Integration of Data

Bruno Kieffer (kieffer@igbmc.fr)
Groupe de RMN biologique
ESBS-IGBMC
Illkirch

Plan

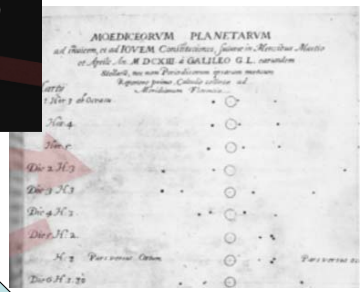
- Measurement and modelling: what do we mean by “model” ?
- Thermodynamics aspects of proteins
- Modelling protein structures from NMR data
- Modelling protein’s dynamics

General Path for Scientific Work

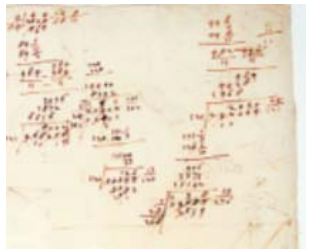
Phenomenon/system

Measure

Data



Interpretation
Computations

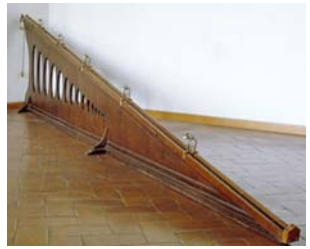


Model

From Delsuc MA
(Ecole de Physique des Houches 2015)

Model can be anything

- Analogic model
  exemple of Galileo ramp



- Analytical equation

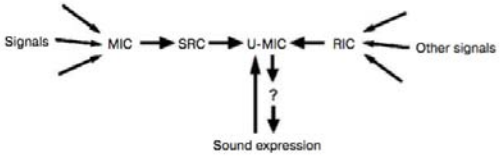
$$\Delta E = \hbar \gamma B_0$$

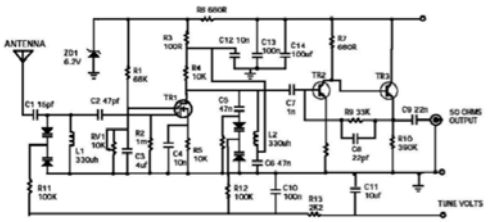
most of the physic we learn / we teach

- Image
  microscopy
- Computer program
  molecular modeling
  any kind of program modeling the system

Model depends on chosen experimental approach

Model A *Model B*





From Lazebnik BIOCHEMISTRY (Moscow) Vol. 69 No. 12 2004

From Data to Models: Two Opposite Ways

Top-Down	Bottom-Up
From Very Large sets of Data On Complex Systems	*From Sparse and accurate Data on Simple systems*
To Simple Predictive Models	*To Model Complex Systems*

Model depends on chosen experimental approach

Another way to say the same thing (Bayes theorem):

$$P(model\ /\ data) = \frac{P(model)}{P(data)} \times P(data\ /\ model)$$

Models are also involved in Measurements

- A **measure** results from a **signal** which has been processed

Transfer function

$$Y^{meas} = F_T(S) + err$$

measure *signal* *noise*

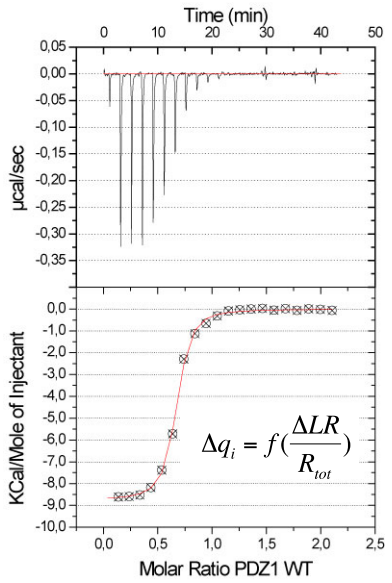
Number of points N P *Number of parameters to define the transfer function*

Models are also involved in Measurements

- N > P : Modelling the ***Phenomenon***
  - ***Fitting*** *the parameters onto the data*
  - *Exp: ITC*
- N = P : Modelling the ***Signal***
  - ***Transforms*** *(exp: Fourier Transform)*
- N < P : Modelling the ***Knowledge***
  - ***Reconstruction*** *of data*
  - *Exp: Structure determination by NMR*

Example of N>P : ITC Measurement

$\frac{dq_i}{dt}$



- Isothermal Titration Calorimetry

$$nL + R \underset{}{\overset{Ka}{\rightleftharpoons}} nLR$$

$$\Delta q_i = n[R]_{tot} V_{cell} \Delta H_{app} \left\{ \frac{[LR]_i}{[R]_{tot}} \right\}$$

- Fit with parameters: n, Ka, ΔH_{app}

→ ΔG ΔH -TΔS

Gibbs Free energy — enthalpy — entropy

Characteristic time of an observation

- All observation method is associated with a characteristic time
  - Interaction light-matter : very short times
    exp: X-FEL 10^{-14} s - 10^{-12} s
  - Measure of heat transfer : 10^{-1} s
- During the observation characteristic time, we measure an average:
  - If the molecules are static: $X = \langle x_i \rangle_{space}$
  - If the molecules are moving: $X = \langle x_i \rangle_{time}$

Ergodicity

- A system is ergodic when time and space averages gave identical results:

$$\langle x_i \rangle_{space} = \langle x_i \rangle_{time}$$

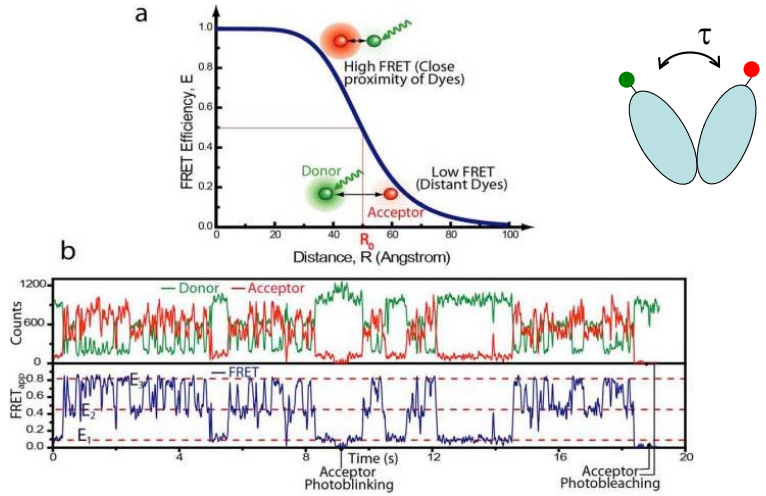
- Methods that are not sensitive (NMR, ITC):

$$X = \langle \langle x \rangle_{time} \rangle_{space}$$

- Integrating information from different biophysical methods requires the knowledge of their characteristic times

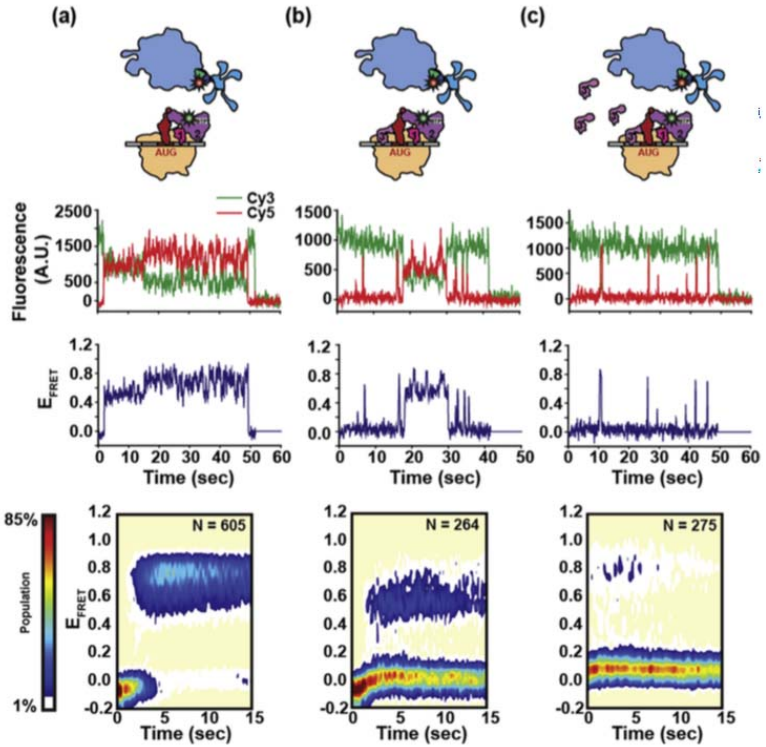
Pure time average methods

- smFRET provides pure time average observations



ReNaFobis 2016

From Nat Methods. 2008 Jun; 5(6): 507–516.

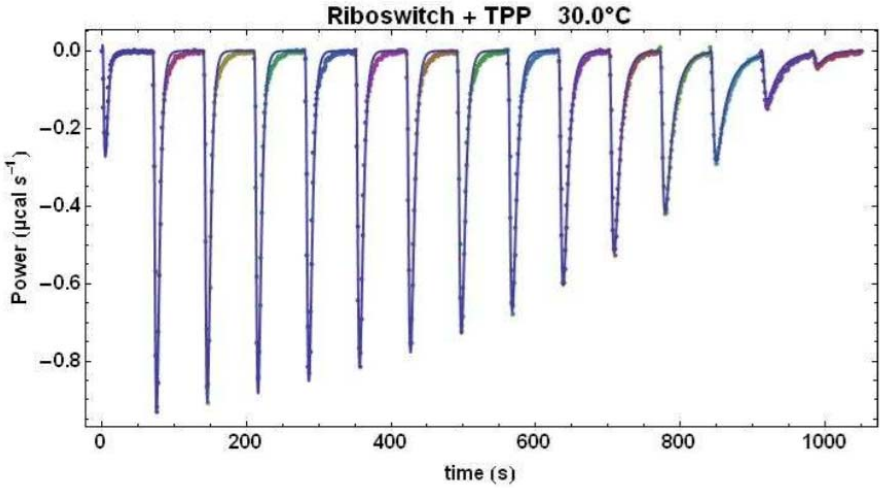


ReNaF

Review in K. Caban, R.L. Gonzalez Jr. / Biochimie 114 (2015) 30-38

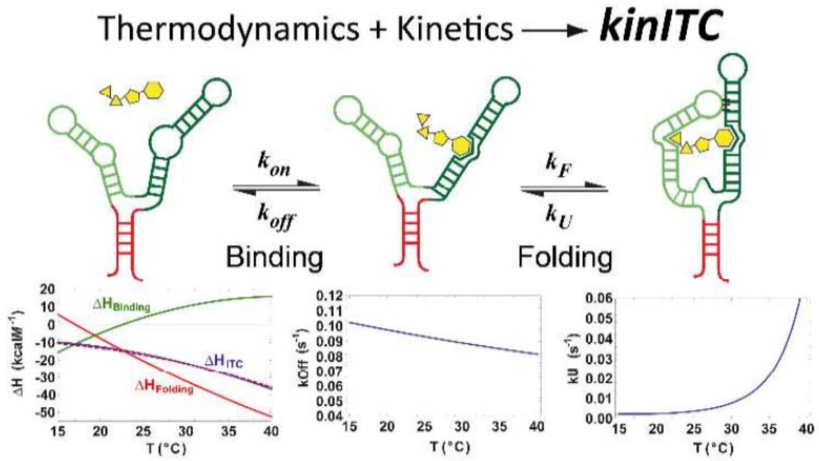
Example of a non-ergodic system

- ITC measurement of the heat release upon addition of a ligand (TPP) to a riboswitch RNA (from Bec et al. JACS 2013 **135** pp9743-52)



Example of a non-ergodic system

- ITC measurement of the heat release upon addition of a ligand (TPP) to a riboswitch RNA (from Bec et al. JACS 2013 **135** pp9743-52)

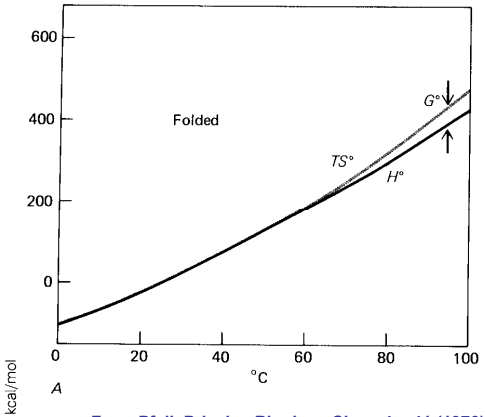


Modelling a complex polymer

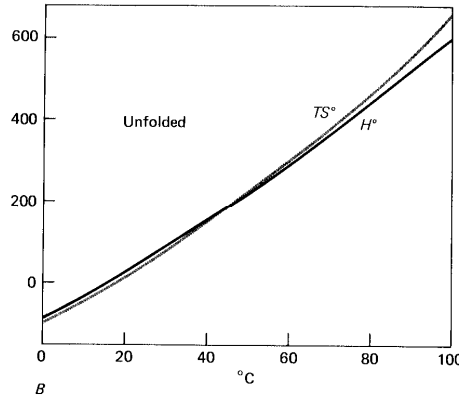
It is all about thermodynamics and kinetics

The internal energy of a protein defines its behaviour

- This energy results from a fine balance between two large reservoirs of energy: ΔH and –TΔS



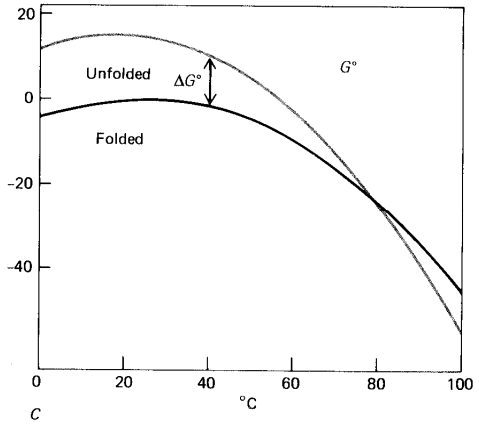
From Pfeil, Privalov Biophys. Chem 4, p41 (1976)



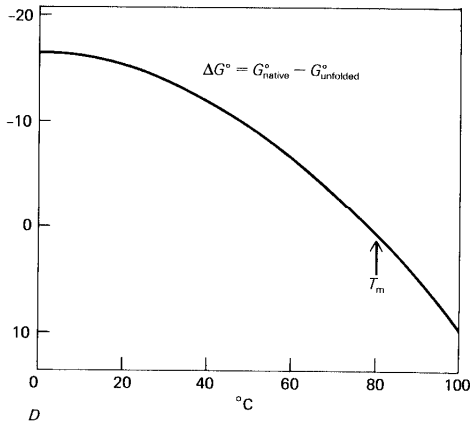
Lysozyme at pH 7

The internal energy of a protein defines its behaviour

- This energy results from a fine balance between two large reservoirs of energy: ΔH and –TΔS



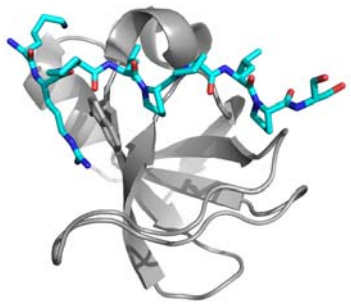
From Pfeil, Privalov Biophys. Chem 4, p41 (1976)



Lysozyme at pH 7

Single Structure representation emphasizes the enthalpic contributions

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



PDB 3I5R Batra-Safferling et al. 2010 JBC 391 p33

First NMR structure BUSI II Wühtrich et al. 1982

What is entropy ?

Entropy is a measure of the degrees of freedom for the system defined by the protein and the solvent

$T\Delta S_{rot,trans}$	*Rotational and translational degree of freedom*
$T\Delta S_{sol}$	*Organization of the water shell*
$T\Delta S_{conf}$	*Conformational entropy*

Disorder is conformational entropy

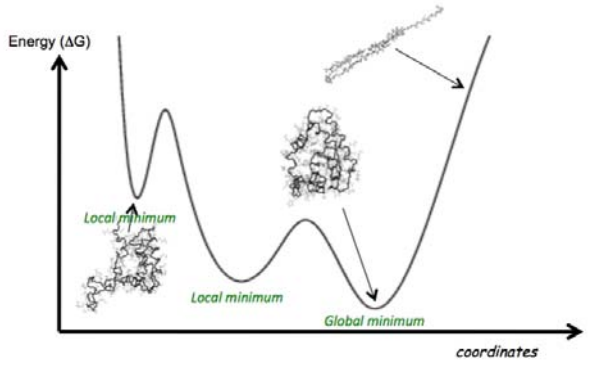
The conformational Entropy is defined by the number of states (conformations) accessible by the protein.

$$S = k_B \ln \Omega$$

Ω Number of microstates of the system

Disordered regions of proteins provides a fine reservoir of energy to fine-tune their function

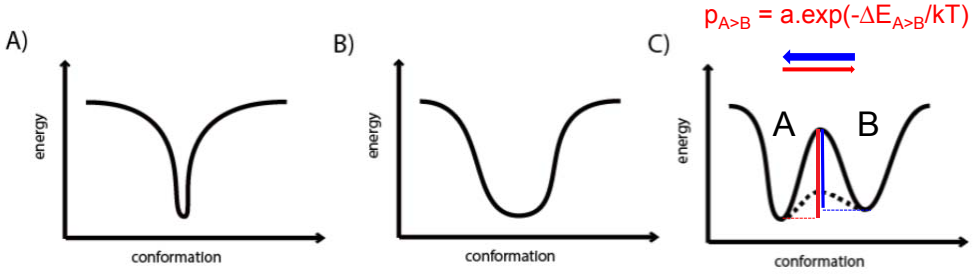
The energy landscape of a protein system



- ☞ The most stable (populated) conformation lies at the global minimum
- ☞ Local minima are “excited” states that are lowly populated
- ☞ The barriers between minima report on the kinetics of conformational transitions

The free energy landscape is a convenient way to represent structure **and** dynamics of proteins

Examples of free energy landscapes



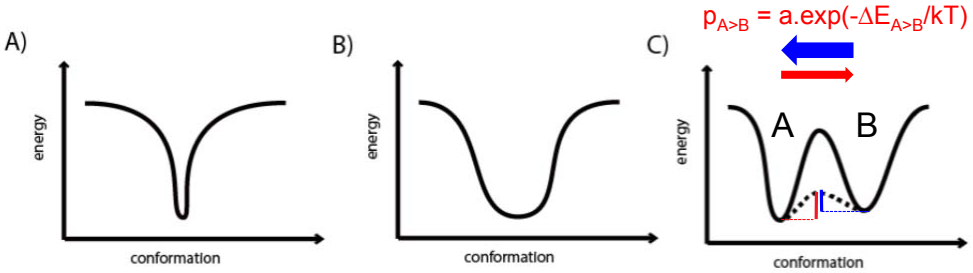
Unique rigid native conformation

Unique flexible native conformation

Two alternative flexible native conformations

Slow or rapid interconversion

Examples of free energy landscapes



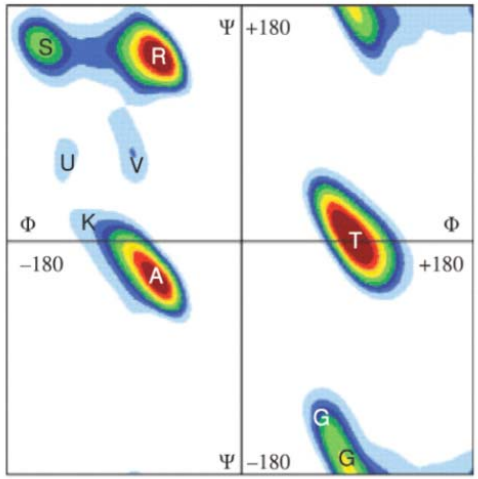
Unique rigid native conformation

Unique flexible native conformation

Two alternative flexible native conformations

Slow or rapid interconversion

Examples of free energy landscapes
The Ramachandran plot



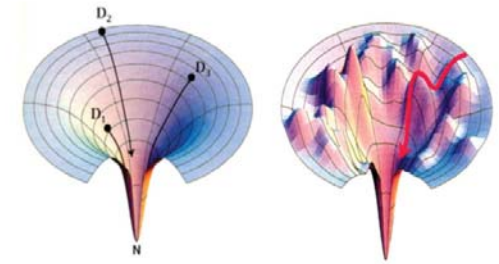
Ramachandran plot provides a 2D representation of the local energy of a dipeptide

A: α-helices
S: β-sheets
R: polyproline type II

K: ends of helices or in 3_{10} helices.
T: turn
G: glycine.

Various protein aspects and Free Energy landscape

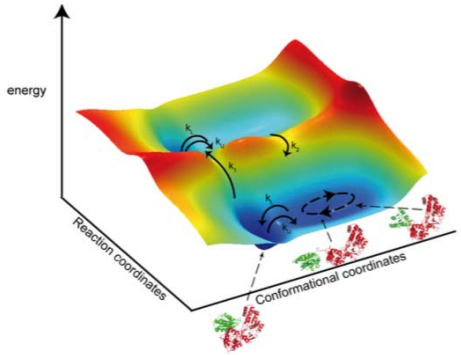
Protein folding



Modern view:
Protein folding results from the exploration of a funnel-type landscape

K.A.Dill & H.S.Chan (1997) Nature Struct.Biol. **4**, 10-19;

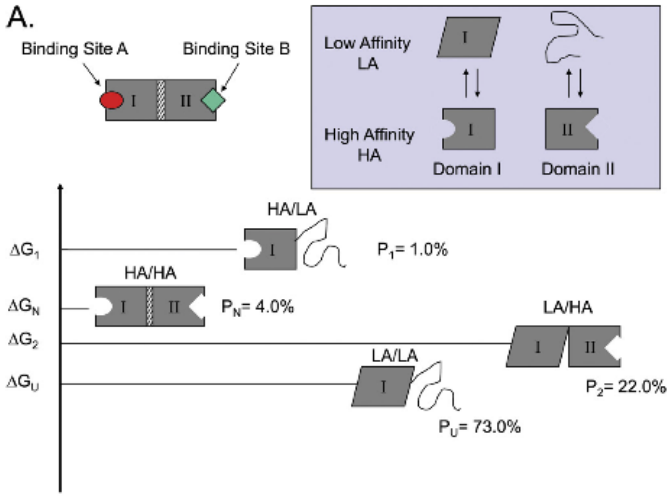
Enzymatic catalysis



Multidimensional energy landscape

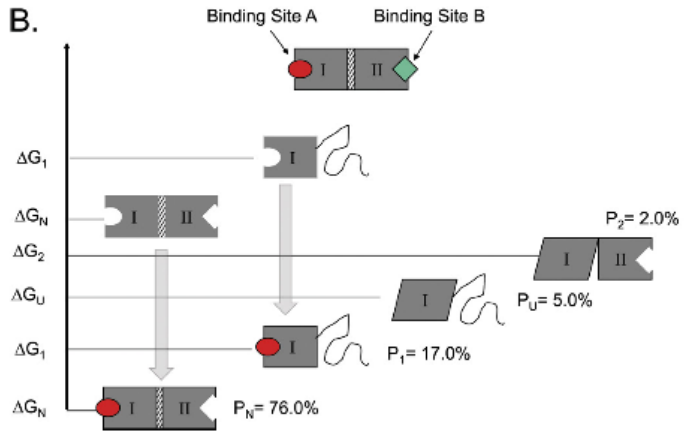
Frances O, Fatemi F, Pompon D, Guittet E, Sizun C, Pérez J, Lescop E, Truan G. Biophys J. 2015;108(6):1527-36

An thermodynamic (ensemble) view of allostery



From V.J. Hilser and E.B. Thompson JBC 286 p3967 (2011)

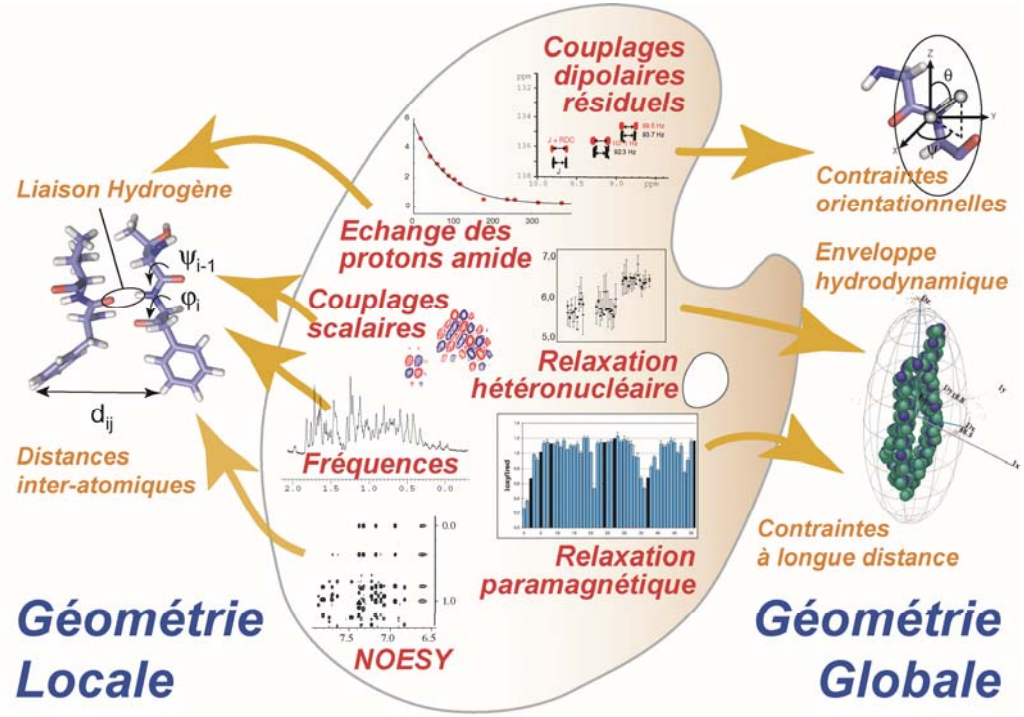
An thermodynamic (ensemble) view of allostery



From V.J. Hilser and E.B. Thompson JBC 286 p3967 (2011)

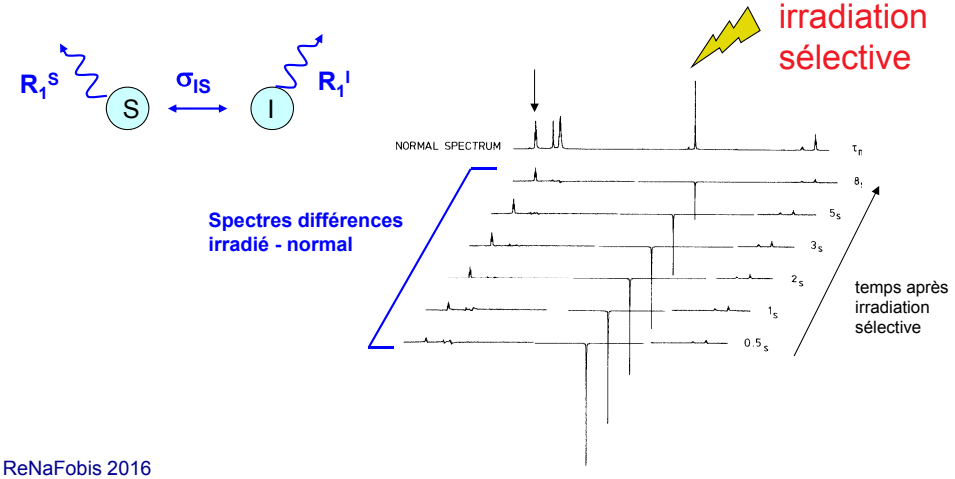
From NMR measurements to protein's models of structures

P. Signac port de La Rochelle 1921



NMR observables used for protein modelling

- Nuclear Overhauser Enhancement (NOE)



Quantification of NOEs

- The cross-relaxation rate allows interproton distances to be measured

$$\sigma_{IS} = A\left(\frac{6\tau_c}{1+4\omega^2\tau_c^2} - \tau_c\right)\frac{1}{r_{IS}^6}$$

Dynamics of the inter-nuclear vector

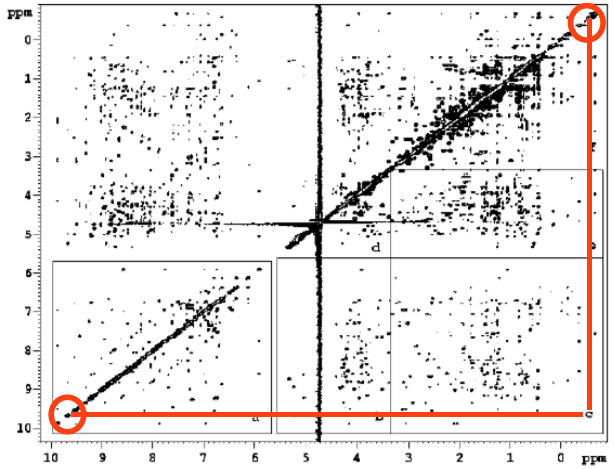
Distance measurement

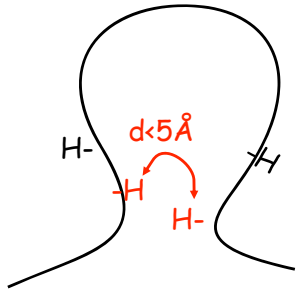
τ_c: *rotational correlation time*

ω : *spectrometer frequency*

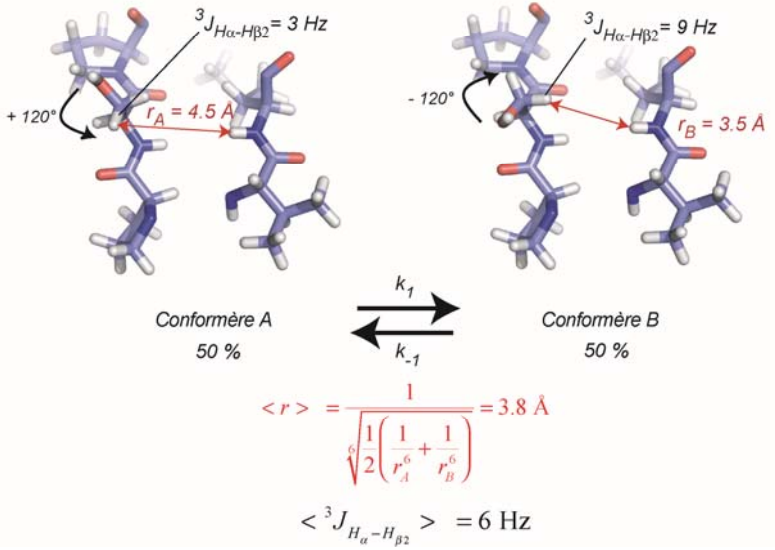
A : *dipolar constant*

The NOESY
A tool to measure NOE

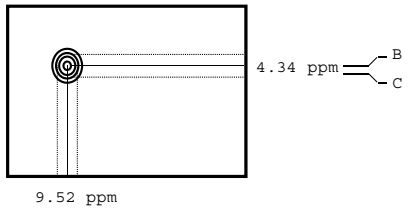


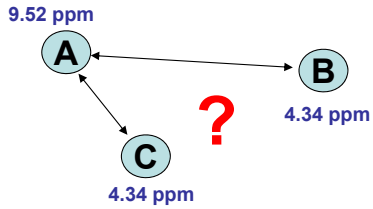


Effect of conformational averaging on NOE and coupling constants



Handling ambiguous NOE information





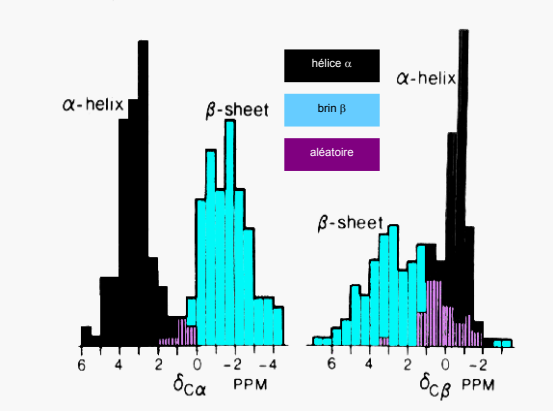
Ambiguous distances are grouped within d^{-6} averages

The redondancy of the data will enable a proper convergence and the distance assignment

$$\langle d \rangle = \left(\sum_{i=1}^{n}\frac{1}{d_i^6}\right)^{-1/6}$$

Integration of dihedral angle constraints in the model

The dependance of backbone nuclei from local structure is exploited to get ϕ and ψ dihedral angles using databases approaches



TALOS

Torsion Angle Likelihood Obtained from Shift and sequence similarity
F. Delaglio & A. Bax J.Biol.NMR 13 p289

Input: Chemical shifts ($^{13}C_\alpha$, $^{13}C_\beta$, $^{13}C'$, ^{15}N, $^{1}H_N$)

Recherche: Based on a similarity score for tripepeptides (frequencies and sequence)

Query ⇨ ..... L A T .....

..... I A T ..... ⇦ Sequence from the database
- Good quality 3D structres (2.2 Å)
- Reliable assignments

i-1 i i+1

Résultat: Return the best ϕ et ψ values for the central residue (i)

CS-ROSETTA

Consistent blind protein structure generation from NMR chemical shift data
Yang Shen, Oliver Lange, Frank Delaglio, et al.
Proc Natl Acad Sci USA, (2008) 105, 4685-4690

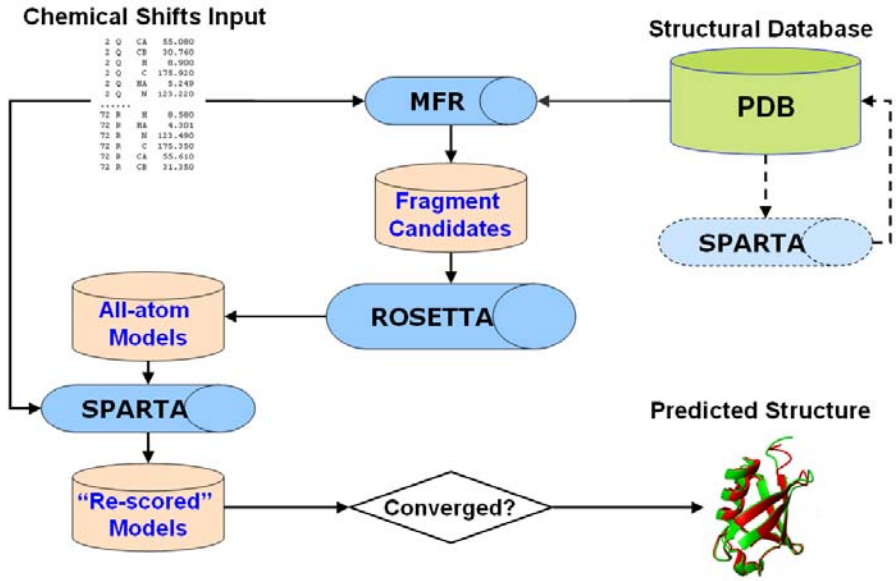
Sélection of rigid fragments based on chemical shifts calculations

Assembly of fragments into a full 3D model using ROSETTA (Monte-Carlo)

Addition of a additional term in the ROSETTA target function to take into account the chemical shift agreement between calculated and measured chemical shifts.

Based on chemical shift calculations from a 3D model using the SPARTA program

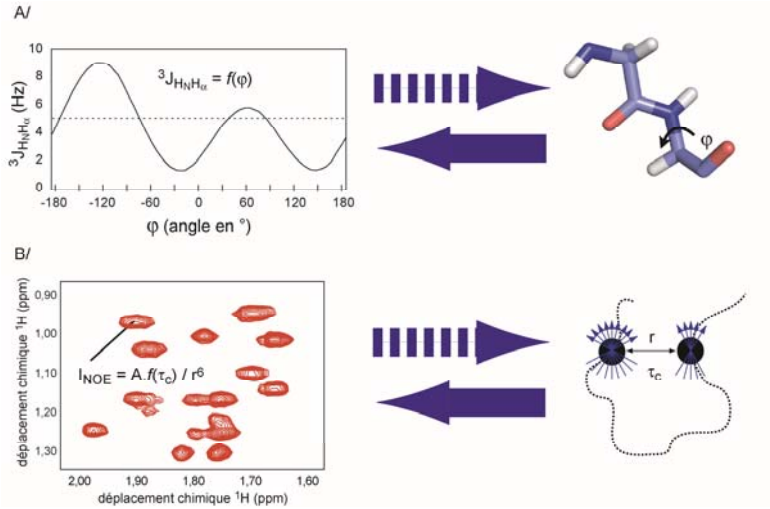
CS-ROSETTA



Structure determination from NMR data is a non-linear inverse problem:

Dihedral angles

NOEs



General procedure for structure determination

Model

NMR Observables

Y^i_{calc} *Back-calculated values*

Y^i_{meas} *Measured values*

Parameters of the model

$[x, y, z]_i$

$[\theta]_i$ *A priori information*

N (number of states)

τ_c (dynamics)

Distance measurement

$$D = \sum_i \left(Y^i_{meas} - Y^i_{calc}\right)^2$$

Global Optimisation Process

General procedure for structure determination

- Optimisation algorithm : Simulated annealing
  - Model: a set of atoms linked by interaction potentials
  - Atoms dynamics is simulated using Newton equations of motion
  - High temperature allows a random exploration of a large conformational space
  - Cooling down the system leads to a minimal value of the interaction energies

The target function: the potential energy of the molecular system

$$E_{hybrid} = E_{phys} + w_{data} E_{data}$$

Geometry

Real physical energy term

$\omega_{bond} E_{bond}$

$\omega_{angle} E_{angle}$

$\omega_{dihedr} E_{dihedr}$

$\omega_{vdw} E_{vdw}$

$\omega_{coulomb} E_{coulomb}$

Experiment

Adhoc energy term to take experimental into account

$$\sum_i \left(Y^i_{meas} - Y^i_{calc}\right)^2$$

Molecular dynamics simulations

From a starting conformation, we compute the following one (a dt time step after) by numerical integration of the equation of motion

$$m_i \frac{d^2 r_i}{dt^2} = \sum_n F_n = -\sum_n \frac{dE_n}{dr_i} = -\frac{dE_{Total}}{dr_i}$$

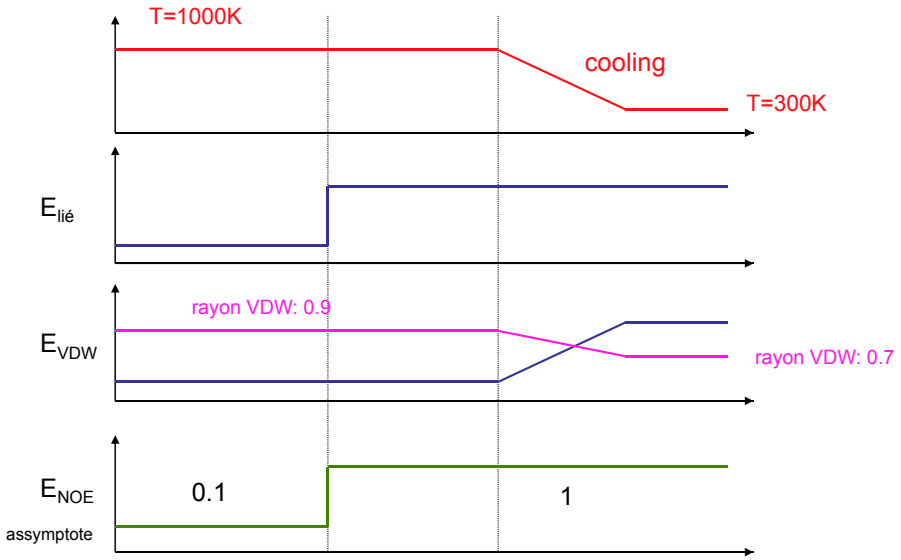
E_n : Potential energies of the system
r_i : Atom coordinates
m_i : Atom mass

The temperature of the system is defined from the atom velocities

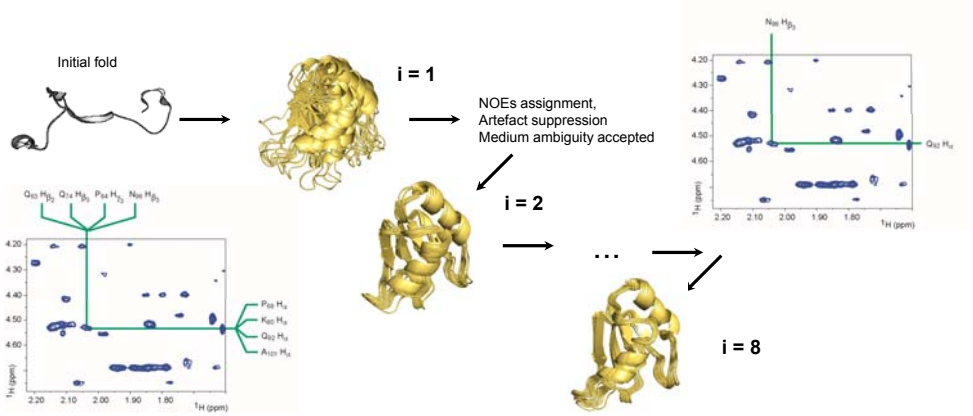
$$T = \frac{1}{(3N)} \sum_{i=1}^{N} \frac{|p_i|}{2m_i}$$

ReNaFobis 2016

A simulated annealing protocol



Iterative interpretation of ambiguous distances



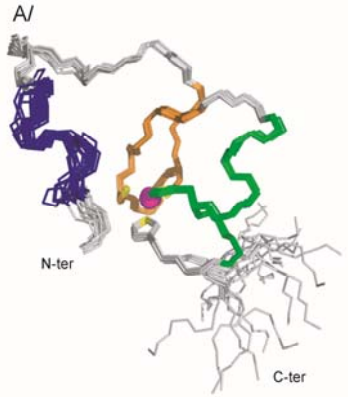
Logiciels ARIA, UNIO

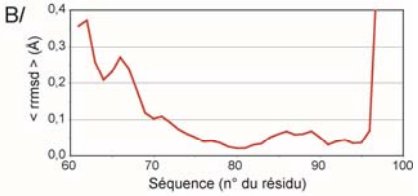
Rieping W., Habeck M., Bardiaux B., Bernard A., Malliavin T.E., Nilges M. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23:381-382.

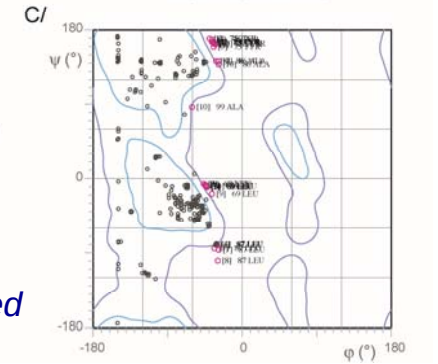
Volk, J.; Herrmann, T.; Wüthrich, K. J. Biomol.NMR. 2008, 41, 127-138..

Analysis of the resulting structures

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{n}}$$

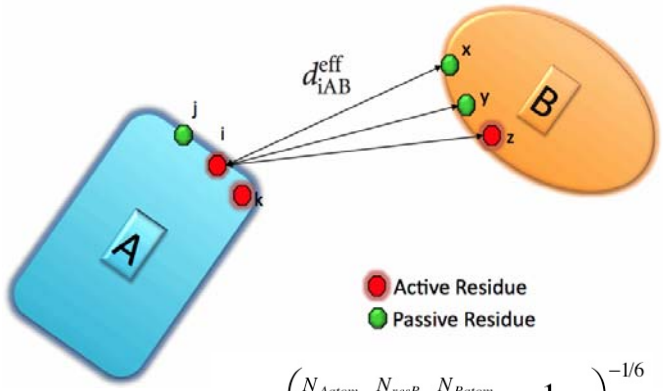






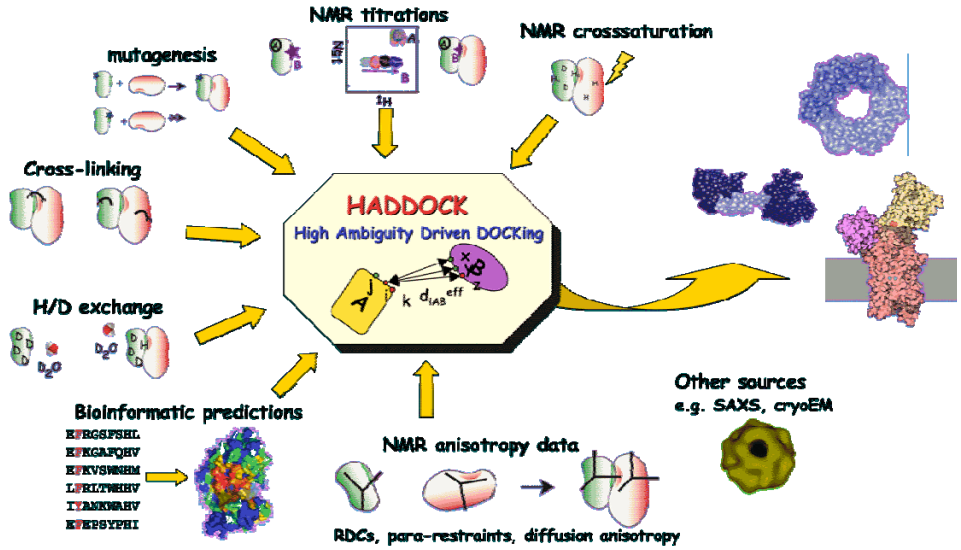
Correct geometry is assessed by Ramachandran plot

HADDOCK



Domingez C, Boelens R, Bonvin A, J. Am. Chem. Soc. 125, 1731-1737 (2003).

Use of Ambiguous Interaction Restraints for soft docking



Conclusions

- Integration of data requires a careful thought about the measurement and modelling processes
- Interesting and biologically relevant features may be hidden by the use of "over-simplified" models of proteins
- NMR provides a unique way to bridge thermodynamics and dynamics
- Complexes motions are captured using single molecule approaches