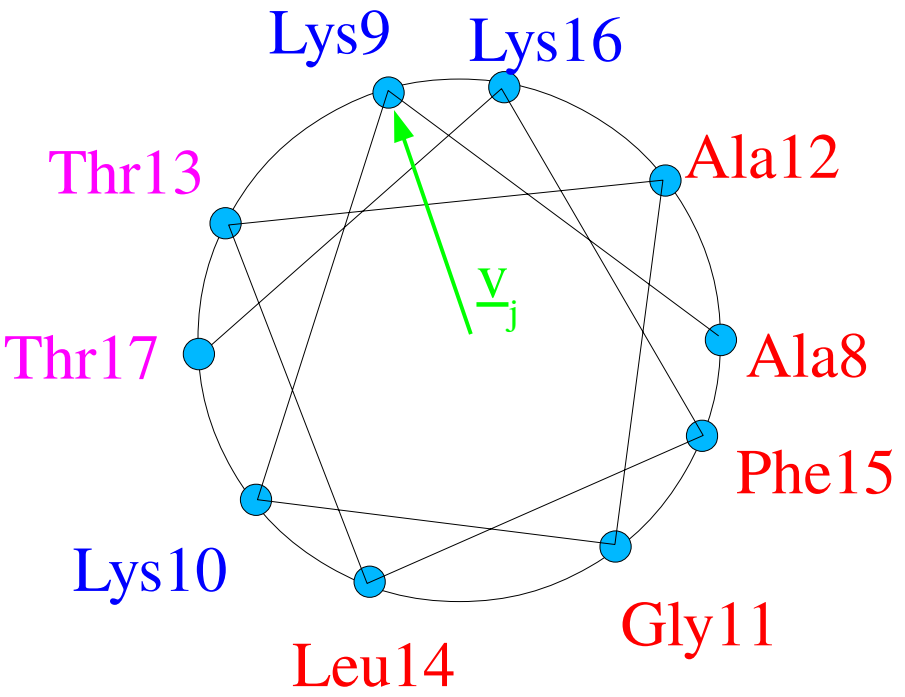
Exercice:

En utilisant XPLOR et le modèle GB, estimer l'énergie (libre) d'une charge ponctuelle placée sur un atome enfoui au centre du Trpcage, toutes les autres charges atomiques étant mises à zéro.
Comparer à ce que vous obtenez avec une charge placée sur un acide aminé seul (le reste de la protéine étant enlevée).

On admet que quand le Trpcage se replie, l'énergie libre varie de 0 à -5 kcal/mol. Quelle est, à l'équilibre, la fraction de protéine repliée?

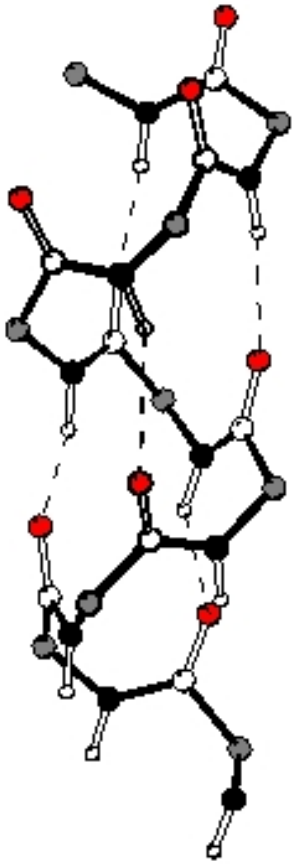
Nous avons vu ci-dessus l'effet d'une “mutation” sur l'énergie de la protéine repliée et celle de la protéine dépliée. (On utilisera le calcul fait pour un acide aminé isolé pour représenter la protéine dépliée; est-ce que c'est raisonnable?) Pour la protéine “mutée”, quelle est la fraction de protéine repliée à l'équilibre?

Secondary structure prediction



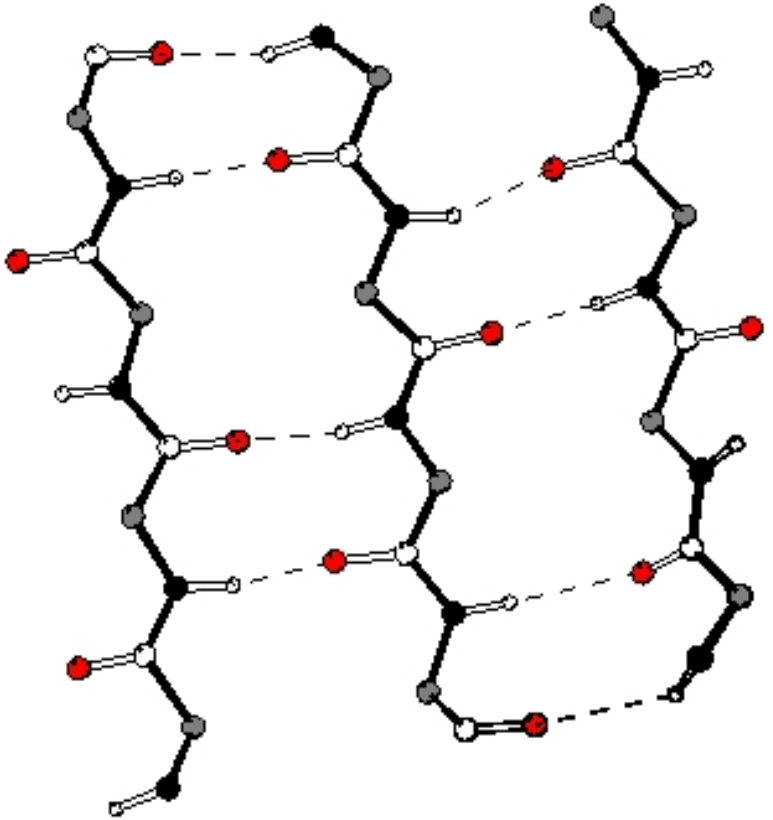
Secondary structure prediction

Hélice α



3.6 residues per turn

Feuillet β

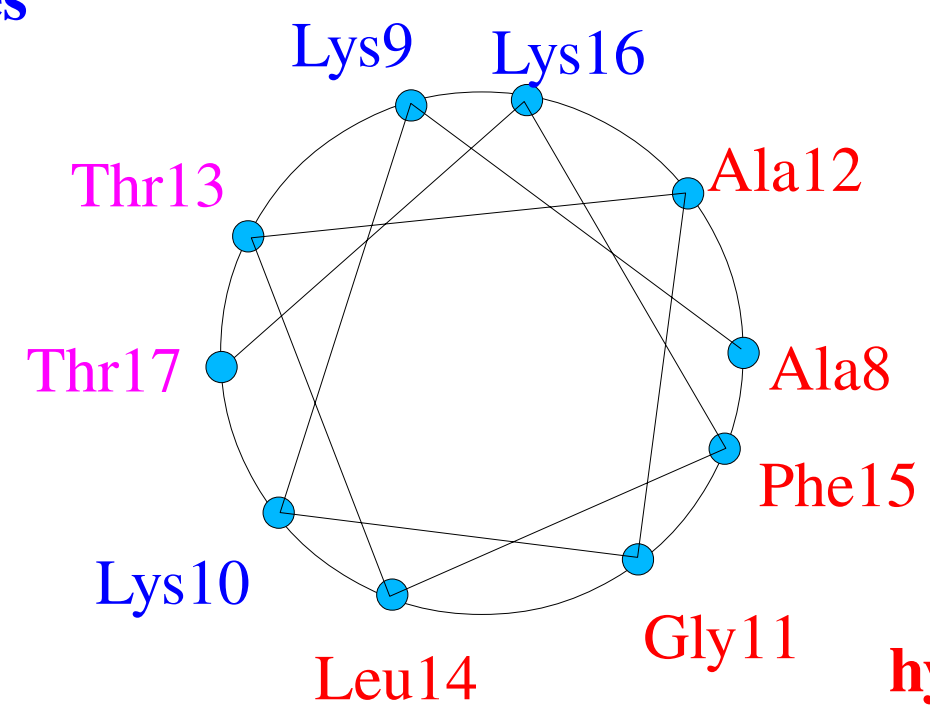


Pseudo-periodicity 2

Secondary structure prediction

Helices are often amphipathic, with a buried and an exposed side

hydrophiles



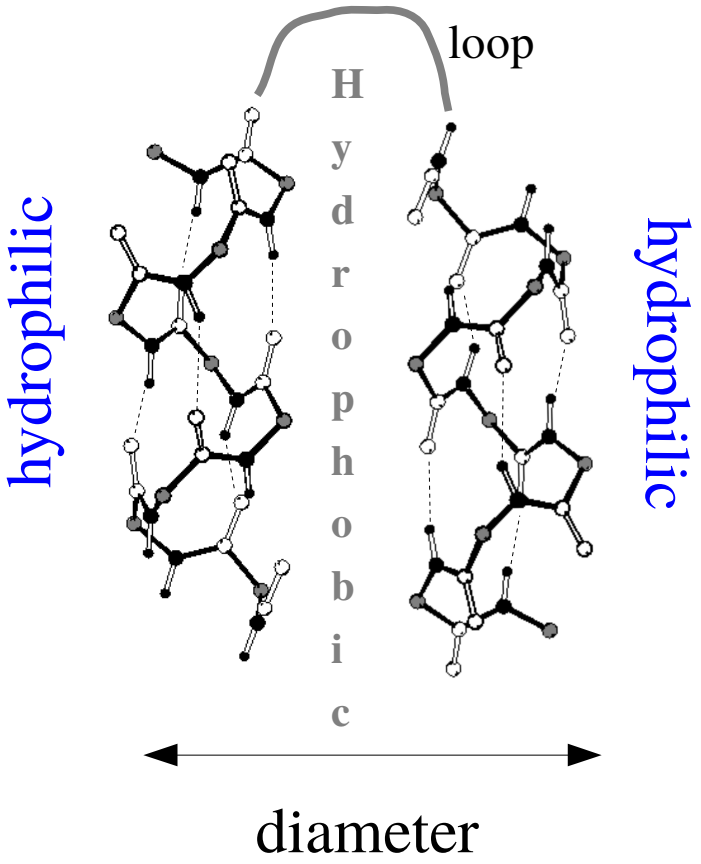
"Helical wheel"

hydrophobes

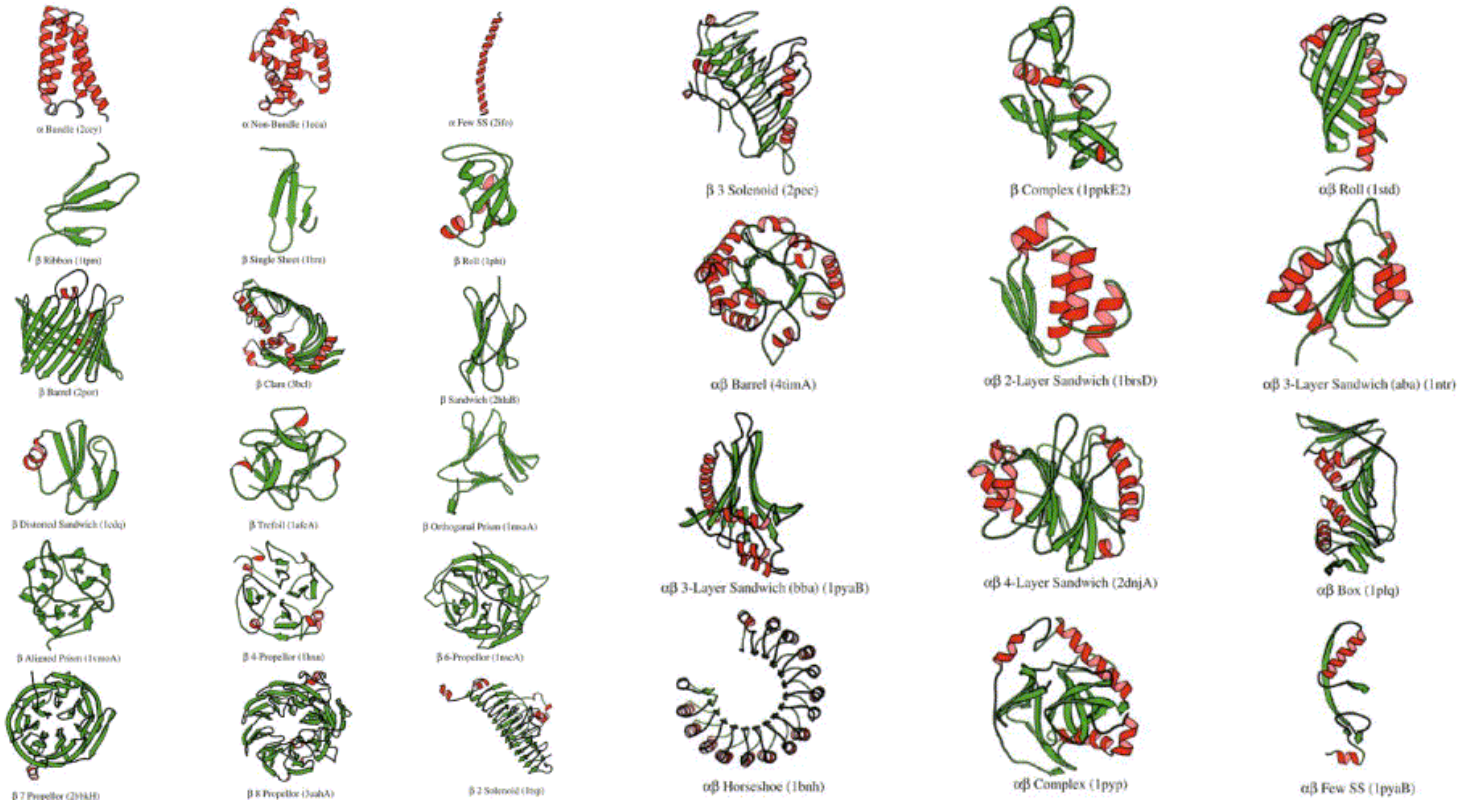
Simplified, amphipathic helix

In proteins, the diameter of a domain often corresponds to a few secondary structure elements

Helices are often amphipathic, with a buried and an exposed side



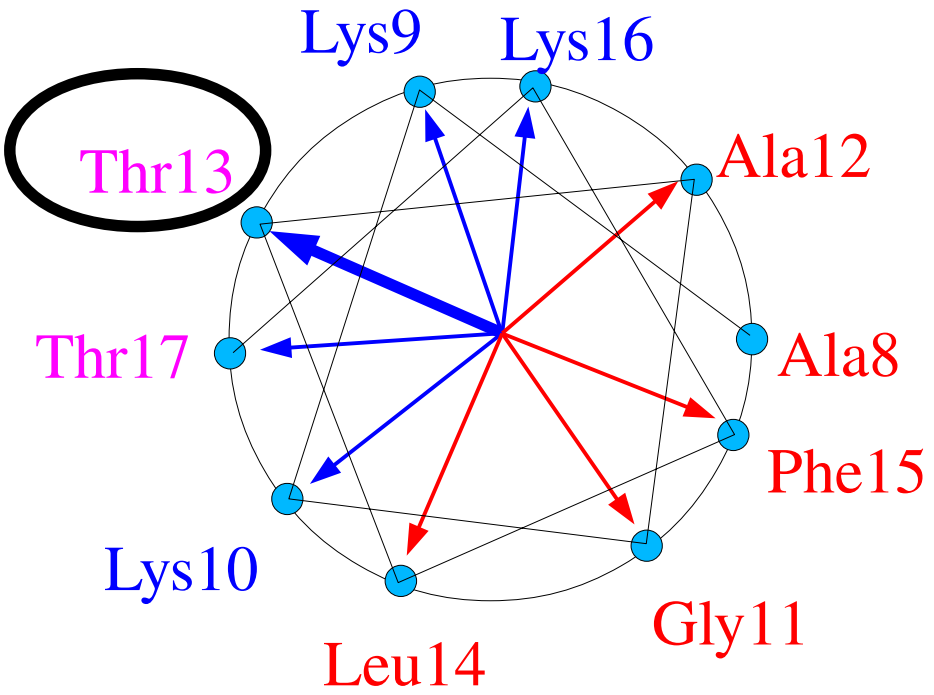
In proteins, the diameter of a domain often corresponds to a few secondary structure elements



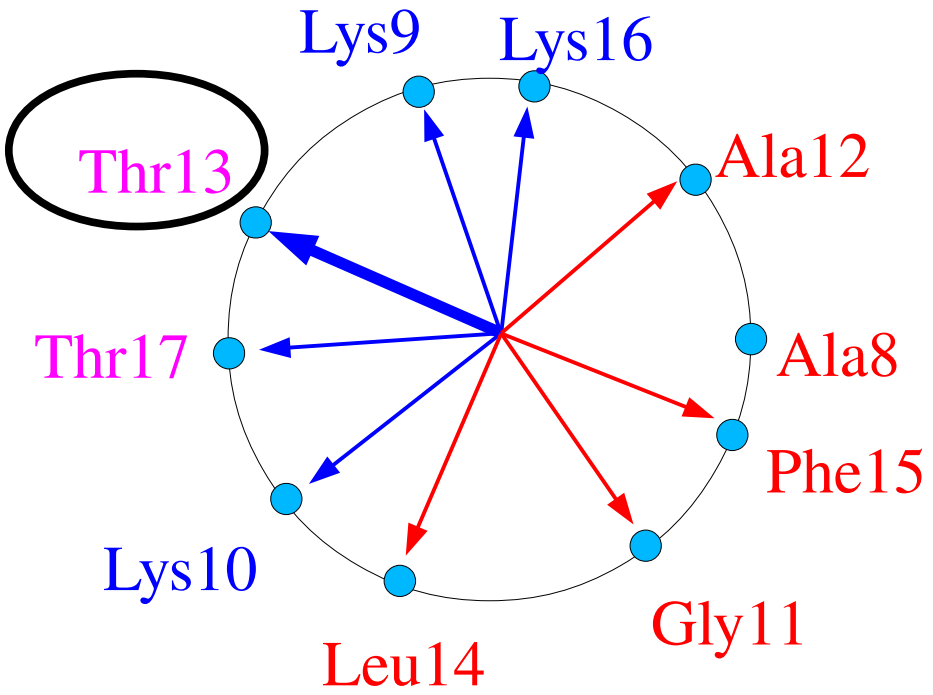
30 “architectures” from CATH

www.cathdb.info

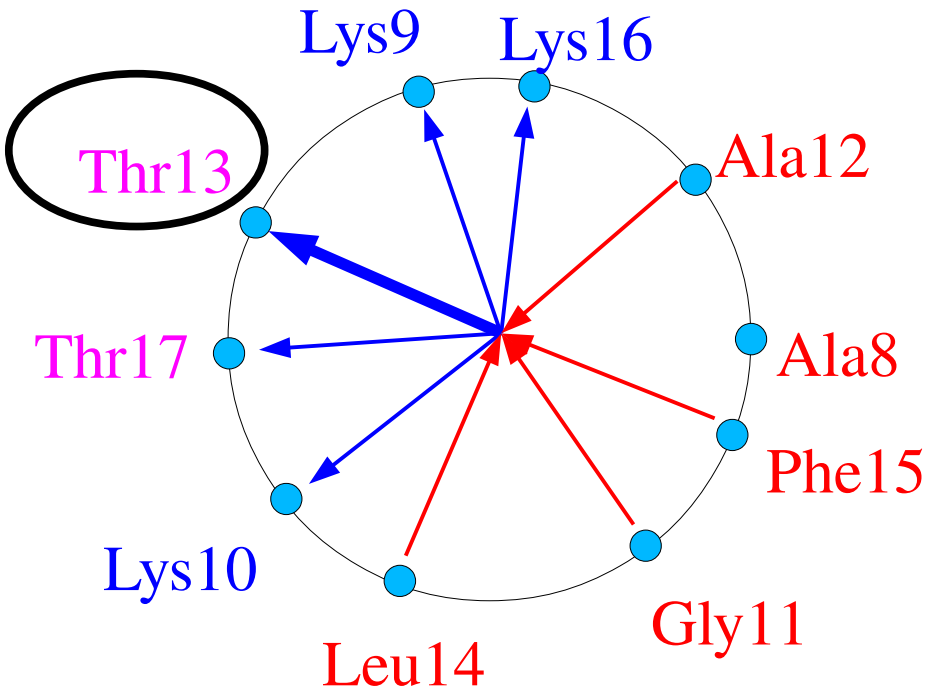
Consider an amino acid, say **Thr13**. We assume it is in a helix, shown below. Its neighbors 9, 10, 11, 12, **13**, 14, 15, 16, 17 define radial vectors:



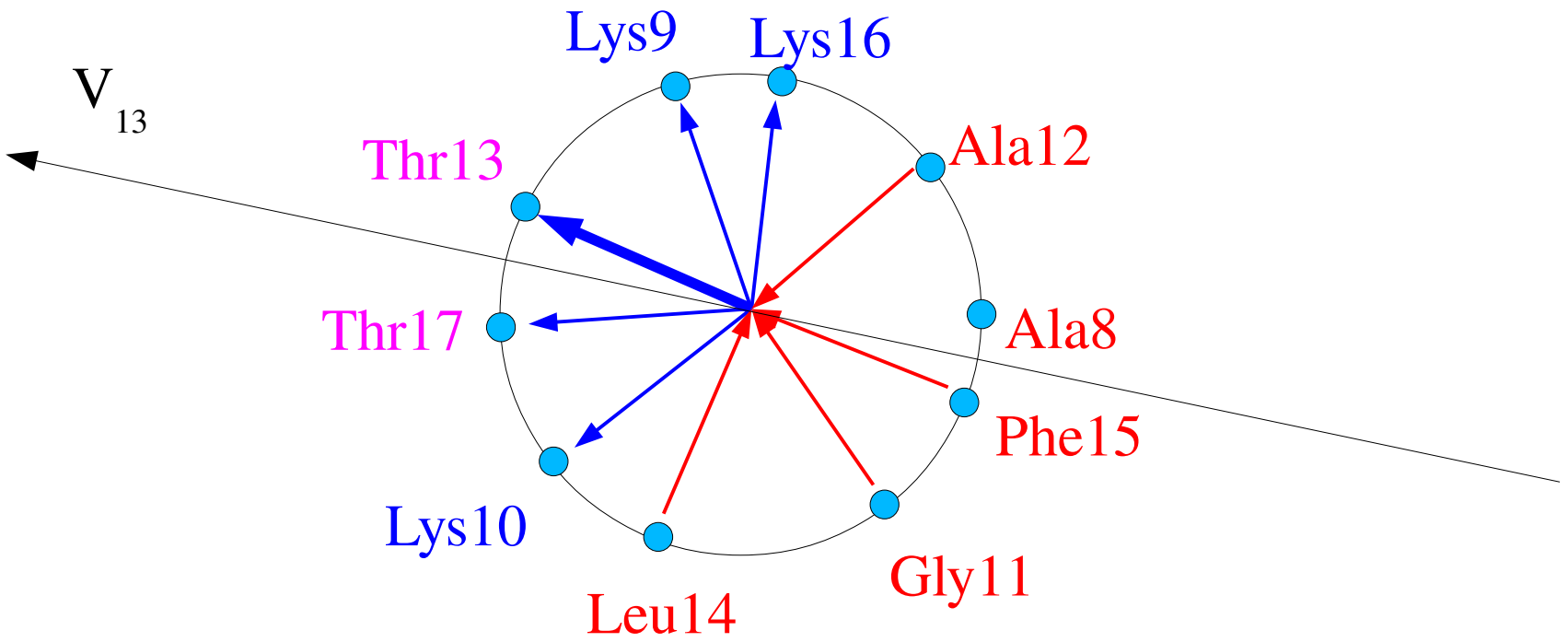
Consider an amino acid, say **Thr13**. We assume it is in a helix, shown below. Its neighbors 9, 10, 11, 12, **13**, 14, 15, 16, 17 define radial vectors:



Consider an amino acid, say **Thr13**. We assume it is in a helix, shown below. Its neighbors 9, 10, 11, 12, **13**, 14, 15, 16, 17 define radial vectors. For hydrophobes, reverse the vectors:



Sum the vectors:



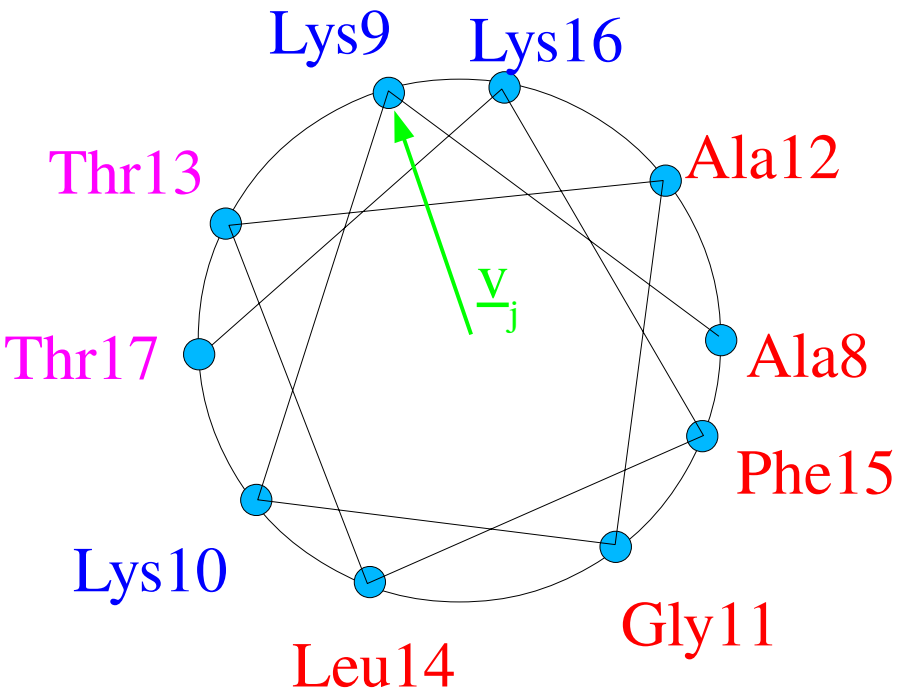
If our hypothesis is correct, we obtain a large vector V_{13} (the v_j add up "constructively"). Else, V_{13} is small.

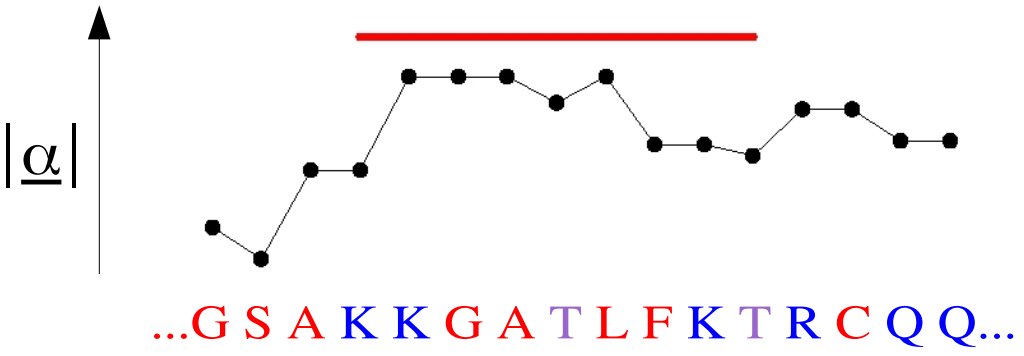
We have defined a “hydrophobic moment”:

$$\underline{\alpha}_i = \sum_{j=\, i-m}^{i+\, m} H_j \; \underline{v}_j$$

H_j = hydrophobicity, depends on amino acid type

m = 3 ou 4





The hydrophobic moment reveals amphipathic regions, which have periodicity 3.6.

Beta sheet hypothesis: we define a different hydrophobic moment

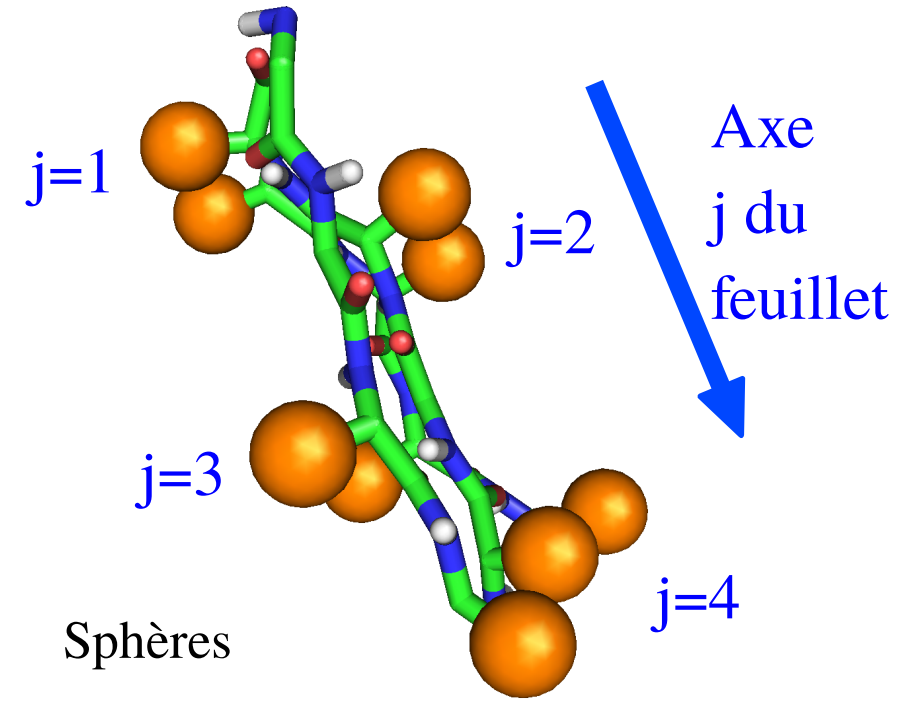
$$\beta_i = \sum_{j=\, i-m}^{i+\, m} H_j \; v_j$$

$$v_j = (-1)^j$$

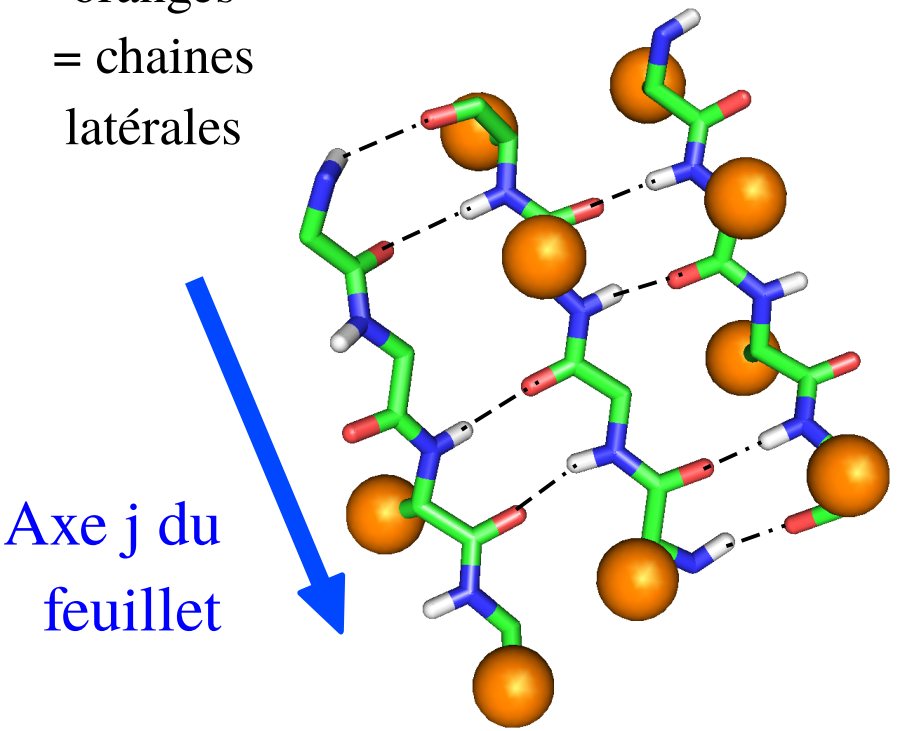
(side chain above or below sheet)

Correct hypothesis:
large $|\beta_i|$

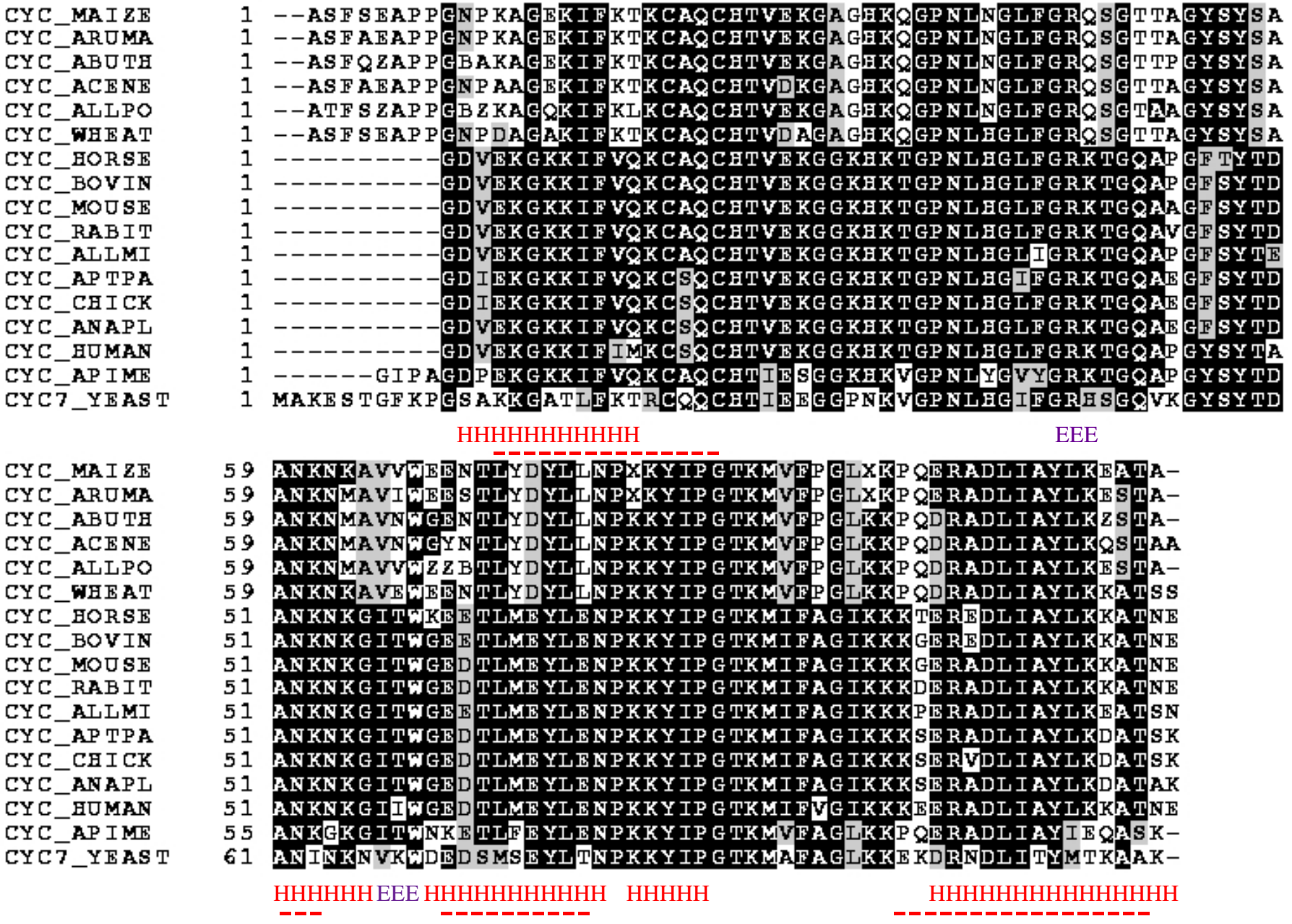
Pseudo-periodicity 2



Sphères oranges = chaines latérales

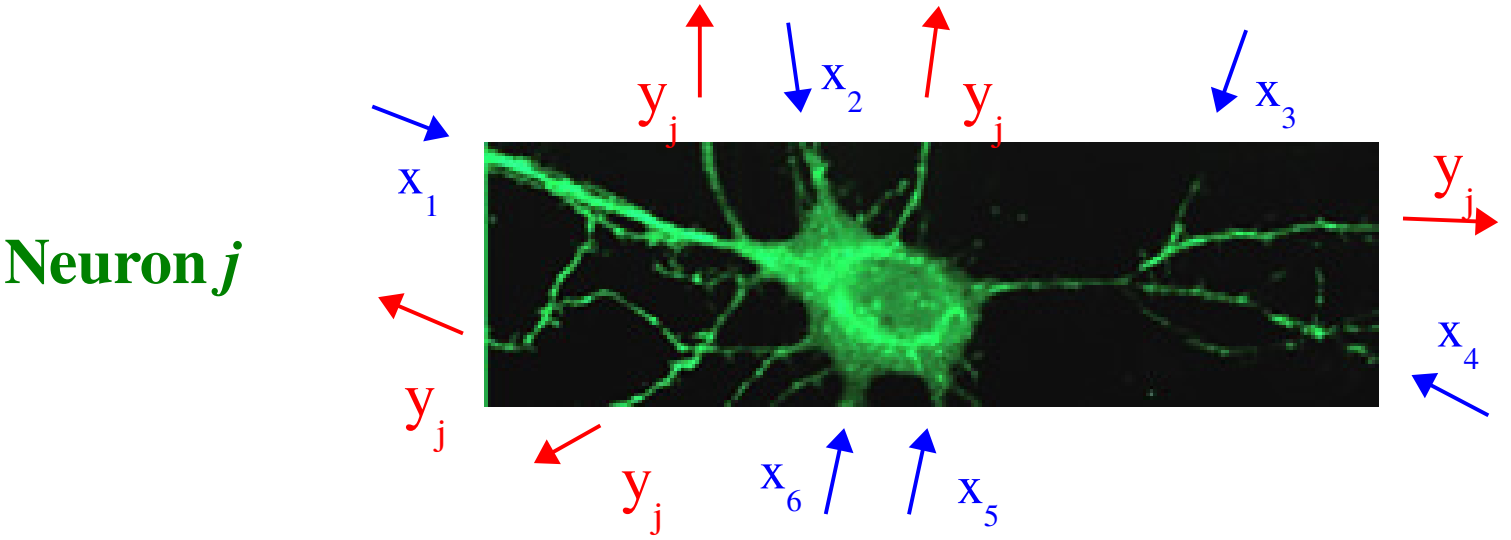


~ 75% success in general



A method inspired by neural networks

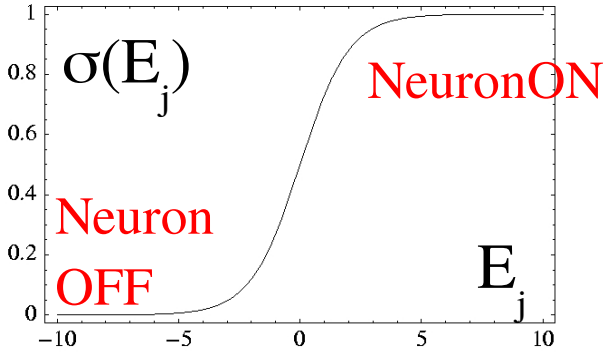
A neuron *j* receives input signals $x_1, x_2, ..., x_n$ from nearby cells *1*, *2*, ..., *n*; sums them, and converts them into an output signal y_j:



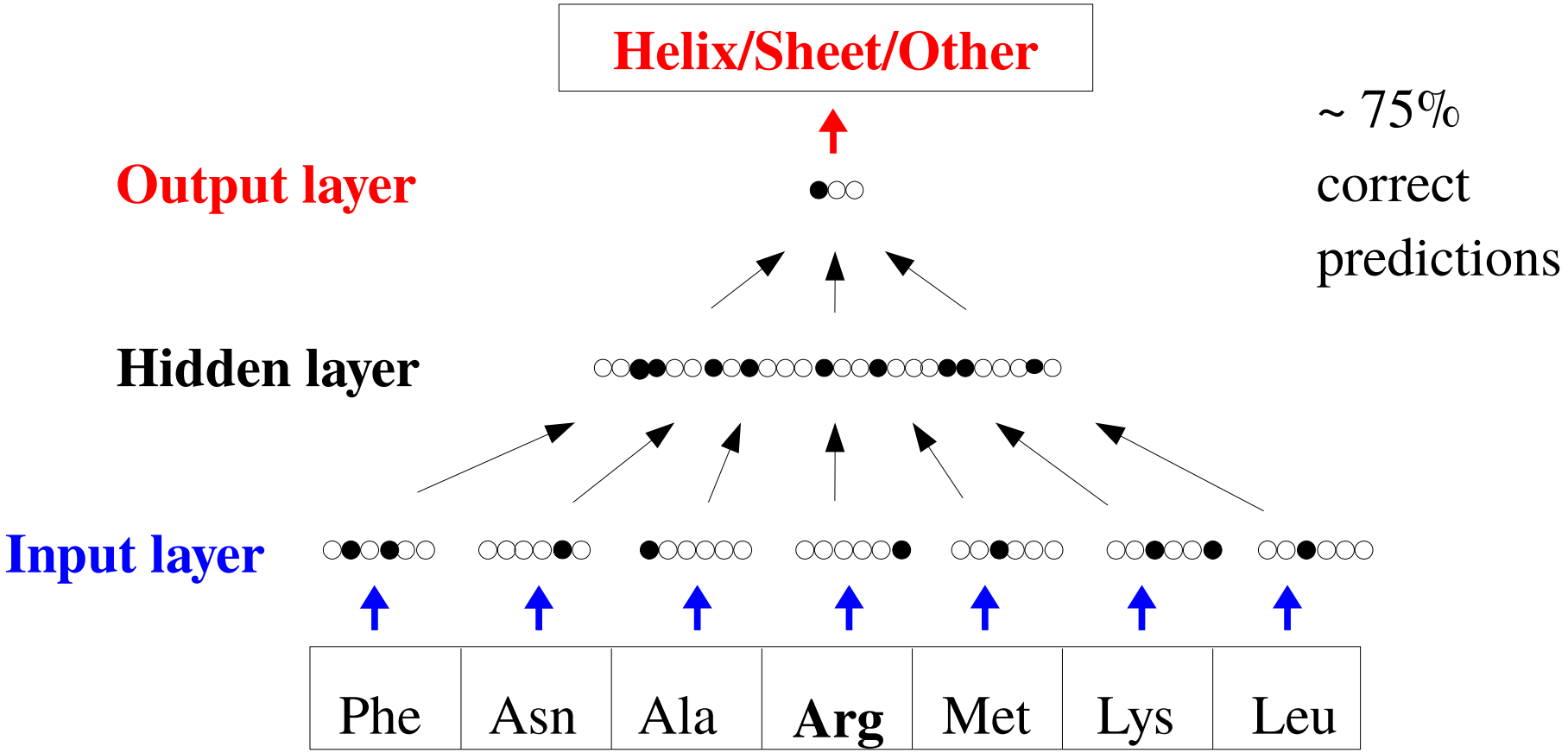
Input signals x_i

are weighted and summed: $E_j = \Sigma_i w_{ij} x_i$

then converted: $y_j = \sigma (E_j)$

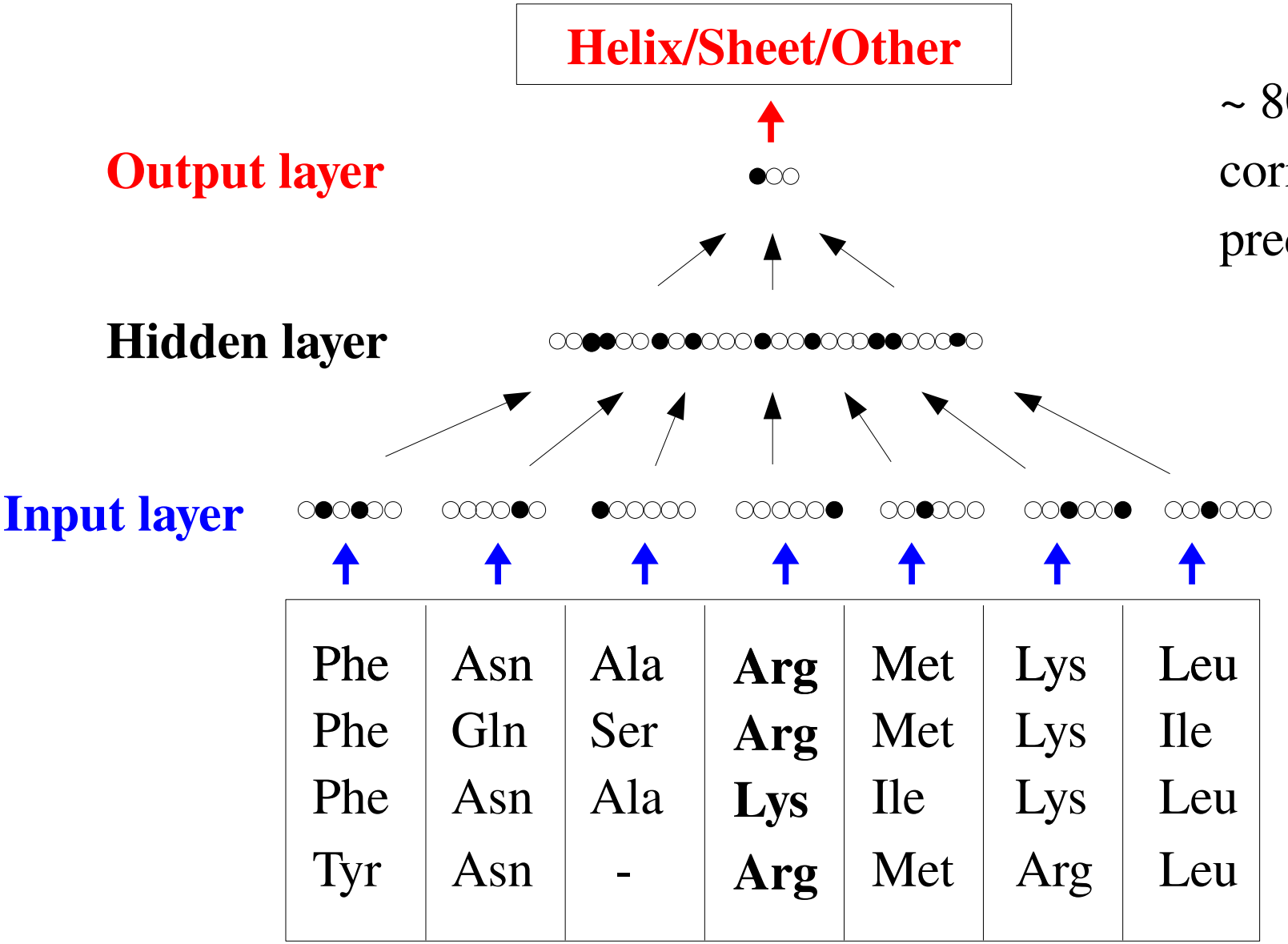


A neural network to predict secondary structure



Network architecture is chosen; the connection weights w_{ij} (1212 in this example) are chosen to give correct predictions for known structures (learning phase). At this point, the network has "learned" to assign a secondary structure to any input sequence. Specifically, it predicts the state of the central residue (**Arg**).

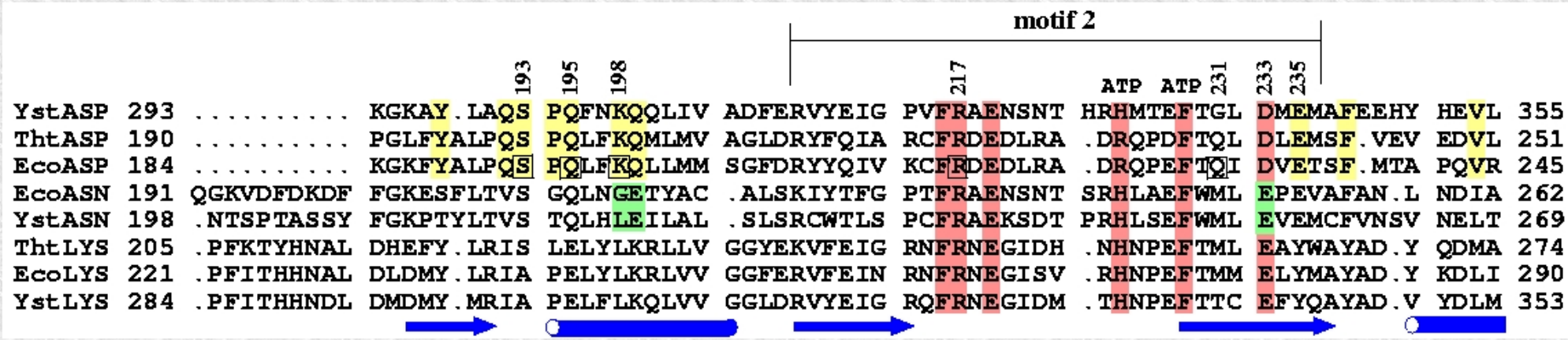
A neural network to predict secondary structure



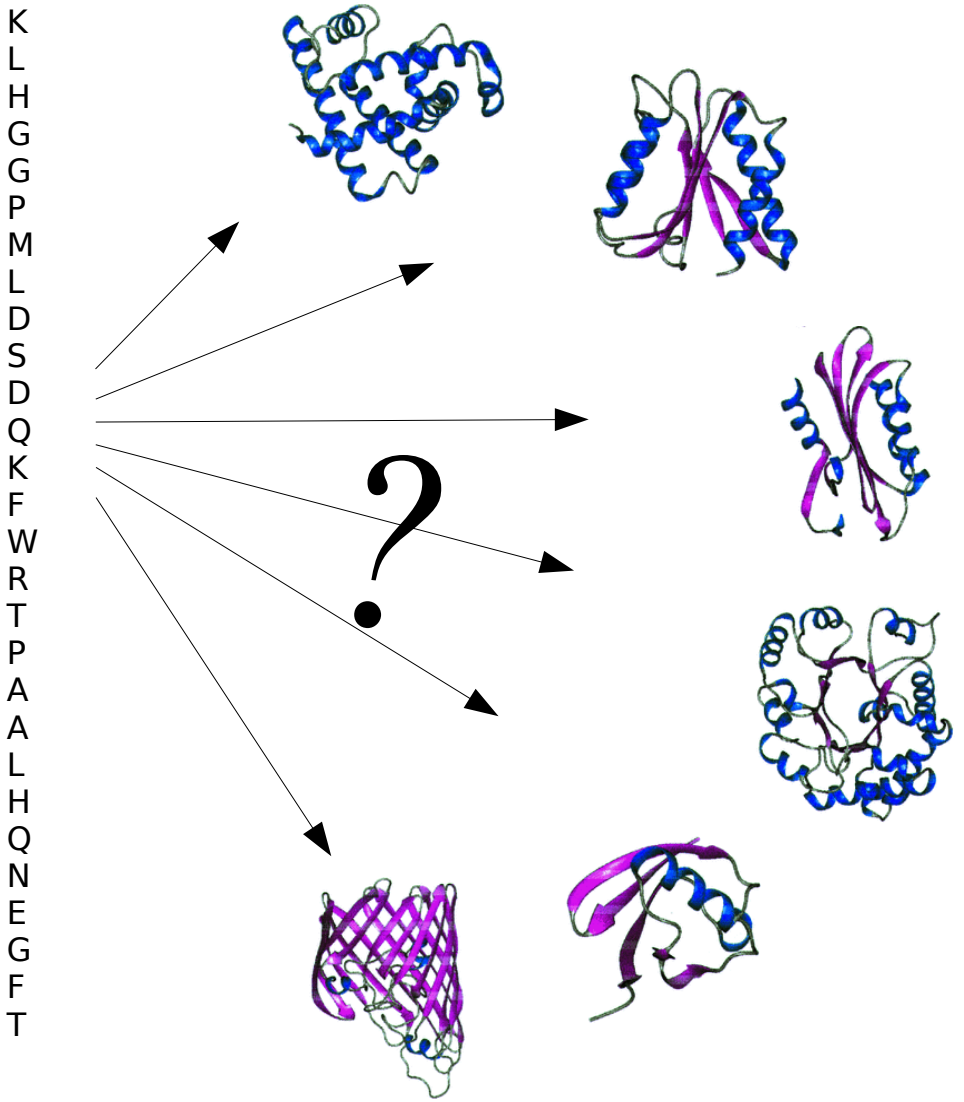
~ 80% correct predictions

Using a multiple alignment as input

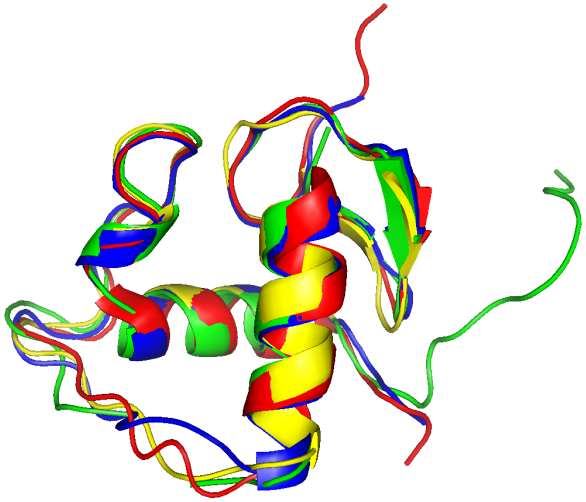
Homology modeling: an overview



Homology-based structure prediction: start from the sequence of a new protein. Check if it has a known fold through sequence comparaisons



If so, perform “Homology modelling”



Identify and align homologues of known structure
(“templates”)

Do multiple sequence alignment with many homologues

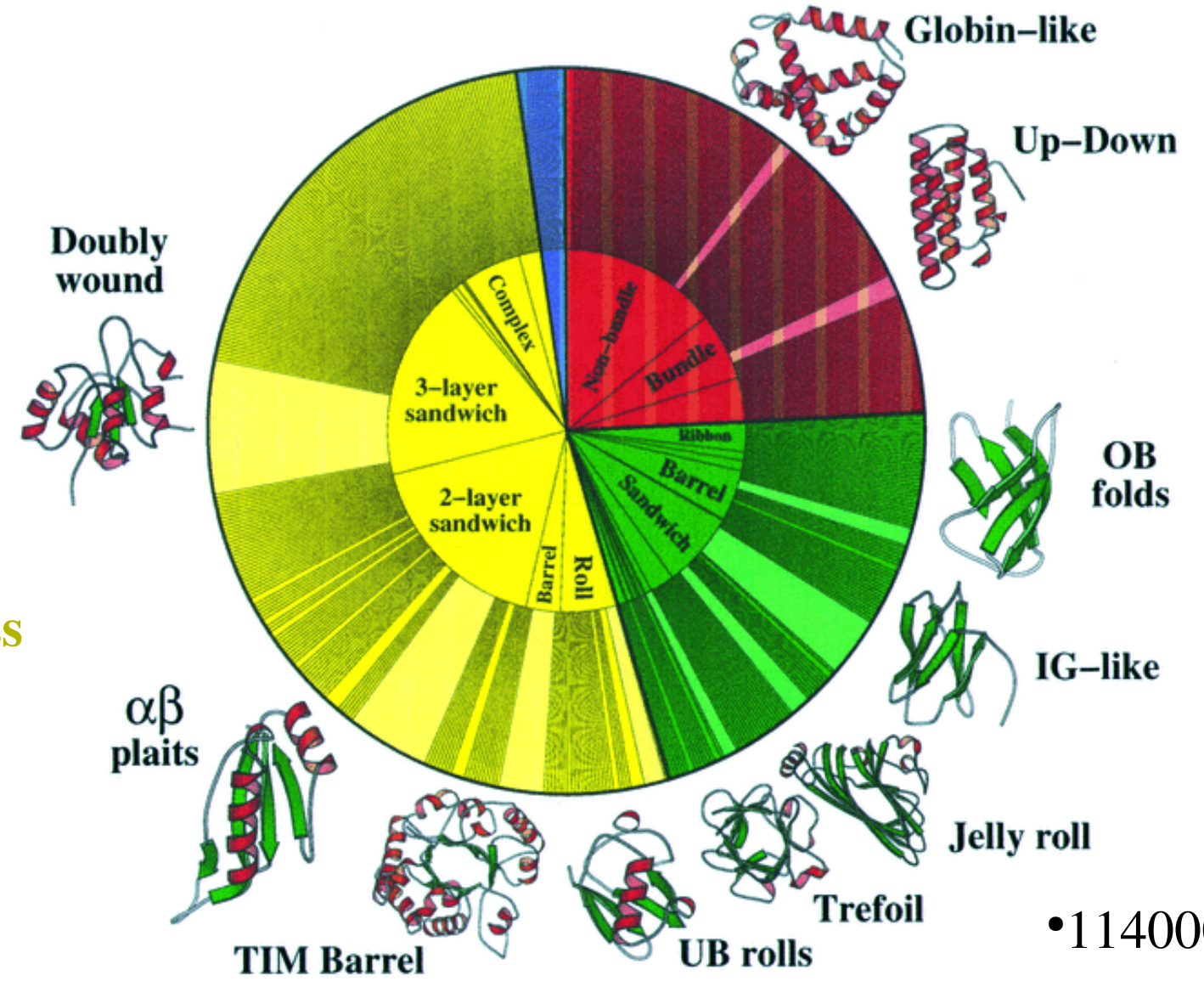
Conserved, well-structured regions: adopt mean backbone of templates

Loops, insertions: - search Protein Data Bank for similar fragments
- if not, use “ab initio” loop modelling methods

Sidechains: explore and score possible rotamers, possibly with a
stochastic approach such as “Monte Carlo”
Model refinement: energy minimization, molecular dynamics

Experimental testing

Fold set is discrete and rather small



- 114000 structures
- **2200 superfamilies**
- 100 contain 60% of all domains

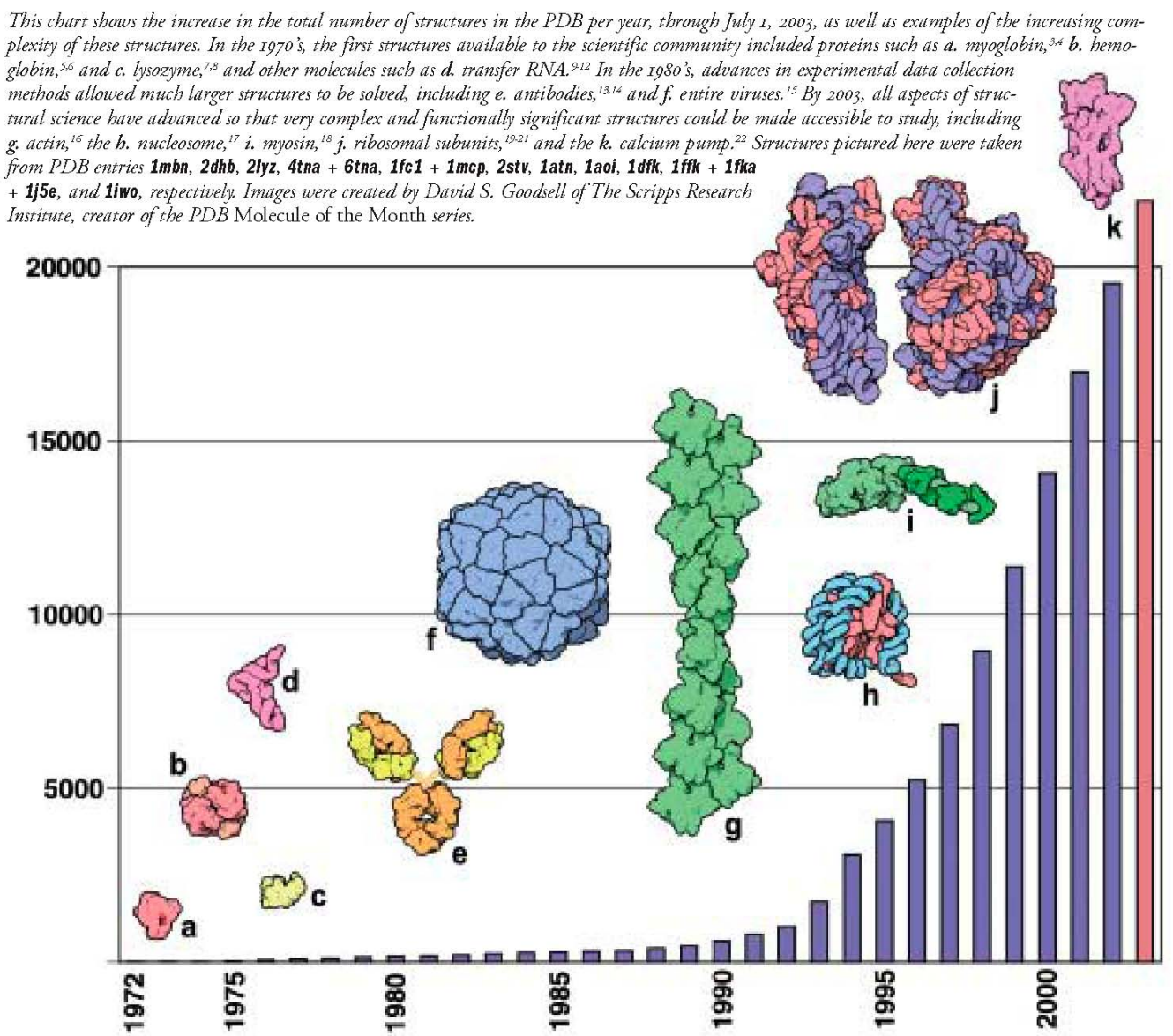
www.cathdb.info

Homologues of androsterone receptor with BLAST

#	ID Swissprot	Hit	Description	Score (bits)	E *	% Identity	Match Length
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor (PR)	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor (PR)	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor (MR)	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor (GR)	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta (ER-beta)	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor (ER-alpha)	98	4e-21	55	72
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47
351	P20659	TRX_DROME	Trithorax protein.	31	0.74	26	49
355	P98164	LRP2_HUMAN	Lipoprotein receptor.	30	1.7	27	65

*E = expectation of number of random alignments with a higher score

Structures in “Protein Data Bank”



This chart shows the increase in the total number of structures in the PDB per year, through July 1, 2003, as well as examples of the increasing complexity of these structures. In the 1970's, the first structures available to the scientific community included proteins such as **a.** *myoglobin,[3,4]* **b.** *hemoglobin,[5,6] and* **c.** *lysozyme,[7,8] and other molecules such as* **d.** *transfer RNA.[9-12] In the 1980's, advances in experimental data collection methods allowed much larger structures to be solved, including* **e.** *antibodies,[13,14] and* **f.** *entire viruses.[15] By 2003, all aspects of structural science have advanced so that very complex and functionally significant structures could be made accessible to study, including* **g.** *actin,[16] the* **h.** *nucleosome,[17]* **i.** *myosin,[18]* **j.** *ribosomal subunits,[19-21] and the* **k.** *calcium pump.[22] Structures pictured here were taken from PDB entries* **1mbn**, **2dhb**, **2lyz**, **4tna** + **6tna**, **1fc1** + **1mcp**, **2stv**, **1atn**, **1aoi**, **1dfk**, **1ffk** + **1fka** + **1j5e**, *and* **1iwo**, *respectively. Images were created by David S. Goodsell of The Scripps Research Institute, creator of the PDB* Molecule of the Month *series.*

Total structures in PDB

88000 structures in 2012

Non-redundant subsets

PDB95

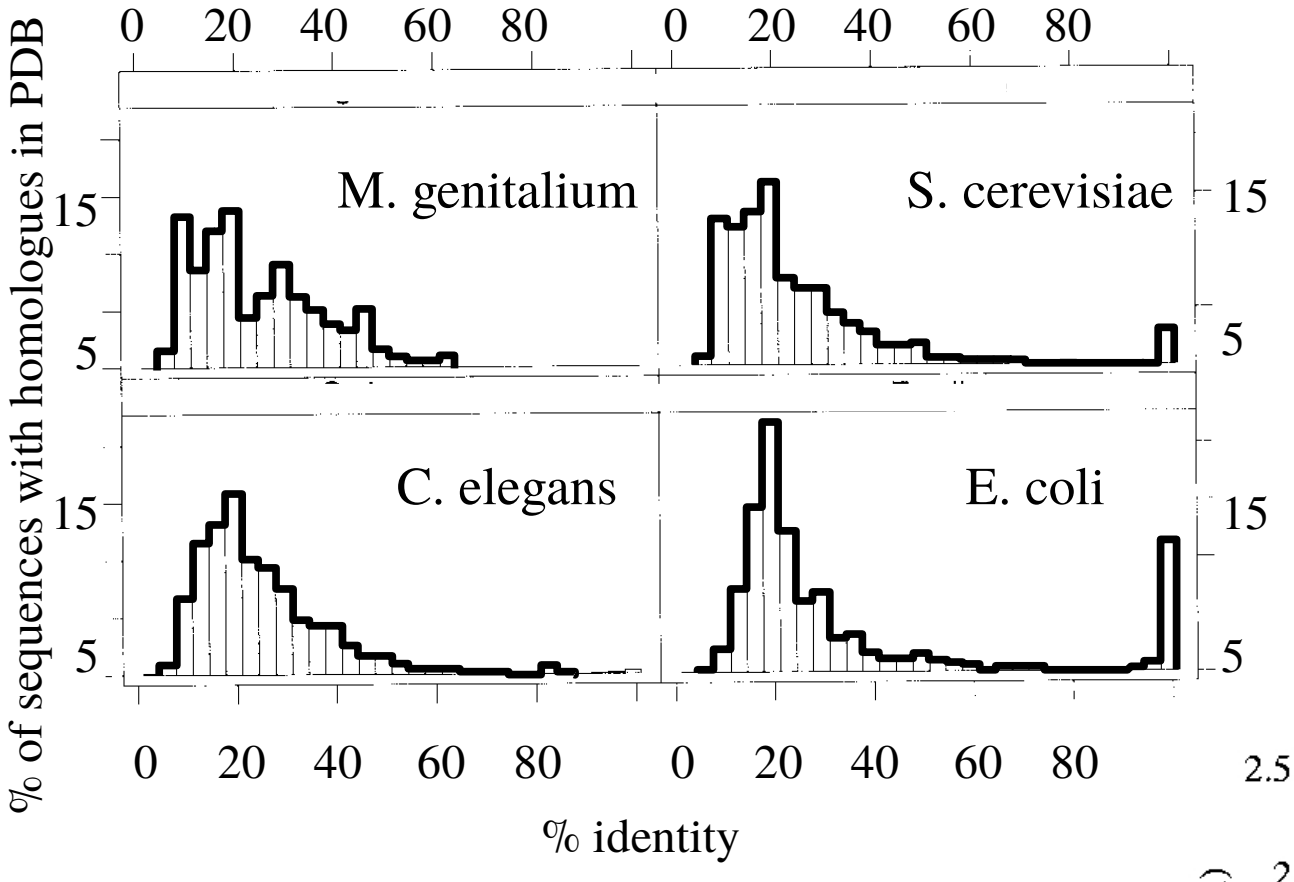
~35000

PDB25

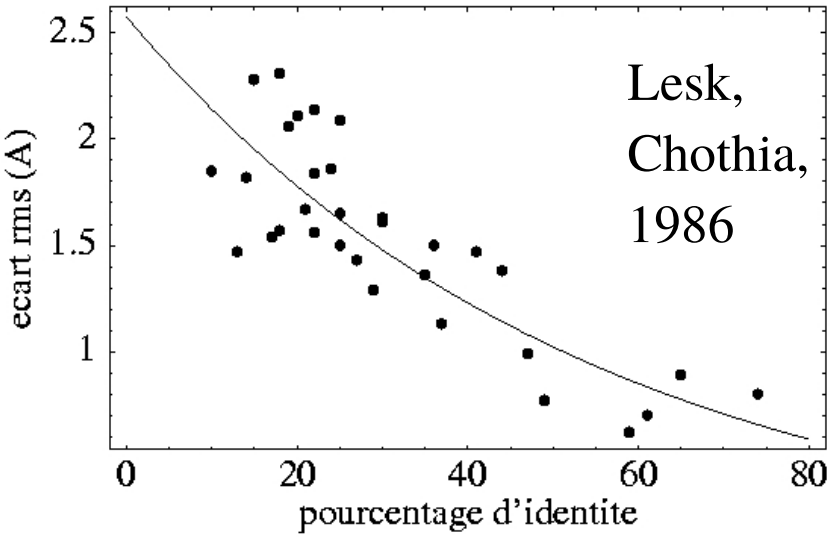
~4000 structures en 2012

année

Homology modelling: a template is not always available



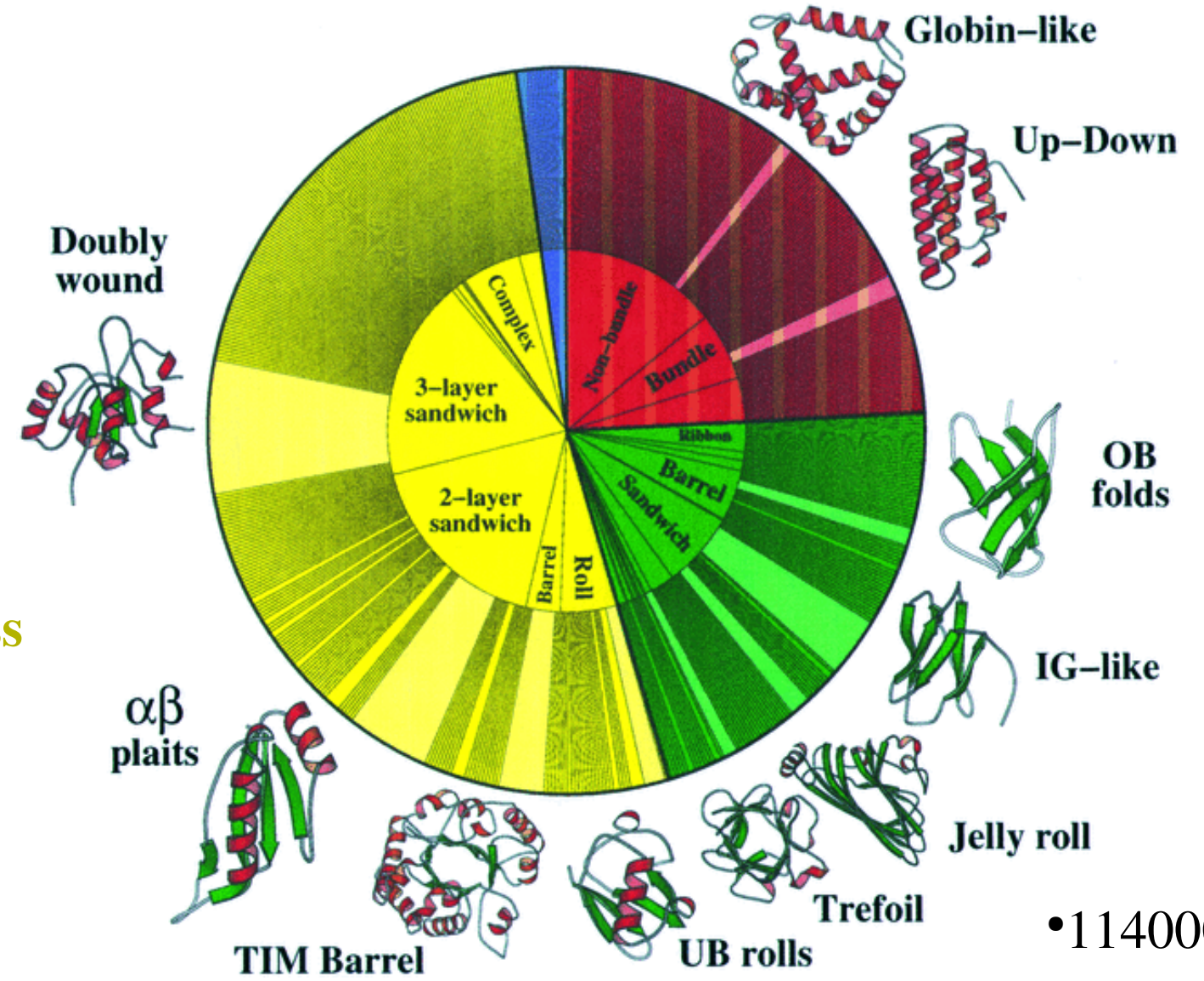
Distribution of sequence identities in Protein Data Bank



Lesk, Chothia, 1986

A template should have > 25% sequence identity with the modelled protein

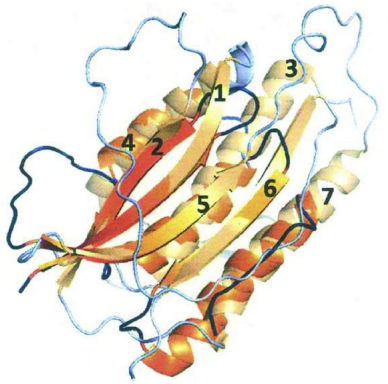
Fold set is discrete and rather small

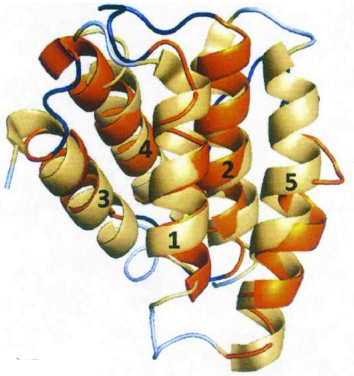


- 114000 structures
- **2200 superfamilies**
- 100 contain 60% of all domains

www.cathdb.info

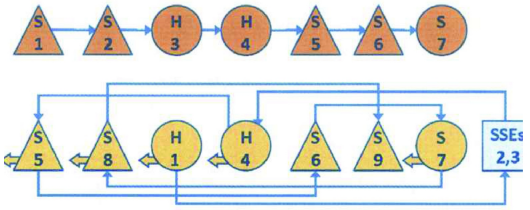
“New” folds aren’t always new....

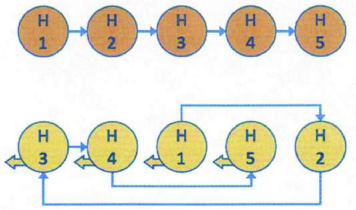




New fold

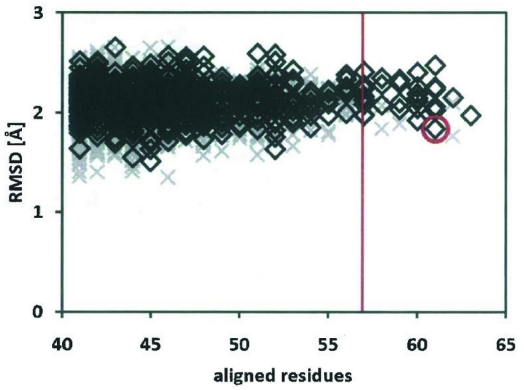
Similar fold

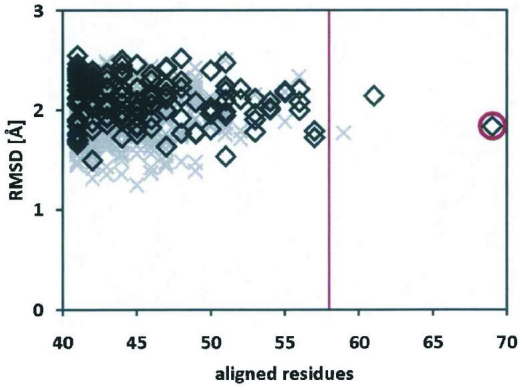




“topologic” view

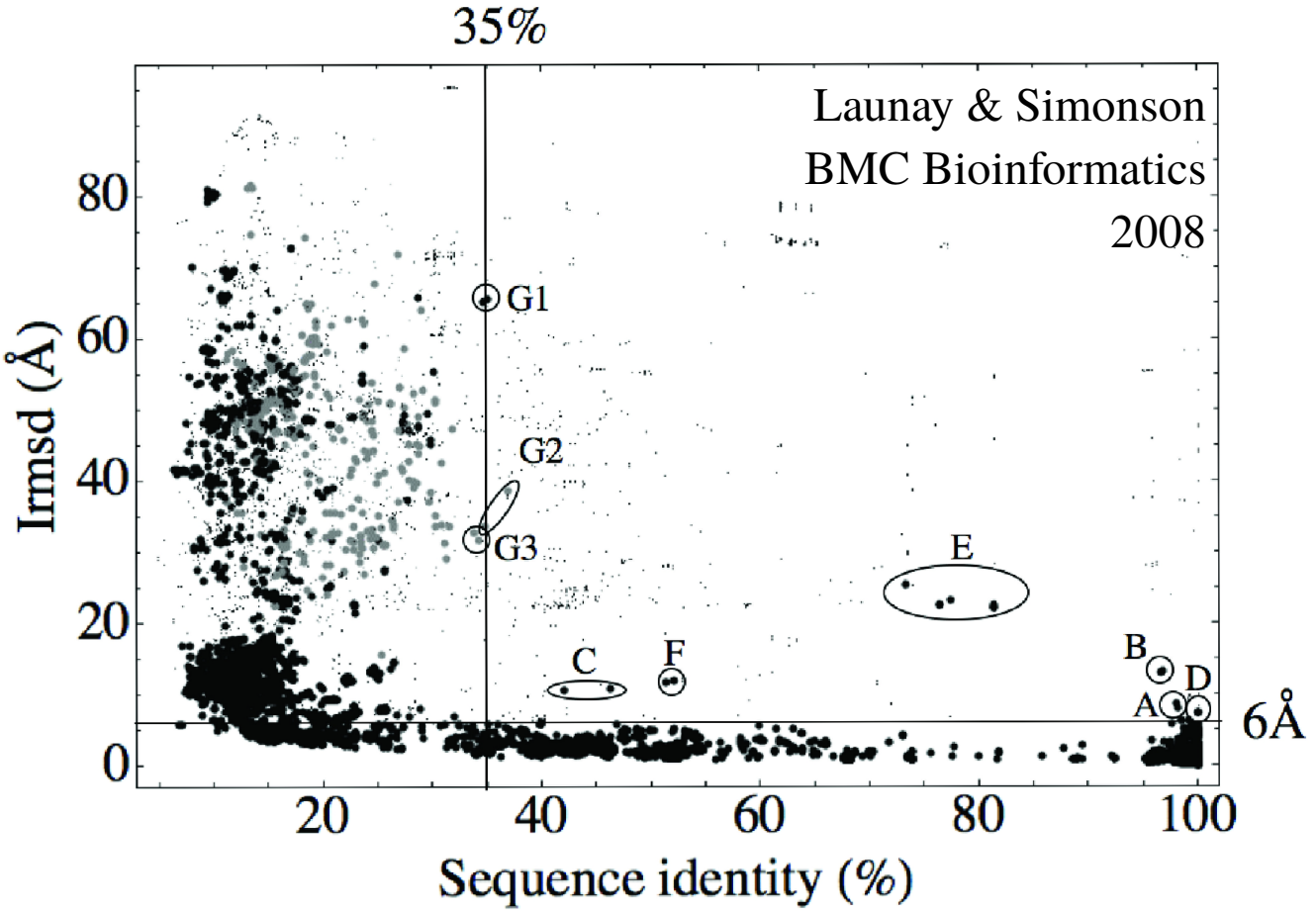
Alignment with known domains





Guerleur et al (2008)
Prot Sci, 17:1374

Homology of protein:protein interactions: 35% threshold



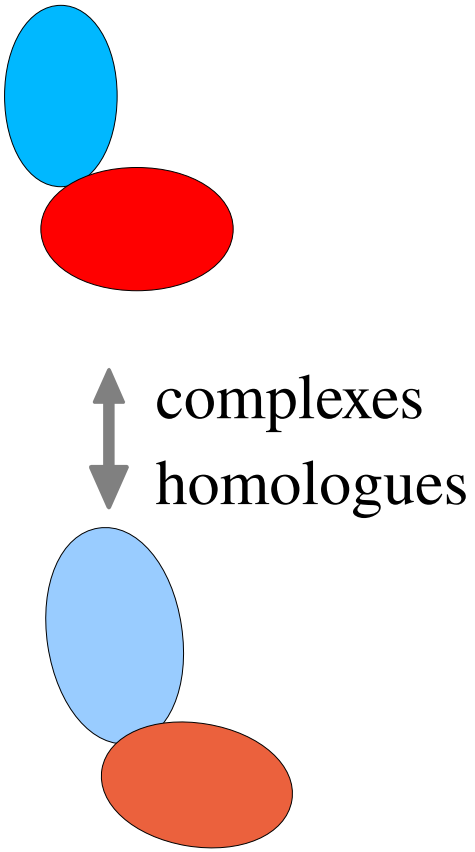
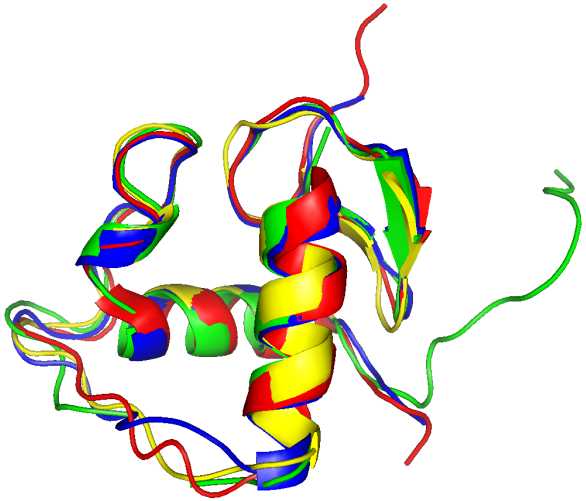


Figure 1
The relationship between sequence and structural similarity. 743 complexes from 40 interacting superfamily groups (ISGs) were analyzed. All pairs within each ISG were compared, for a total of 9630 pairwise comparisons. Small points correspond to comparisons involving at least one complex with either a small interface (buried area < 600 Å^2) or a weak association energy (E_{int} > -10 kcal/mol; see text). Points labelled A-G are discussed in the text. The horizontal line corresponds to Irmsd = 6 Å; the vertical line corresponds to a 35% sequence identity. Gray points correspond to comparisons where the MATRAS structural alignment provided fewer than 80% of the equivalent residues used for the Irmsd calculation. All the gray points lie below the 35% similarity threshold.

Perform “Homology modelling”



Identify and align homologues of known structure
(“templates”)

Do multiple sequence alignment with many homologues

Conserved, well-structured regions: adopt mean backbone of templates

Loops, insertions: - search Protein Data Bank for similar fragments
- if not, use “ab initio” loop modelling methods

Sidechains: explore and score possible rotamers, possibly with a
stochastic approach such as “Monte Carlo”
Model refinement: energy minimization, molecular dynamics

Experimental testing

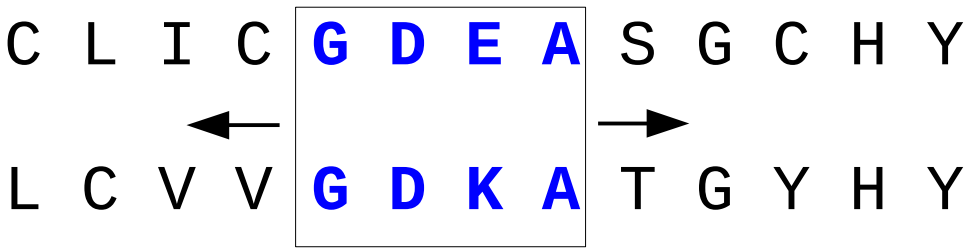
Homologues of androsterone receptor with BLAST

#	ID Swissprot	Hit	Description	Score (bits)	E *	% Identity	Match Length
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor (PR)	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor (PR)	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor (MR)	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor (GR)	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta (ER-beta)	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor (ER-alpha)	98	4e-21	55	72
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47
351	P20659	TRX_DROME	Trithorax protein.	31	0.74	26	49
355	P98164	LRP2_HUMAN	Lipoprotein receptor.	30	1.7	27	65

*E = expectation of number of random alignments with a higher score

BLAST: *Basic Local Alignment Search Tool*

- Find homologous tetrapeptides
- Extend as long as similarity > threshold



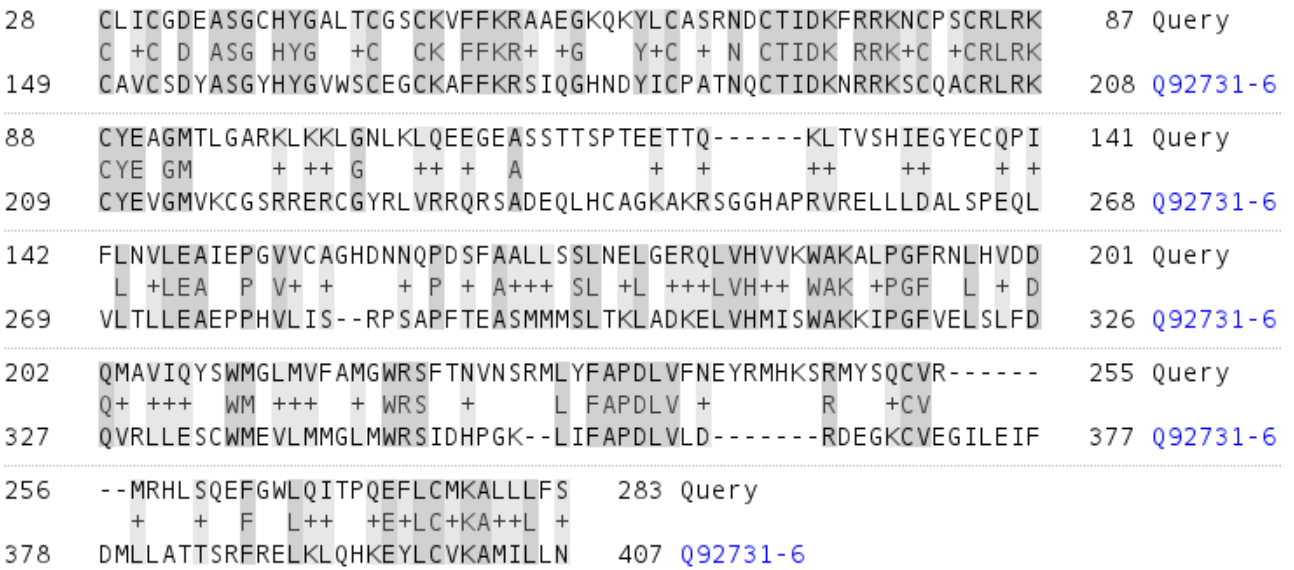
Récepteur de l'androstérone

Récepteur de l'hormone thyroïdienne

Blast output

Query = Human androgen receptor

Hit = Human Estrogen receptor beta



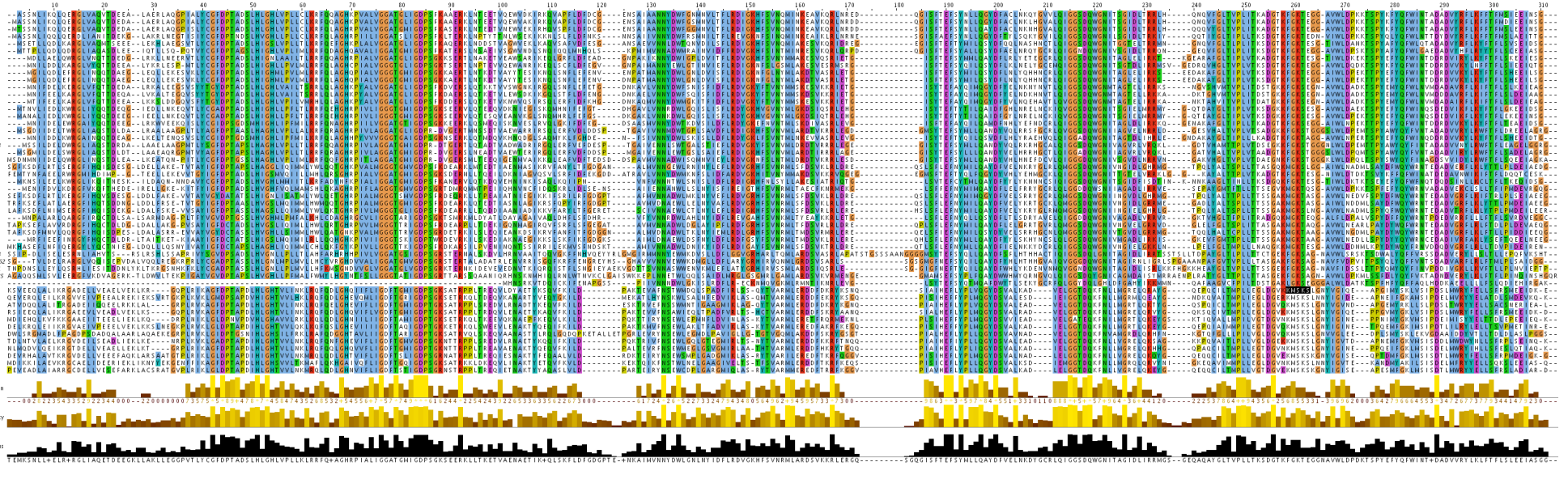
Similarité en gris

```
Job information
Query sequence
>P10275
   etc
Date of job execution  Dec 1, 2010
Running time  23.3 seconds
Program  blastp (BLASTP 2.2.23 [Feb-03-2010])
Database  uniprotkb (Protein) generated for BLAST on Nov 2, 2010
Sequences  12,898,884 sequences consisting of 4,176,319,342 letters
Matrix blosum62   Threshold  0.001  Gapped true
```

Heuristic alignment method: 3 phases

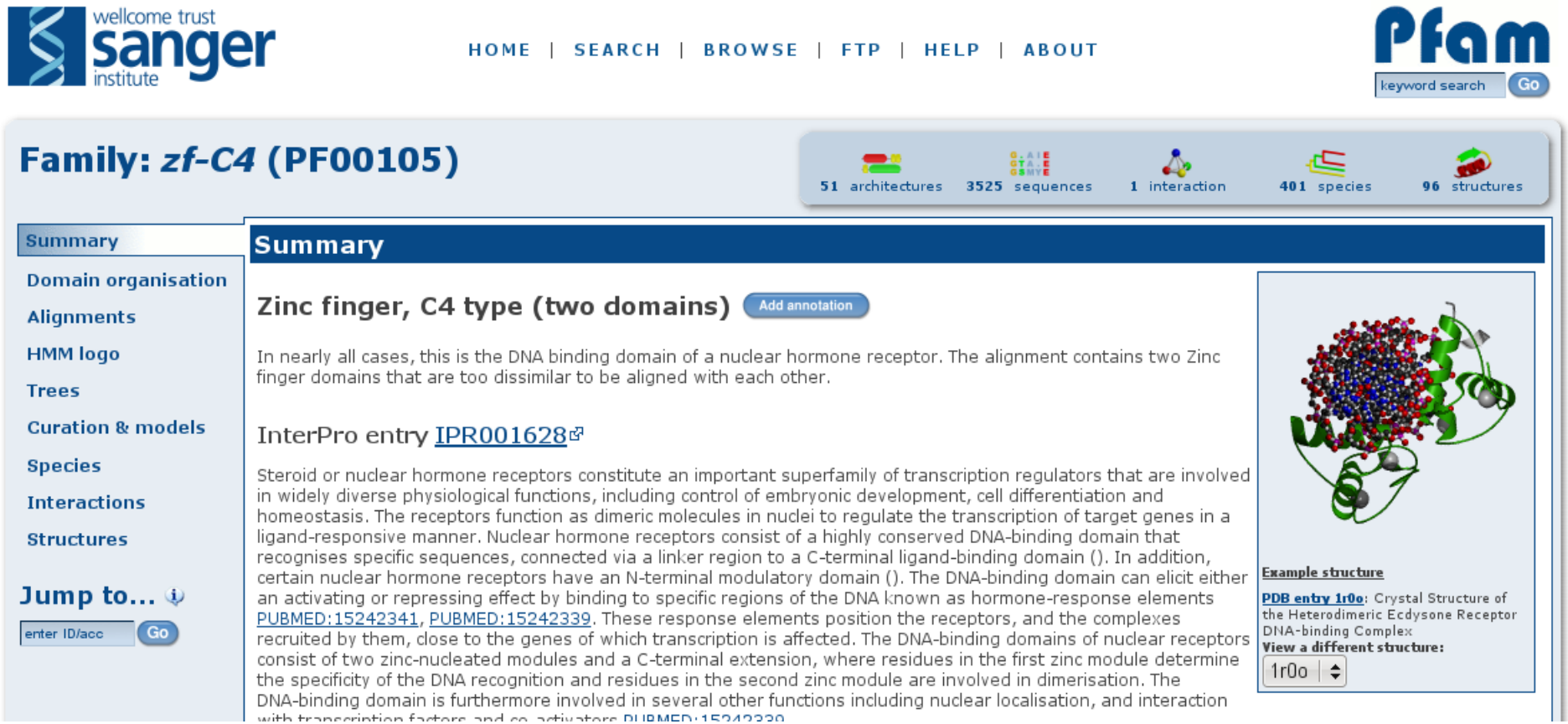
- Align sequence pairs
- Construct “guide” tree
- Progressive sequence-profile alignment

Alignment of 30 tyrosyl-tRNA synthetases, catalytic domain



Alignment libraries

http://pfam.sanger.ac.uk **Pfam** 24.0 (October 2009, 11912 families)



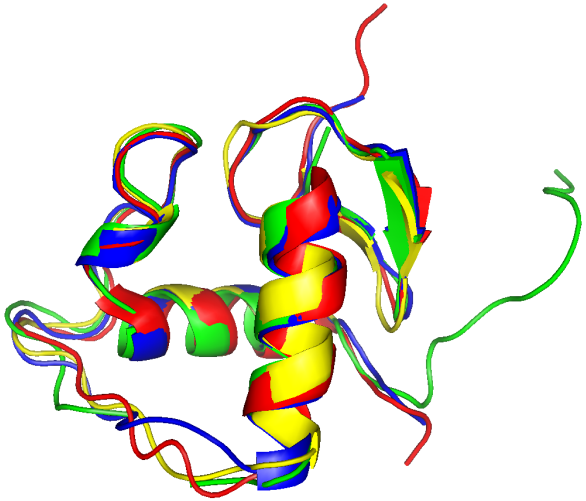
Famille zf-C4 → “expert” alignment “expert” of 26 sequences; “automatic” of 3525 séquences

---

Balibase: 217 expert alignments benchmark set to tester methods/programs

http://www-bio3d-igbmc.u-strasbg.fr/balibase Thompson et al (1999) Bioinformatics, 57, 87-88

Perform “Homology modelling”



Identify and align homologues of known structure
(“templates”)

Do multiple sequence alignment with many homologues

Structural alignment with available homologues

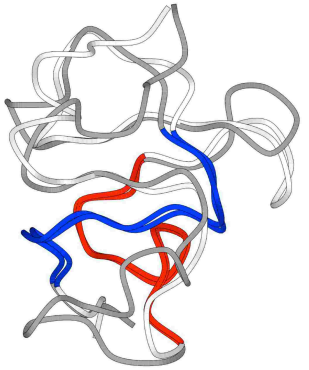
Loops, insertions: - search Protein Data Bank for similar fragments
- if not, use “ab initio” loop modelling methods

Sidechains: explore and score possible rotamers, possibly with a stochastic approach such as “Monte Carlo”

Model refinement: energy minimization, molecular dynamics

Experimental testing

Remarks on structural alignments



1) Differ from sequence alignments: an amino acid is not coded by a letter

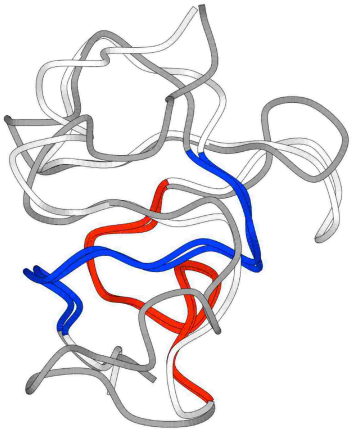
2) A “trivial”method is possible: align sequences then superimpose homologous C_α atoms

3) We can align complex objects, like alignments... Code an amino acid by a vector of properties: type, secondary structure, burial, ...

4) In general, various heuristic approaches are used.

5) For multiple alignments, use a progressive method.

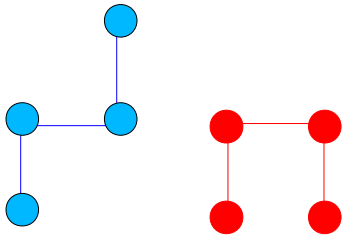
Remarks on structural alignments



STRUCTAL: a) align sequences; **b)** *superimpose homologous* Cα atoms;
c) compute scoring matrix; **d)** repeat a-c.
Score M(i,j) = -distance between i, j (Levitt, 1993)

SSAP: describe each amino acid by a vector of properties;
Align with Needleman-Wunsch (Taylor & Orengo, 1989)

CE: find two similar octapeptide; extend with gaps. (Bourne, 1998)

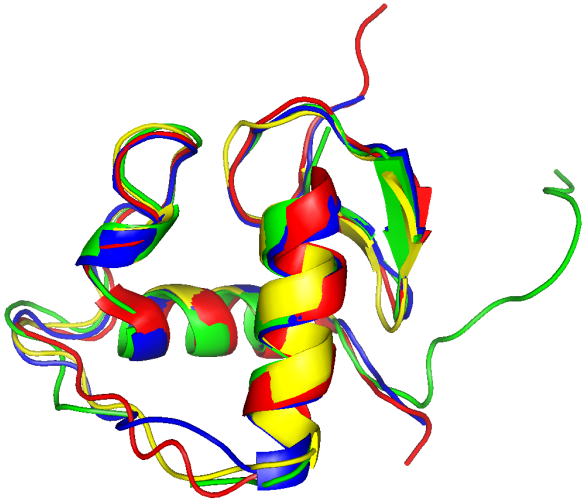


DALI: similar to CE (Holm & Sander, 1993).

Large-scale test: Kolodny et al (2005) J Mol Biol, 346:1173

Over 80 methods in Wikipedia...

Perform “Homology modelling”



Identify and align homologues of known structure (“templates”)

Do multiple sequence alignment with many homologues

Structural alignment with available homologues

Loops modeling methods

Side chains: explore and score possible rotamers

Model refinement: energy minimization, molecular dynamics

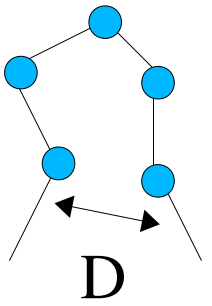
Experimental testing

Loop modeling with an experimental conformation library

Wojcik et al (1999) J Mol Biol, 289:1469

13,563 loops of 3-8 amino acids from the PDB

Ranked by:
-length n
-distance D between ends
-"height" and "width"



n = 5

To model a new loop X of length n:
-estimate D
-find loops in library compatible with n, D;
-find those with sequence similar to X
-refine with molecular dynamics

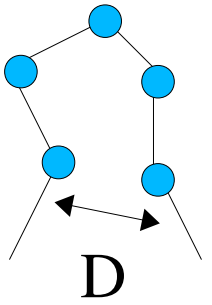
What are the main hypotheses of the method?

Loop modeling with an experimental conformation library

Wojcik et al (1999) J Mol Biol, 289:1469

13,563 loops of 3-8 amino acids from the PDB

Ranked by:
-length n
-distance D between ends
-"height" and "width"

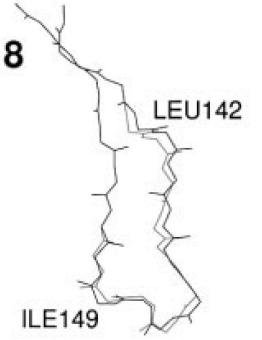


n = 5

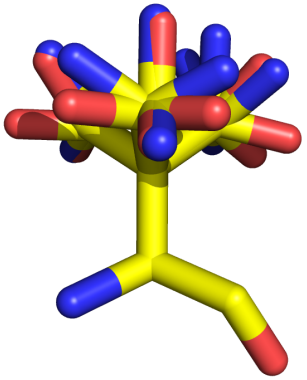
To model a new loop X of length n:
-estimate D
-find loops in library compatible with n, D;
-find those with sequence similar to X
-refine with molecular dynamics

Quality of results

Longueur de boucle	3	4	5	6	7	8
Nombre de cas testés	92	81	54	61	37	37
Rmsd des C_α (A)	1.0	1.2	1.6	2.1	2.7	2.3
Avec Modeller:		0.6				1.2



Side chain placement by simulated annealing



Side chains explore rotamers.

Explore with **Monte Carlo**:

Many elementary steps (10^6).

At each step:

-pick an amino acid randomly

-pick a new rotamer randomly

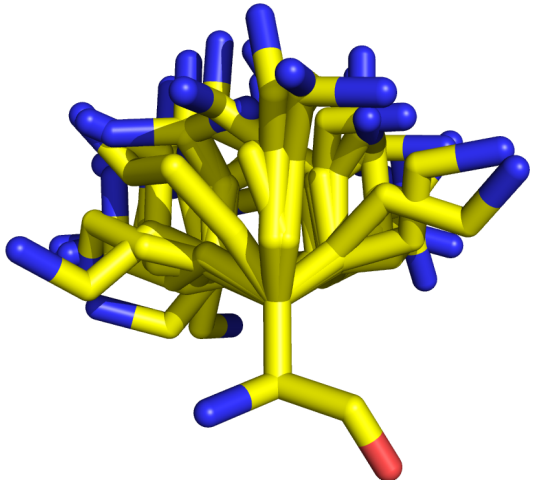
-compare new energy E_n to previous energy E_p

-if $E_n < E_p$, keep the new rotamer

-if $E_n > E_p$,pick a random number between 0 and 1

-if $\exp[-(E_n - E_p)/RT] > x$, keep the new rotamer

-else keep the previous one



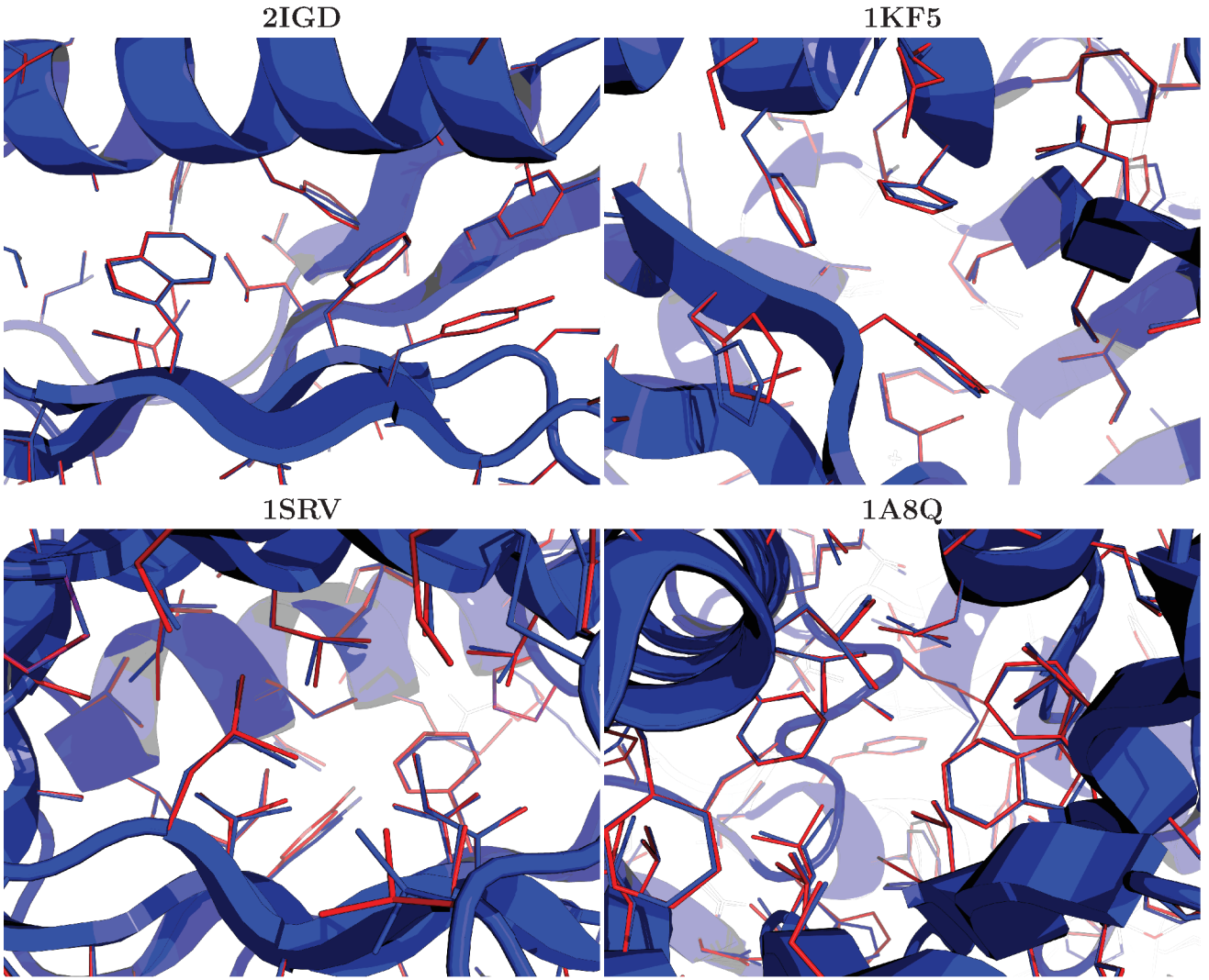
Start at high temperature T and lower T progressively.

Remarkably, the method explores conformations with probability $p(Conf) = A \exp(-E_{Conf}/RT)$, the experimental, Boltzmann distribution.

Prediction exercice: delete, then reconstruct side chains; main chain known exactly.

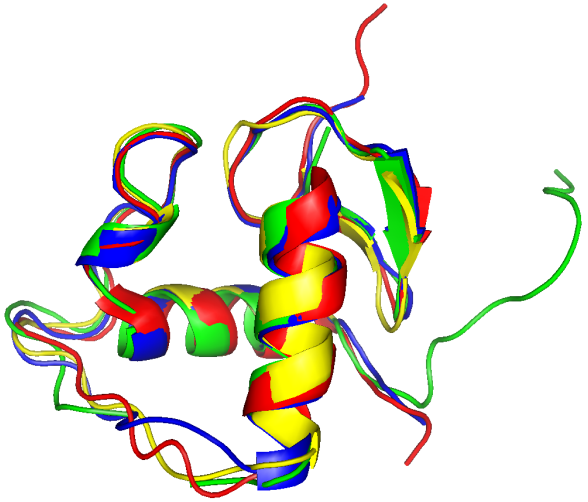
86% chi1 correct
73% chi1 et chi2

With perfect backbone, side chain predictions are good but not perfect.



Gaillard, Panel, Simonson (2016) Proteins

Perform “Homology modelling”



Identify and align homologues of known structure (“templates”)

Do multiple sequence alignment with many homologues

Structural alignment with available homologues

Loops modeling methods

Side chains: explore and score possible rotamers

Model refinement: energy minimization, molecular dynamics

Experimental testing

Performance of selected methods

	% correct predictions		
	chi1	chi2+chi1	
Gaillard et al, 2016	86%	73%	
Koehl & Delarue, 1994	72%	62%	
Yang et al, 2002	80%	66%	
Dunbrack et al, 1999	80%	?	
Abagyan et al, 1993	80%	67%	
Mendes et al, 1999	87%		(improved rotamer library

In these tests, the side chains were placed on the known backbone...

In homology modeling, the backbone is not exact...

Homology modeling. Krieger E, Nabuurs SB, Vriend G.
In Structural Bioinformatics; editors PE Bourne, H Weissig; Wiley, 2003

Homology modelling in biology and medecine. R Dunbrack.
In Bioinformatics: from genomes to drugs; editor T Lengauer; Wiley, 2002; Vol. 1.

Structure comparison and alignment. Bourne & Shindyalov
In Structural Bioinformatics; editors PE Bourne, H Weissig; Wiley, 2003

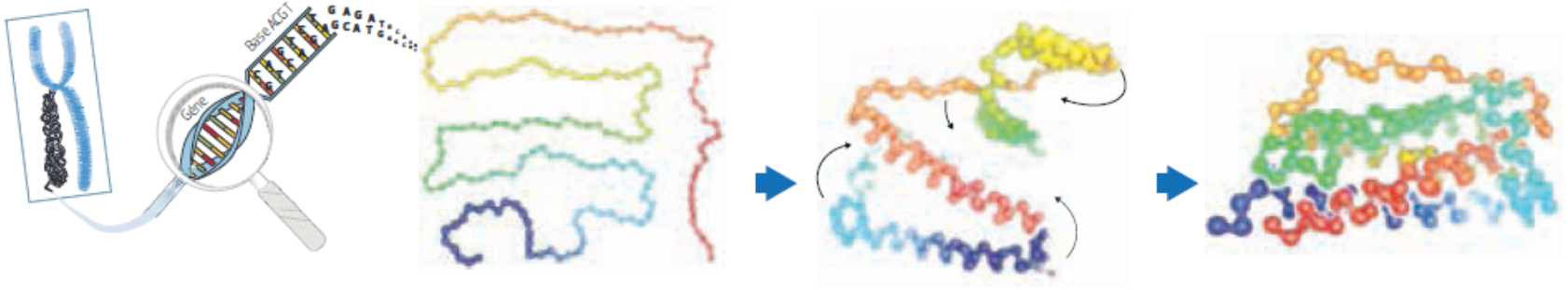
Proteins: Structure, Function, and Bioinformatics; Volume 77, Issue S9, pages 1-228.
Numéro spécial décrivant la compétition CASP8:

"Critical Assessment of Structure Prediction"

64 cibles à prédire; 159 méthodes/équipes

Le secret du pliage des protéines dévoilé

Une intelligence artificielle de Google arrive à prédire la structure 3D d'à peu près n'importe quelle molécule codée par le génome humain. Une avancée majeure qui ouvre un champ des possibles considérable.



(1) Chaque chromosome contient une multitude de bases (A, C, G ou T). Combinées entre elles, ces bases forment le **code génétique.**

(2) Le code génétique permet la constitution de longues **chaînes d'acides aminés...**

(3) ... qui se replient naturellement sur elles-mêmes pour constituer la **protéine** et lui donner sa **forme spécifique.**

(4) Grâce à l'intelligence artificielle Alpha Fold, les chercheurs peuvent désormais **décoder ce « schéma de pliage »** et donc prédire la structure 3D de la protéine.

Infographie LE FIGARO

STÉPHANY GARDIER @S_Gardier

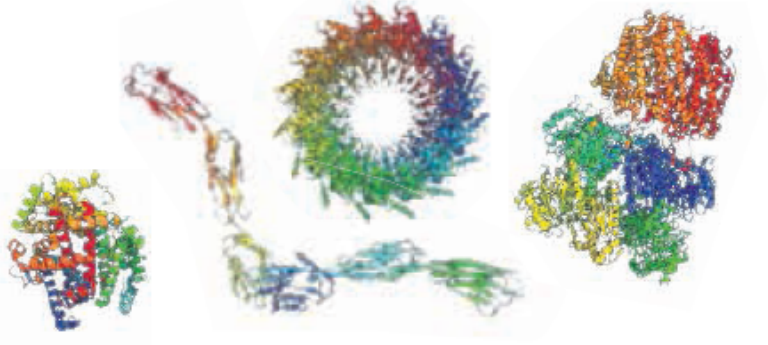
BIOLOGIE *« Incroyable », « révolutionnaire », « bluffant »...* Quand ils évoquent la plateforme AlphaFold, dévoilée il y a quelques jours par DeepMind (une filiale d'Alphabet, maison mère de Google), les spécialistes ne manquent pas de superlatifs. Cet outil d'intelligence artificielle est capable de prédire la structure de la quasi-totalité des protéines codées par le génome humain. De quoi ouvrir de nouvelles perspectives notamment en médecine, la plupart des produits thérapeutiques ayant pour cible des protéines. Après la publication du génome humain en 2003, la mise à disposition de la quasi-totalité de ce « protéome » humain pourrait marquer un nouveau tournant pour la recherche biomédicale.

> **« Pouvoir prédire la structure d'une protéine offre un gain de temps considérable »**
> THOMAS SIMONSON, DIRECTEUR DE RECHERCHE SPÉCIALISÉ EN BIOLOGIE DES PROTÉINES À L'ÉCOLE POLYTECHNIQUE (PALAISEAU)

Les protéines sont des molécules présentes dans toutes les cellules grâce à une machinerie (elle-même faite de protéines !) capable de traduire le code génétique en acides aminés. Ces petites briques élémentaires sont assemblées de manière linéaire un peu comme les perles d'un collier. Sous l'effet de forces qui s'exercent entre ces perles, la protéine va se replier pour former une structure en trois dimensions. Et c'est le résultat de cet origami de précision qui détermine les fonctionnalités de la protéine, dont la forme globale joue un rôle crucial dans son fonctionnement. *« Pouvoir prédire la structure d'une protéine offre un gain de temps considérable, en permettant de cibler les domaines qui ont un intérêt particulier et sur lesquels il faut réaliser des vérifications expérimentales »*, souligne Thomas Simonson, directeur de recherche spécialisé en biologie des protéines à l'École polytechnique (Palaiseau), qui imagine déjà comment AlphaFold va accélérer le rythme des travaux de son laboratoire.

Une grande diversité de structures possibles

Une fois repliées sur elles-mêmes, les chaînes d'acides aminés peuvent former toutes sortes de **protéines**, qui constituent la base du vivant.



Infographie LE FIGARO

Rapidité, simplicité et fiabilité sont trois des atouts d'AlphaFold, qui a supplanté tous ses concurrents lors de la dernière édition du CASP, la compétition internationale bisannuelle de prédiction de structure de protéines. *« Ce qui est encore plus surprenant, c'est que DeepMind ne s'est attaqué à cette thématique du repliement des protéines que récemment*, souligne Thomas Simonson. *On les a vus la première fois à la 13e édition en 2019, et deux ans plus tard, ils sont les premiers, loin devant les autres ! »* Développés majoritairement par des laboratoires académiques, les autres programmes n'ont pas fait le poids face à la puissance de calcul et aux investissements du nouveau venu. Interrogé sur le coût de ce projet, Demis Hassabis, le fondateur de DeepMind, n'a pas donné de chiffres, mais a souligné les énormes efforts consentis pour ce qu'il a présenté comme *« le plus gros investissement »* de sa société, qui a monopolisé une équipe de 30 personnes durant cinq ans. *« Il est évident qu'aucun laboratoire académique ne peut déployer autant de ressources pour tenter de coder un programme »*, commente Cécile Breyton.

Comme sa consœur, Alexis Verger, chercheur à l'Institut des sciences biologiques du CNRS, salue la décision de DeepMind d'avoir opté pour un accès libre à AlphaFold et d'avoir publié le code source du programme. *« Mais n'oublions pas que, sans le travail de tous les biologistes structuraux des cinquante dernières années, AlphaFold n'existerait pas, car ce sont leurs résultats qui ont servi à "entraîner" l'intelligence artificielle du programme. Et c'est grâce à nos futures données expérimentales que l'IA d'AlphaFold deviendra encore plus performante. C'est un excellent exemple du cercle vertueux des sciences ouvertes. »*

> **AlphaFold contient les structures des protéines de la bactérie de la tuberculose et de *Plasmodium falciparum*, responsable du paludisme**

AlphaFold devrait intéresser de nombreux laboratoires car, en plus

Article

Highly accurate protein structure prediction with AlphaFold

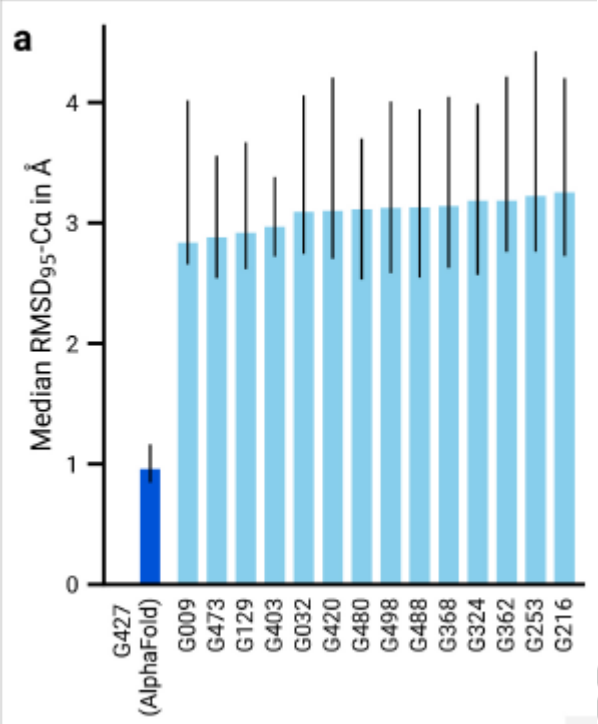
https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

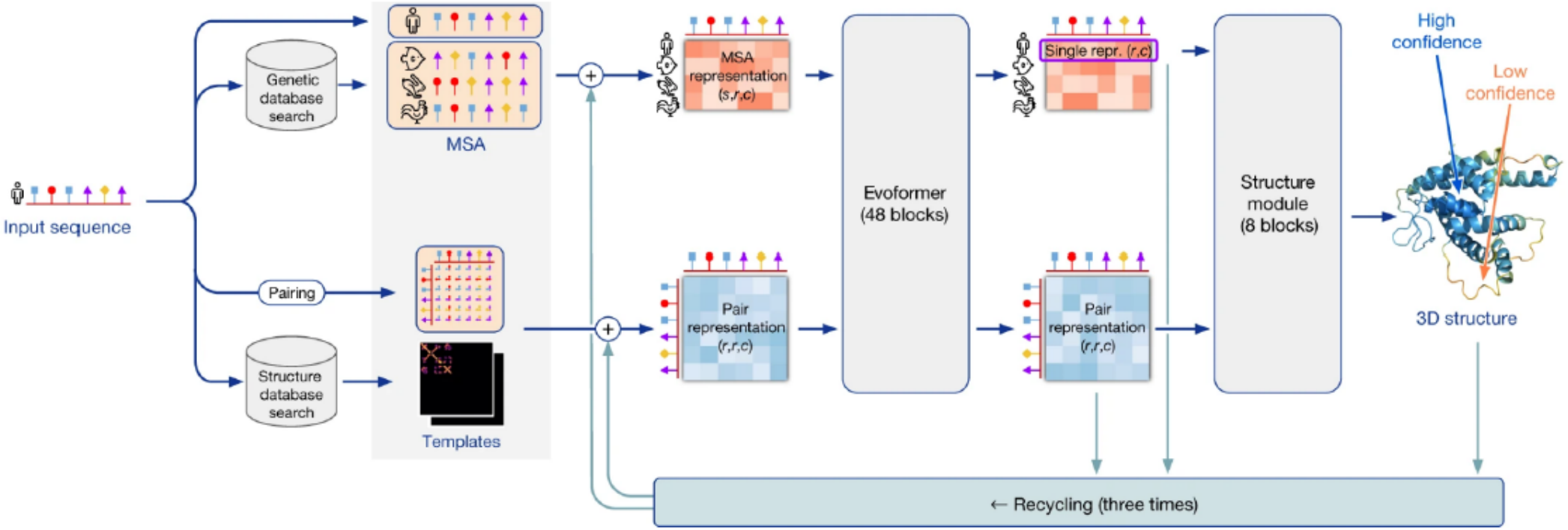
John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉



Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the 3-D structure that a protein will adopt based solely on its amino acid sequence, the structure prediction component of the 'protein folding problem'[8], has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even where no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experiment in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

Nature, 2021

AlphaFold: the general structure



search in databases

- for MSA construction
- template search

construction of the initial pair représentation

simultaneous optimization of the MSA representation and the pair representation thanks to a neural network

Determination of 3D position for each atom thanks to a neural network