Part 8

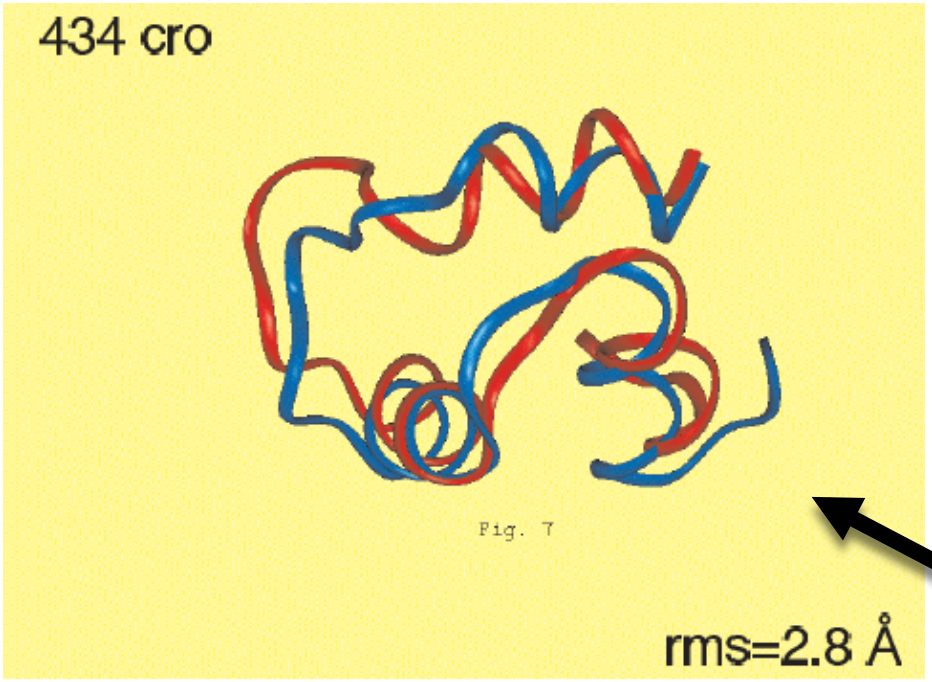
- *Ab initio* structure prediction approaches

- Thermodynamic stability prediction methods/ rational design of modified proteins

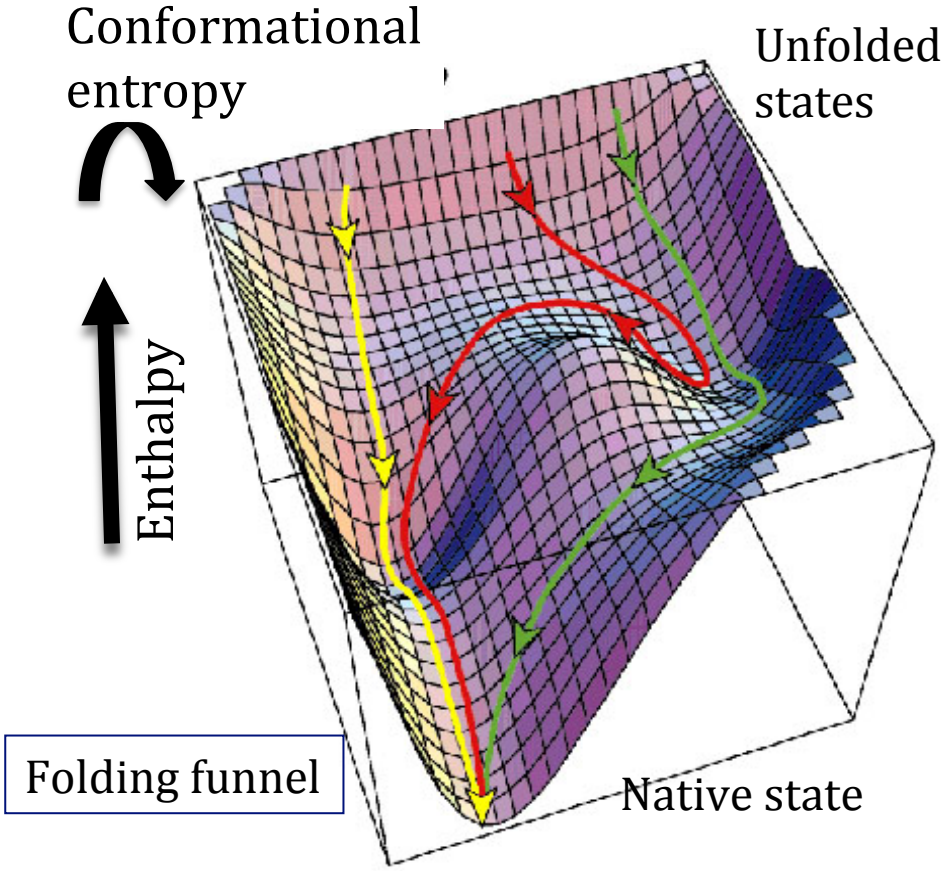
Ab initio structure prediction

Meaning: prediction from the sequence alone, without additional information
Not to be confused with *ab initio* in quantum chemistry, where it means resolution of the Schrodinger equation

Most difficult method, most costly in computer time: search for the native structure among all possible structures, or along folding pathways



Example of good results



Ab initio structure prediction

Steps:

- Choice of a (simplified) structural representation
- Choice of an algorithm for searching the conformational space
- Choice of an energy function for evaluating the sequence-structure compatibility that is adapted to the level of simplification

Concept of simplification:

It is assumed that, despite the simplification of the structure (and of the associated energy functions), it is possible to find a structure close to the native structure

--- Not necessarily as the solution of lowest free energy, but as a low free energy solution.

In a second step, simulations with a more detailed representation can be performed, starting from one (or more) solution(s) predicted in the first step – in order to refine the model - to get closer to the native structure

Ab initio structure prediction

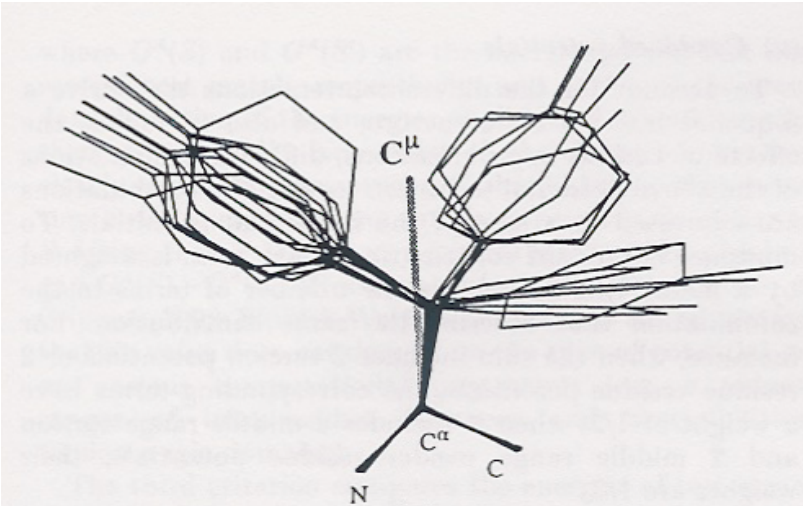
Choice of a (simplified) structural representation

- Most detailed model: all-atom model
  the most accurate, but the search in the conformational space is very costly in computer time (unless one remains in the vicinity of a structure, or unless one biases the search towards certain structures)

- Most simplified models : residue represented by a point, *e.g.* Cα, Cβ, average side chain centroid, …
  not very precise but faster

- Intermediate models: residue represented by the main chain and by an atom or pseudoatom representing the side chain, such as Cβ, average side chain centroid, …

Generally, the degrees freedom of the side chain are neglected



Ab initio structure prediction

Conformations are represented by:

- Cartesian coordinates of all atoms considered, or of the point(s) representing the residues.

-> Difficulty: ignores the polypeptide chain: if the position of an atom/residue is changed, the bond lengths/bond angles/torsion angles will in general no longer be correct

- Distances between atoms/residues

-> Difficulty: if the distances between N atoms/residues are (randomly) specified, they can in general not be represented in 3 dimensions – but they can always be embedded in (N-1) dimensions (if the triangular inequalities are satisfied) !!!

- The values of the main chain torsion angles (ϕ,ψ,ω) (optionally also of the side chains), considering the bond lengths and angles constant

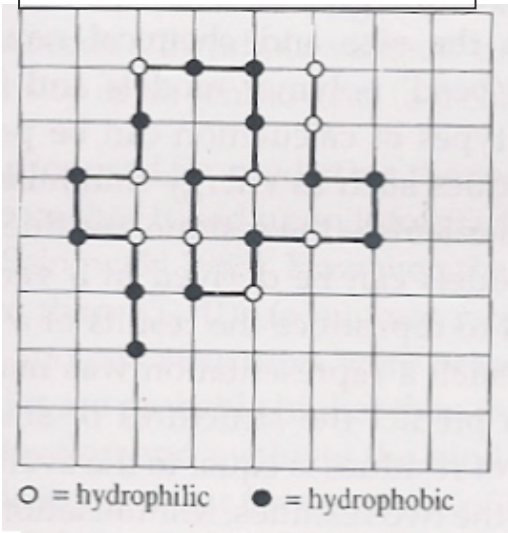
-> Difficulty: changing a torsion angle of one residue entails a movement of the entire part of the chain that follows the residue -> steric clashes

- Regular lattices: cubic, tetrahedral, etc..

-> Difficulty: impossible to represent a real protein structure on a regular lattice, unless the lattice spacing is reduced -> but then the advantage of the lattice is lost.

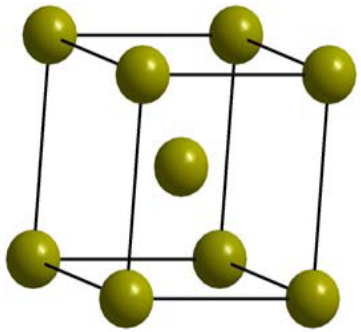
Ab initio structure prediction - lattices

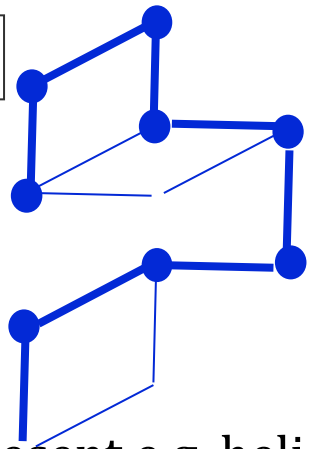
2D regular lattices

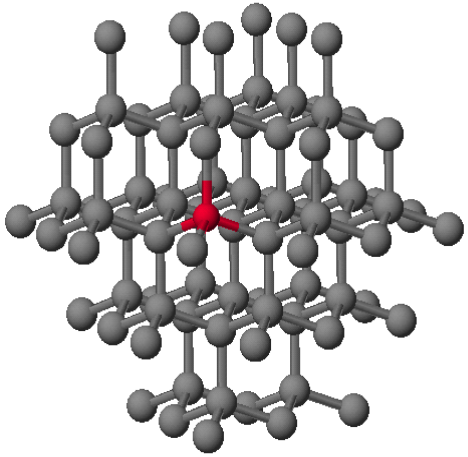


Unrealistic but allows for qualitative information

3D regular lattices



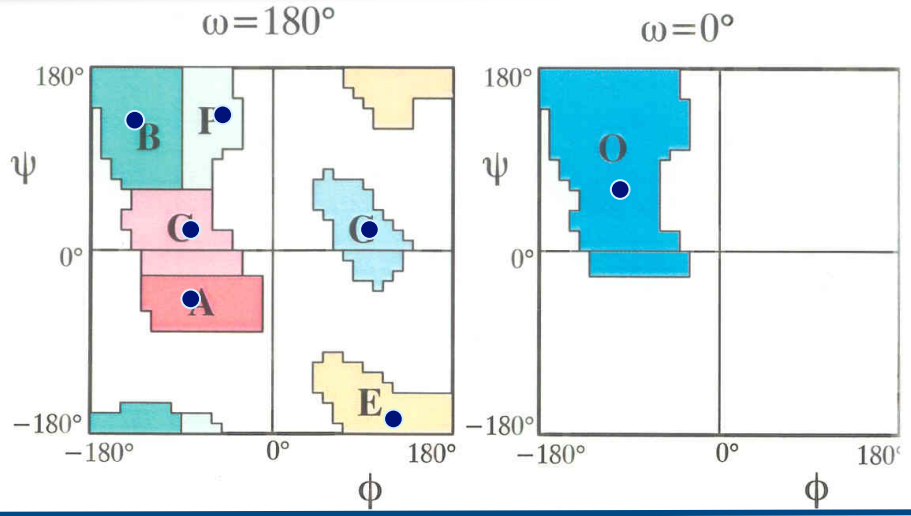




- Impossible to represent e.g. helices (at least in all directions)
- But allows compensated movements, which involve only a part of the chain.

3D irregular lattices

- Allows a perfect representation of α-helices and β-strands
- But does not allow compensated movements



***Ab initio* structure prediction**
Search in conformational space

Categories :
- Molecular dynamics
- Systematic search
- Model-building approaches
- Random sampling

Molecular dynamics

Principle:

Equations of motion (Newton)

$\mathbf{F} = m\,\mathbf{a} = m\,d^2\mathbf{r}/dt^2$ → numerical resolution

A $t = t_0$, initial conditions : position $\mathbf{r}(t_0)$ + velocity $\mathbf{v}(t_0)$ + force $\mathbf{F}(t_0)$

Calculation of the positions $\mathbf{r}(t_1)$, $\mathbf{r}(t_2)$ … - trajectory $\mathbf{r}(t)$

$\mathbf{F}$ → derives from the potential energy $\mathbf{F} = -\nabla E$

E estimated with semi-empirical energy functions

- Typically, all-atom representation + solvent molecules
- Very small movements at every step - in other words, very small time steps
  => Requires supercomputers and small proteins if you wish to fully explore the conformational space

Systematic search

Explores the conformational space by regular and predictable changes in conformation

For example:

Assume all bond lengths and bond angles constant, all residues in trans conformation, no degrees of freedom of the side chain.

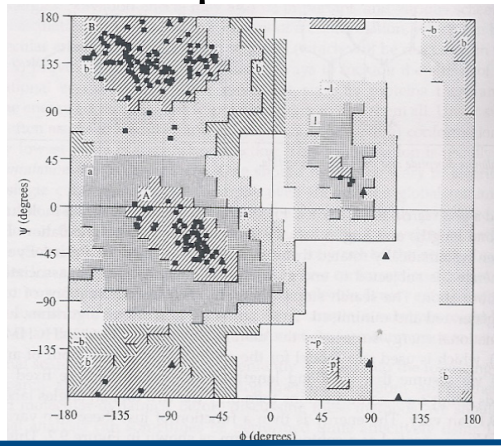
⇒ Only the torsion angles ϕ and ψ vary.

Search on a lattice : ϕ and ψ varies from 0 to 360 ° with a fixed increment θ.

⇒ combinatorial explosion: the number of conformations tested
= $\Pi_{i=1,\,2N-2}\ 360/\theta$, with N = number of residues, factor 2 because of the two angles ϕ and ψ and -2 because no ϕ for the first residue and not ψ for the last.

e.g. N = 2 and θ = 30 ° => 144 structures
N = 3 => ~ 21 000 structures;
N = 5 => ~ 430 million structures;

If 1 sec is needed to minimize 1structure:
N = 2 -> 2 min, N = 3 -> 6 hours, N = 5 -> 5000 days!

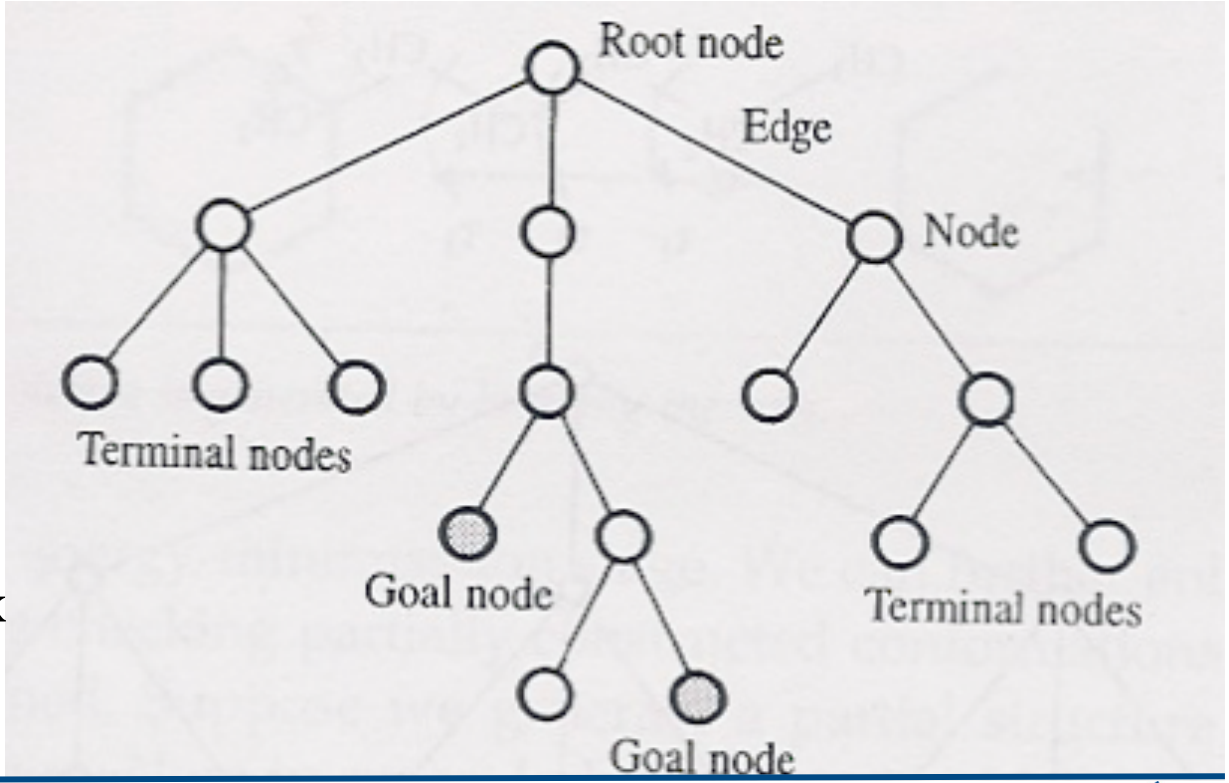


Improved systematic search

• Elimination of the energy minimization step for structures having a very high energy (e.g. with steric clashes)

• Checking partially built conformations -> all structures including substructures that are 'incorrect' or have too high energy values are eliminated (pruning the tree) - criteria!

**Any systematic search requires
a balance between
the resolution of the network
and the computation time
available.



Ab initio structure prediction - Search in conformational space

Model-building approaches

One way to reduce the combinatorial explosion of systematic research:

Using larger molecular fragments - already with a 3D structure – e.g. secondary structures -

Then assembling these fragments to obtain a 3D structure of the entire protein

Based on the following assumptions:

1) The conformation of each fragment is independent of those of other fragments -

This is only partially true! For example, tertiary interactions affect the secondary structure

2) To circumvent this hypothesis:

Keep several conformations for each fragment, and construct different global 3D structures from combining these fragments.

-> Conformations retained for each fragment must contain at least the structures observed in the native structures.

Ab initio structure prediction - Search in conformational space

Random sampling

Antithesis of systematic research:

systematic research explores the energy surface in a predictable way

<->

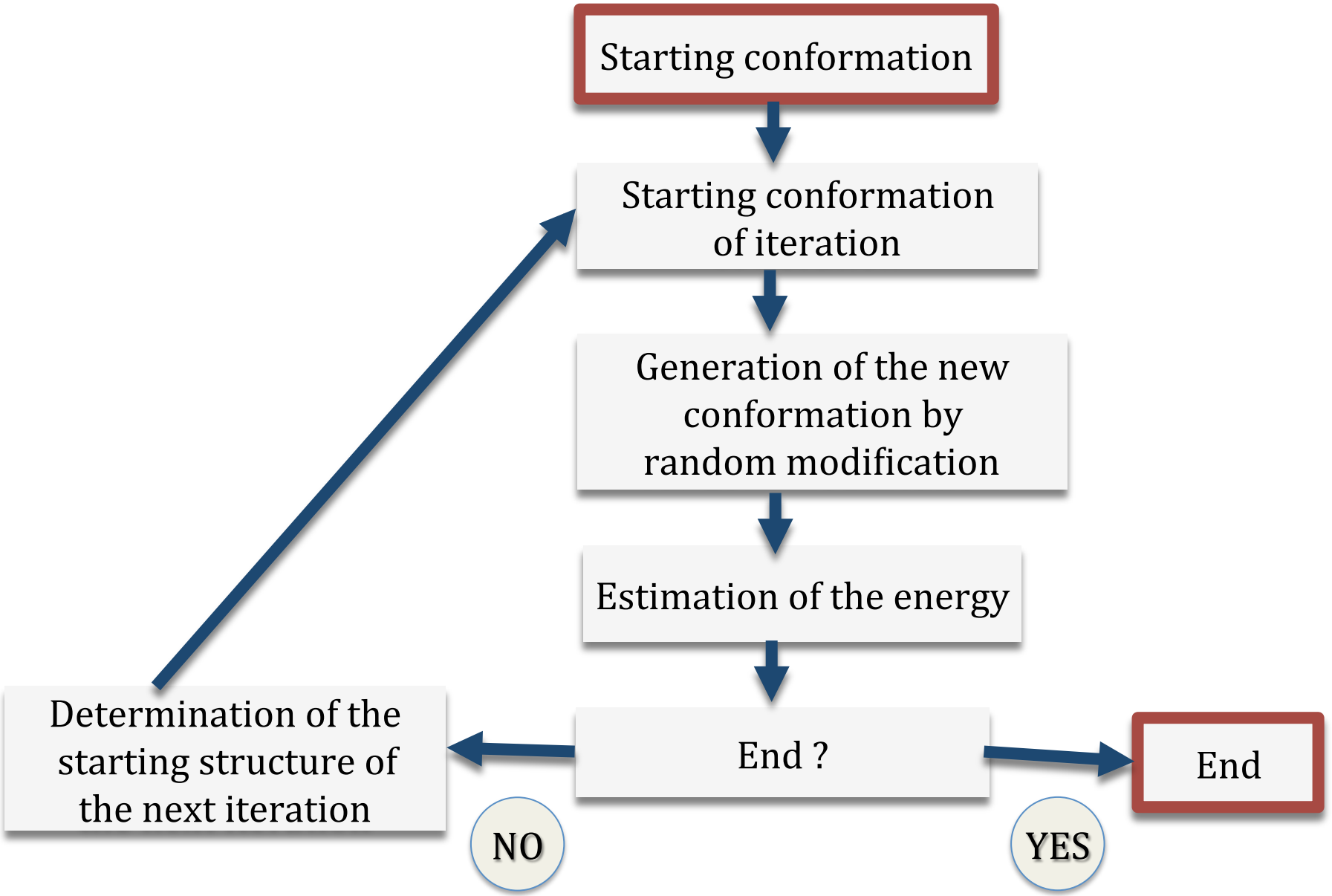
no order in random search

Random search can pass from a certain region in the conformational space to another region, completely disconnected, in a single step.

Movements applied during the search: for example

- In Cartesian coordinates: select randomly one (or more) atom(s)/residue(s), and add random quantities to its (their) (x, y, z) coordinates;
- In main chain torsion angle space: select randomly one (or more) torsion angles, and add random quantities to them or transfer them from one torsion angle domain to another, while keeping bond lengths and bond angles fixed => limits the number of degrees of freedom
- Same in distance space - except that the structure must be possible! Or embedded in 3D (see later)

Ab initio structure prediction - Random sampling of conformational space



Ab initio structure prediction - Random sampling of conformational space

Several ways to select a conformation for the next iteration:

- Choose the structure generated at the previous step
- Select randomly a structure from those previously generated by biasing the choice towards those that were selected less often (uniform usage protocol)
- Choose the lowest free energy structure generated so far, or bias towards low energy structures (Metropolis criterion , see next slides)

There is no fundamental reason to prefer one or the other of these procedures

** Some are more effective than others to explore the conformational space or to find the conformation corresponding to the global free energy minimum - depends on the problem

In systematic search procedures there is a natural end - when all conformations are tested.

In random search procedures, there is no natural end – impossible to be certain that the minimum free energy conformation is found.

Usual strategy: generating conformations until no new structures appear. This usually requires that each structure is generated several times => random methods do not explore the whole conformational space, but explore some areas several times.

Random sampling: Monte Carlo method

At each iteration, a new conformation is generated by performing a random modification, for example of the Cartesian coordinates or the main chain torsion angles of one or more residues, using a random number generator.

The free energy of the new conformation ΔW is calculated.

Metropolis criterion:

At each iteration i, the new conformation is kept as a starting point for the next iteration if:

- It has a lower free energy than its predecessor: $\Delta\Delta W \equiv \Delta W_i - \Delta W_{i-1} \leq 0$
- When it has a higher energy than its predecessor: $\Delta\Delta W > 0$:
  => We calculate the Boltzmann factor $B \equiv exp(-\Delta\Delta W / (RT))$
  => We choose a random number r between 0 and 1
  If $B \geq r$ => the structure is accepted

- If the free energy of the new state is close to that of the previous state, B is close to 1, and the new state is very likely to be accepted.
- If this energy is much larger, B is close to 0, and there is little chance that the new conformation is accepted.

Ab initio structure prediction - Random sampling of conformational space

BUT any conformation – even the worst - has a non-zero chance to be accepted!

=> Can climb over energy barriers to reach other minima.

$B \equiv \exp(- \Delta\Delta W / (RT))$: value of B depends on the value of T:

- For low T, most new conformations are rejected and the system remains close to a given free energy minimum
- For high T, many among the new conformations are accepted, high free energy barriers are crossed, and different regions of the conformational space are sampled.

⇒ Change the regime by varying T

⇒ This is called simulated annealing

Also applies to other techniques, such as molecular dynamics

Usual procedure:

- First high T to sample large portions of the conformational space, and cross energy barriers
- Then gradual decrease of T to further explore certain free energy minima

Typically, several successive phases of increase and decrease of T, or several independent simulations, to have a chance of finding the global minimum.

Ab initio structure prediction - Random sampling of conformational space

Where does the name "simulated annealing" come from? Annealing is the process by which the T of a molten material is slowly reduced until the material crystallizes to give a single large crystal, corresponding to the global minimum of the free energy. This process requires precise control of T near the transition.

Simulated annealing mimics this process to find optimal solutions to problems that have many suboptimal solutions.

The Metropolis criterion is obtained by imposing the microscopic reversibility condition :

At equilibrium, there are as many transitions in the two directions, from state m to state n, et from n to m => $R_{mn} = R_{nm}$

The transition rate R_{mn} is equal to the product of the population in state m (ρ_m) and the element π_{mn} of the transition matrix :

$$R_{mn} = \rho_m \pi_{mn} = \rho_n \pi_{nm} = R_{nm}$$

The population in a given state is given by its Boltzmann factor:

$$\pi_{mn} / \pi_{nm} = \rho_n / \rho_m = \exp(-\Delta W_n/RT) / \exp(-\Delta W_m/RT) = \exp(-\Delta\Delta W/RT)$$

The Monte Carlo algorithm generates a Markov chain of states. Indeed:

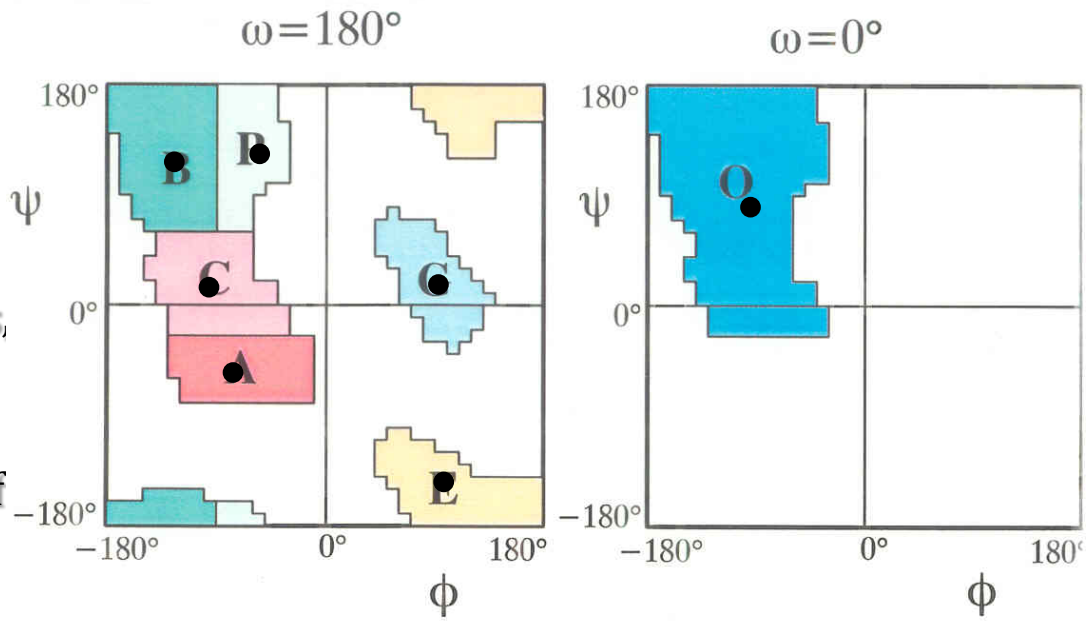
- The result of each iteration only depends on the preceding one (in contrast to molecular dynamics where all states are connected in time)
- Only a finite number of possible outcomes for each trial

Ab initio structure prediction - Random sampling of conformational space

The size of the moves at each iteration is controlled by the maximum displacement δr, which is an adjustable parameter whose value is chosen so that approximately 50% of all tested moves are accepted.
- If δr is too small, many moves are accepted but the successive states are very similar and the conformational space is explored very slowly.
- If δr is too large, many moves are rejected because they lead to steric clashes, which is very unfavorable.
* δr can be adjusted automatically while the program is running to get the right fraction of accepted moves

Moves in (ϕ−ψ−ω) space :
- Increment of e.g. 30°

or
- Choose among (e.g.) 7 (ϕ−ψ−ω) values => in this case, large moves, rapid exploration of conformational space - but many conformations rejected because of steric overlaps.



Evolutionary algorithms

Group of methods for sampling (conformational) space to find the optimal solution - based on ideas of biological evolution

Three classes:

- Genetic algorithm
- Evolutionary programming
- Evolutionary strategies

Ideas in common:

- Create a "population" of possible solutions;
- Members of the population are evaluated using an energy/cost/score function measuring their quality
- Population changes over time and (hopefully) evolves toward better solutions

Ab initio structure prediction - Random sampling of conformational space

Evolutionary algorithms: Genetic algorithm

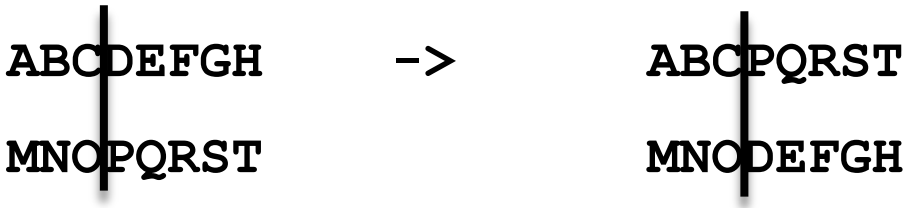
- Step 1: Create a (parent) population of m possible conformations (e.g. randomly generated) - each member of the population is encoded by a "chromosome", usually a linear chain (letters, or bits) representing e.g. main chain torsion angle domains .

  => Calculate the score/energy of each member – using energy function

- Step 2: Create a new population (children) - e.g.: m/2 members of the parent population are randomly selected (with a bias towards individuals with the highest energy). The new population is subjected to several "genetic" operators:

- mutation: at one or several randomly chosen positions, 1 letter/bit is modified

- recombination: consider two individuals, and randomly select a crossing position i. Two new individuals are created by swapping the string after position i. For example, i=3:



This is one cycle of the genetic algorithm. The new (children) population becomes the current (parent) population, ready to undergo a new cycle.

Evolutionary algorithms

Evolutionary programming

Similar to the genetic algorithm, but does not use the recombination operator - the new population is obtained only by mutation (cf Monte Carlo)

Evolutionary strategies

Different from evolutionary programming by:

- recombination operators are allowed
- no random choice of individuals maintained in the next generation. Rather, all individuals are classified according to their energy/score, and the best are selected.

This type of evolutionary algorithms are designed to global optimization - but they contain a significant element of randomness => so there is no guarantee that they produce the same solution each time (the solution corresponding to the absolute free energy minimum) - except for simple problems - but usually they give solutions very close to the global minimum in a reasonable time.

Ab initio structure prediction - Random sampling of conformational space

Using ideas from biology: <u>ant colonies</u>

Sampling the conformational space by mimicking ant colonies (Dorigo et al.)

◉ The algorithm is based on the observation that ants always find the shortest path between their nest and a food source.

◉ Principle: an ant deposits on its way indicators perceived by its peers: pheromones.

◉ When an ant has to choose a path, it chooses the one that has the highest pheromone concentration.

◉ Consequence: after going back and forth to various nest-food sources, the path that is the most dense in pheromones is the shortest.

Ab initio structure prediction
Random sampling of conformational space

Ant colony optimization

- pheromones brought by an ant

Ant colony optimization

Example: Using an algorithm of ant colony optimization for *ab initio* prediction of protein structures

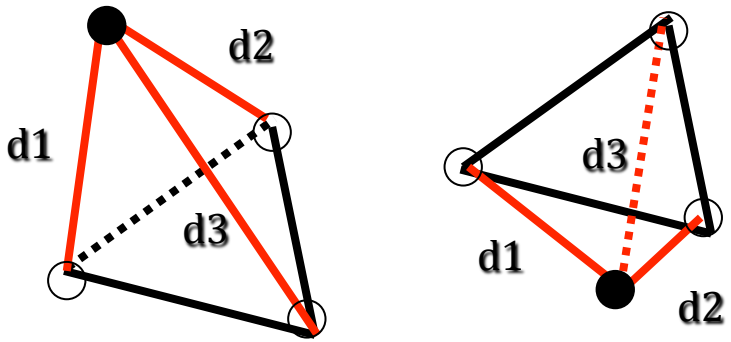
- Establish of a colony of artificial ants. Each ant builds a part of conformation of the protein of unknown structure.
- Deposit pheromones on the best solutions. Conformations are evaluated using an energy function.
- Create a new generation of artificial ants that build new structures taking into account the pheromones deposited.

Result:

- Turquoise: predicted structure
- Green: experimental structure

Ab initio structure prediction - Distance geometry

- Another way to describe the conformations of a molecule: in terms of distances between all pairs of atoms/residues.
- For N atoms/residues: N (N-1) / 2 distances => symmetric NxN matrix with 0 on the diagonal.
- Distance geometry explores the conformational space distances by generating a large number of distance matrices, which are then converted into conformations represented by Cartesian coordinates.
- But: inter-residue/atomic distances are correlated - many combination of distances are geometrically impossible.
  For example: for a triangle ABC, | AB | + | BC | ≥ | AC |



The position of each atom is determined by its distance to three other (non collinear) atoms – up to a reflection indeterminacy

Ab initio structure prediction - Distance geometry

Principle

Step 1: a matrix of lower and upper limits of each distance between atoms/ residues is calculated - based on simple chemical principles

Step 2: values are randomly assigned to each distance between atoms/residues between the upper and lower limits.

Step 3: the distance matrix NxN is converted into a test set of Cartesian coordinates for N atoms in a process called "embedding"
=> search for 3D coordinates that are the most compatible with the distances
=> embedding of structures from N-1 dimensions into 3 dimensions.

Step 4: Refinement of the coordinates

This technique is used in NMR, which provides distance constraints between atoms, with a certain experimental error margin => search for 3D structures that are compatible with the experimental distance constraints, taking the errors margins into account.

Embedding

From the NxN distance matrix, composed of all distances between residues/atoms i and j (d_{ij}), compute the NxN real symmetric metric matrix :

$$G_{ij} = = \frac{1}{2} (d_{io}^2 + d_{jo}^2 - d_{ij}^2)$$

The origin o is generally taken as one of the residues/atoms near the center of the molecule.

Diagonalize **G** => $\mathbf{G = V \Lambda V^{-1}}$; **V** et $\mathbf{\Lambda}$ are NxN matrices; $\mathbf{\Lambda}$ is diagonal.

G is square symmetric => $\mathbf{G = G^T}$ => $\mathbf{V^{-1} = V^T}$ => $\mathbf{G = V \Lambda V^T}$ (All real symmetric matrices are diagonalizable)

Express $\mathbf{\Lambda = L^2}$ => $\mathbf{G = (V L)\ (V L)^T = X X^T}$ where $\mathbf{X = V L}$ is an NxN matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_N \end{pmatrix}$$

⇒ **X** "contains" the N atomic coordinates.

What does that mean?

In general, the conformation cannot be exactly embedded in 3D. The « best » 3D structure compatible with the distances d_{ij} is obtained from the eigenvectors that correspond to the three largest eigenvalues Λ, denoted λ_1, λ_2, λ_3

Embedding

$$X = VL = \begin{pmatrix} V_{11}\lambda_1^{1/2} & V_{12}\lambda_2^{1/2} & V_{13}\lambda_3^{1/2} & V_{14}\lambda_4^{1/2} & \dots & V_{1N}\lambda_N^{1/2} \\ V_{21}\lambda_1^{1/2} & V_{22}\lambda_2^{1/2} & V_{23}\lambda_3^{1/2} & V_{23}\lambda_3^{1/2} & \dots & V_{2N}\lambda_N^{1/2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ V_{N1}\lambda_1^{1/2} & V_{N2}\lambda_2^{1/2} & V_{N3}\lambda_3^{1/2} & V_{N3}\lambda_3^{1/2} & \dots & V_{NN}\lambda_N^{1/2} \end{pmatrix}$$

⇒ Coordinates of each residue/atom i are: $(\lambda_1^{1/2} V_{i1}, \lambda_2^{1/2} V_{i2}, \lambda_3^{1/2} V_{i3})$

⇒ Best way of embedding an object living in a space in N-1 dimensions to a space in 3 dimensions.

cf Principal component analysis –

Embedding

Check

If the object/conformation is already compatible with 3 dimensions:

$G_{ij} = \frac{1}{2} (d_{io}^2 + d_{jo}^2 - d_{ij}^2) = (x_i - x_0)(x_j - x_0) + (y_i - y_0)(y_j - y_0) + (z_i - z_0)(z_j - z_0) = \mathbf{i}.\mathbf{j}$

where **i** and **j** are vectors linking the origin o to the residues/atoms i et j:

$\mathbf{i} = (x_i - x_0, y_i - y_0, z_i - z_0)$, $\mathbf{j} = (x_j - x_0, y_j - y_0, z_j - z_0)$

Diagonalize: $\mathbf{G = V \Lambda V^{-1} = V \Lambda V^T}$

Here $\mathbf{\Lambda}$ is the identity matrix and $\mathbf{X = V}$

$$\mathbf{X} = \begin{pmatrix} x_1 - x_0 & y_1 - y_0 & z_1 - z_0 & 0 & \dots & 0 & 0 \\ x_2 - x_0 & y_2 - y_0 & z_2 - z_0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & 0 & 0 \\ x_N - x_0 & y_N - y_0 & z_N - z_0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

Evaluation of the structure prediction methods

- Very important point!
By predicting the structure of proteins of known structures, by performing cross-validation

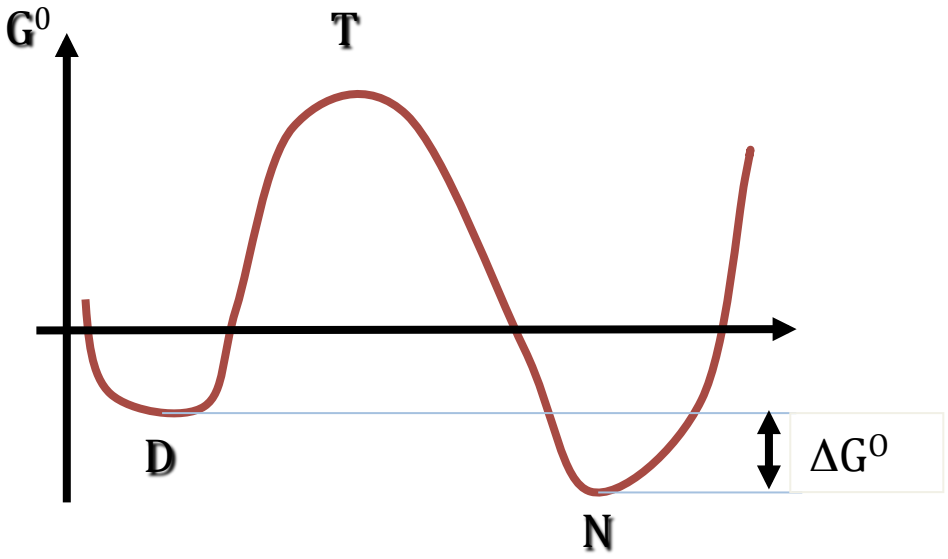
Good way to test the prediction methods:
Biannual meetings "Critical Assessment of Techniques for Protein Structure Prediction" (CASP):
Sessions:
- Secondary structure prediction
- Comparative modeling
- Fold recognition
- *Ab initio* prediction

Predictors/modelers are challenged to predict/model the structures of target proteins of unknown 3D structure and submit their predictions/models to the organizers before the meeting.

Simultaneously, the 3D structures of the targets are determined by X-ray crystallography or NMR. They only become accessible after the predictions/models are submitted.

=> Blind tests

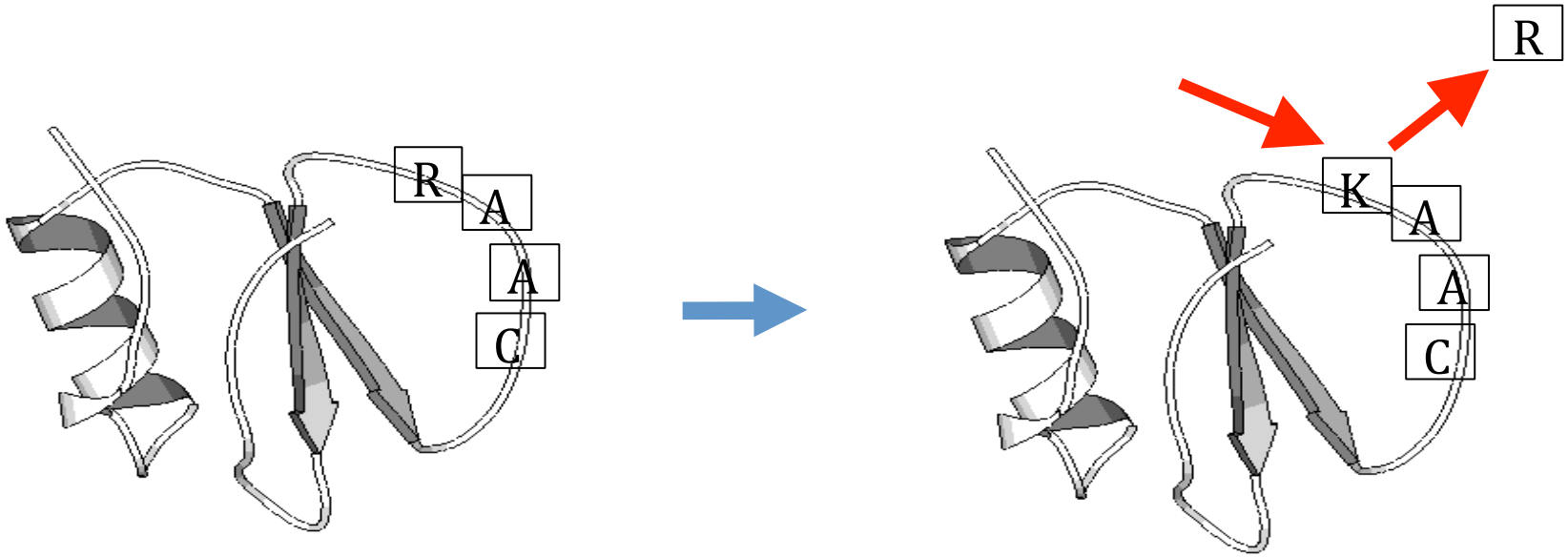
Prediction of the thermodynamic stability of proteins (ΔG^0)



Easier: Prediction of protein stability changes upon point mutations

Objective: Rational design of modified proteins

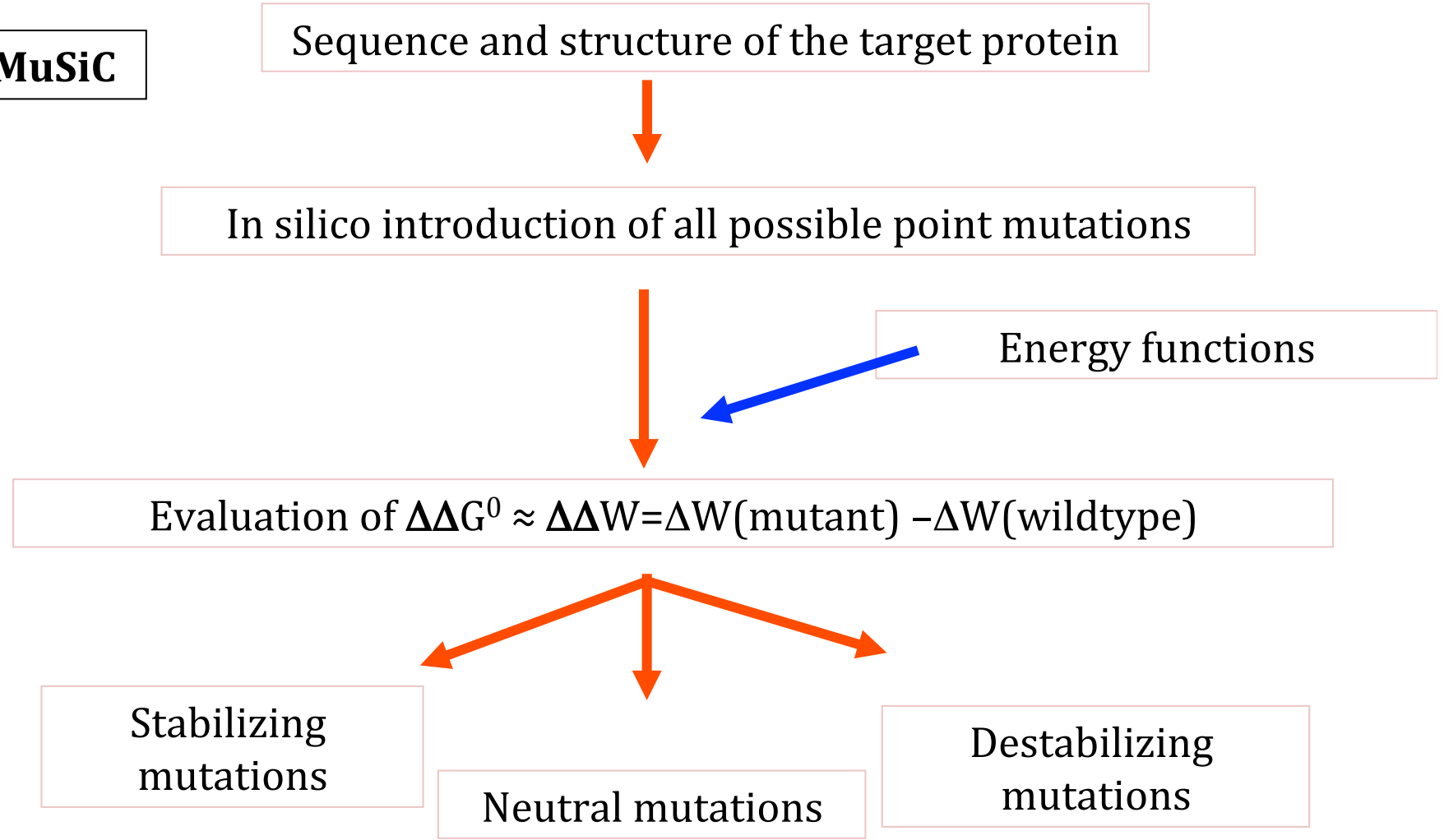
** Thermal stability, thermodynamic stability, solubility, function, structure, ...



Prediction of the thermodynamic stability changes upon ($\Delta\Delta G^0$)

Example of program that predicts changes in thermodynamic stability upon point mutations

PoPMuSiC



Prediction of the thermodynamic stability changes upon ($\Delta\Delta G^0$)

Energy functions

A: solvent accessibility of residue
V: volume of residue

$$\Delta\Delta W = \sum_{i=1} \alpha_i(A)\Delta\Delta W_i + \beta(A, sign(\Delta V))\Delta V + \gamma(A)$$

Linear combination of several potentials, e.g. torsion, distance, accessibility potentials, … + terms related to changes in volume of the wildtype/mutant residue

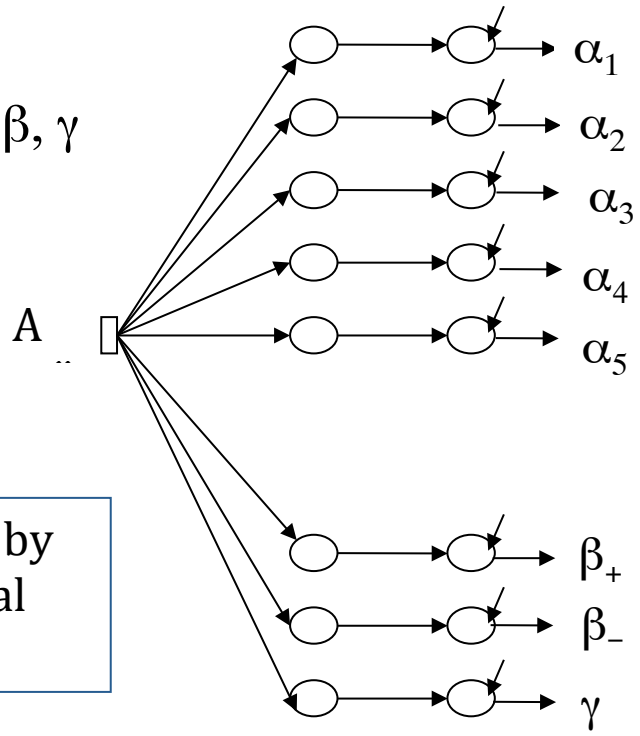
-> Determine the values of the parameters α, β, γ

- Based on all the mutant proteins whose $\Delta\Delta G^0$ have been experimentally measured
- Try to minimize the difference between $\Delta\Delta W$ et $\Delta\Delta G^0$

Cost function :

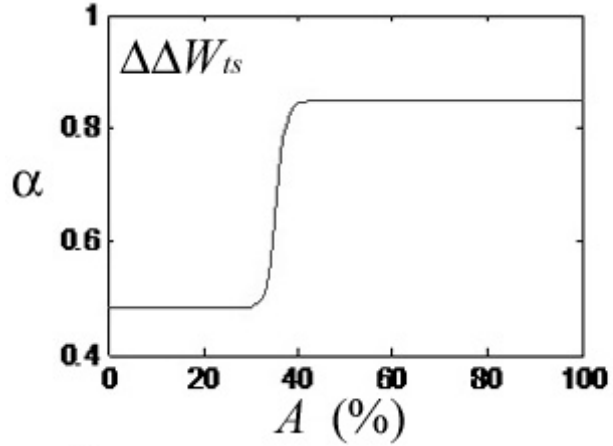
$J = \Sigma_{mutants} (\Delta\Delta G_m - \Delta\Delta W_m)^2$

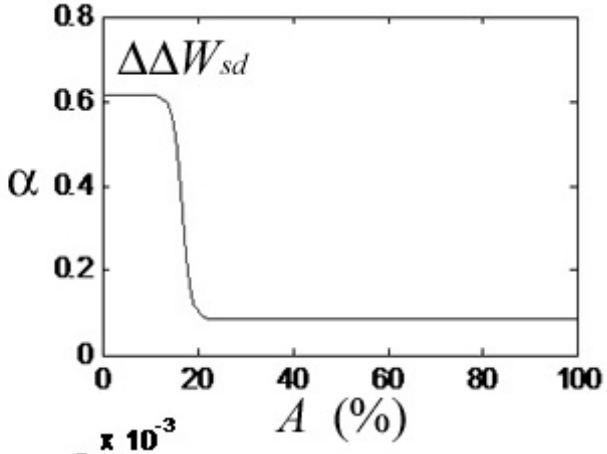
For example by using a neural network



Prediction of the thermodynamic stability changes upon ($\Delta\Delta G^0$)

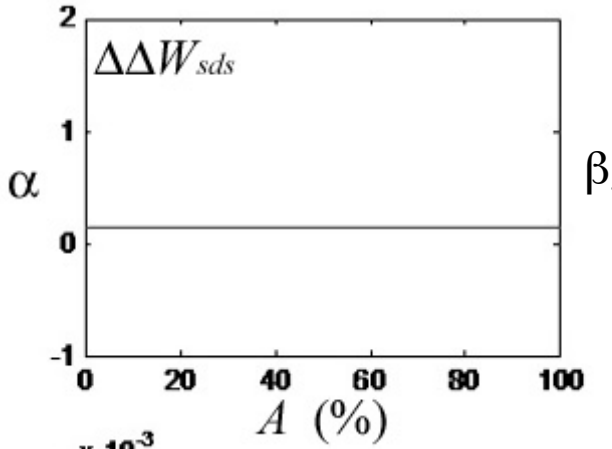
Torsion potential : more important at the surface

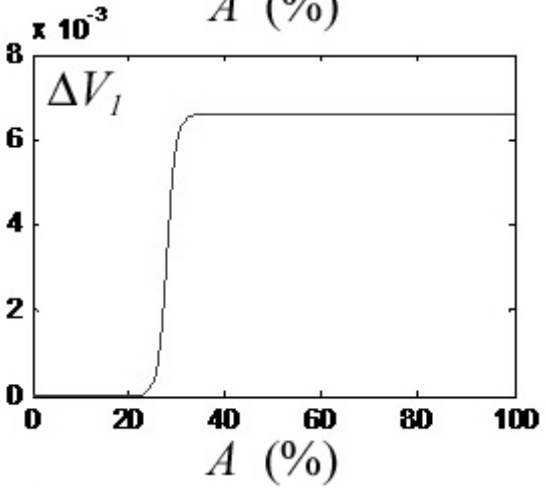




Distance potential ~ $\Delta W(d,s)$: more important in the core

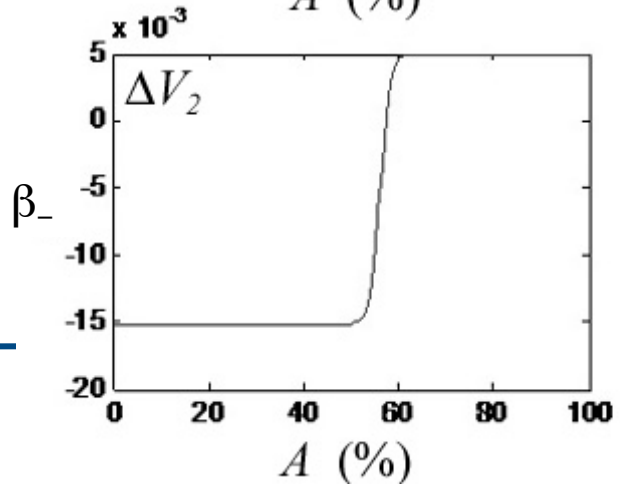
Other distance potential ~ $\Delta W(d,s_1,s_2)$ - $\Delta W(d,s_1)$: constant

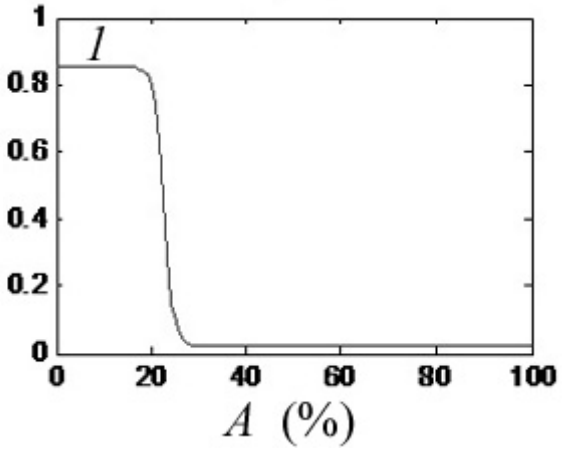




Volume term $V_{mut} > V_{sauv}$ or $\Delta V > 0$: more unfavorable at the surface !!!???

Volume term $V_{mut} < V_{sauv}$ or $\Delta V < 0$: more unfavorable in the core





Additive term: more destabilizing in the core

Prediction of the thermodynamic stability of proteins (ΔG^o)

Conclusion:

• Importance of tertiary interactions (distance potential) decreases from core to surface - in particular the potentials of the type

$$\Delta W(d, s) = -RT \ln [P(d, s) / (P(s) P(s))]$$

which are dominated by the hydrophobic effect

• Potentials of the type

$$\Delta W(d, s, s) = -RT \ln [P(s, s, s) / (P(s, s) P(s))] - \Delta W(d, s)$$

describe specific interactions - all are important

• Importance of local interactions along the chain (torsion potential) increases from the core to the surface – but they also play a role in the core

• Mutation of a residue into a smaller one -> cavity formation -> generally unfavorable in the core

• Mutation of a residue in a larger one -> stress -> generally unfavorable in the core. We see the opposite in the corresponding potential!
May be a compensation between this term and the constant term - the terms are not independent

• Constant term: a mutation is on average always worse in the core than on the surface.

Rational design of modified proteins

Program can be used to predict (de)stabilizing mutations -> rational design of proteins of modified stability.

Can also be used to identify regions where:

- Most mutations are destabilizing => regions for which the sequence is relatively optimal for the stability of the native structure

- Many mutations are stabilizing => regions for which the sequence is not optimal for the stability of the native structure

=> structural weaknesses

can be:

- a functional region (active site, ligand binding site, ...)
- a region that undergoes conformational changes (e.g. when binding to a ligand)
- region that initiates a local unfolding -> conformational change -> conformational diseases
- .....

Rational design of modified proteins

To estimate the optimality of a sequence with respect to the stability of its native structure :

At each position i: mutate the native residues into the 19 others, noted r, and compute the $\Delta\Delta W_{ir}$'s

O_i= sum of $\Delta\Delta W_{ir}$ over all r for which $\Delta\Delta W_{ir} > 0$

N_i= sum of $\Delta\Delta W_{ir}$ over all r for which $\Delta\Delta W_{ir} < 0$

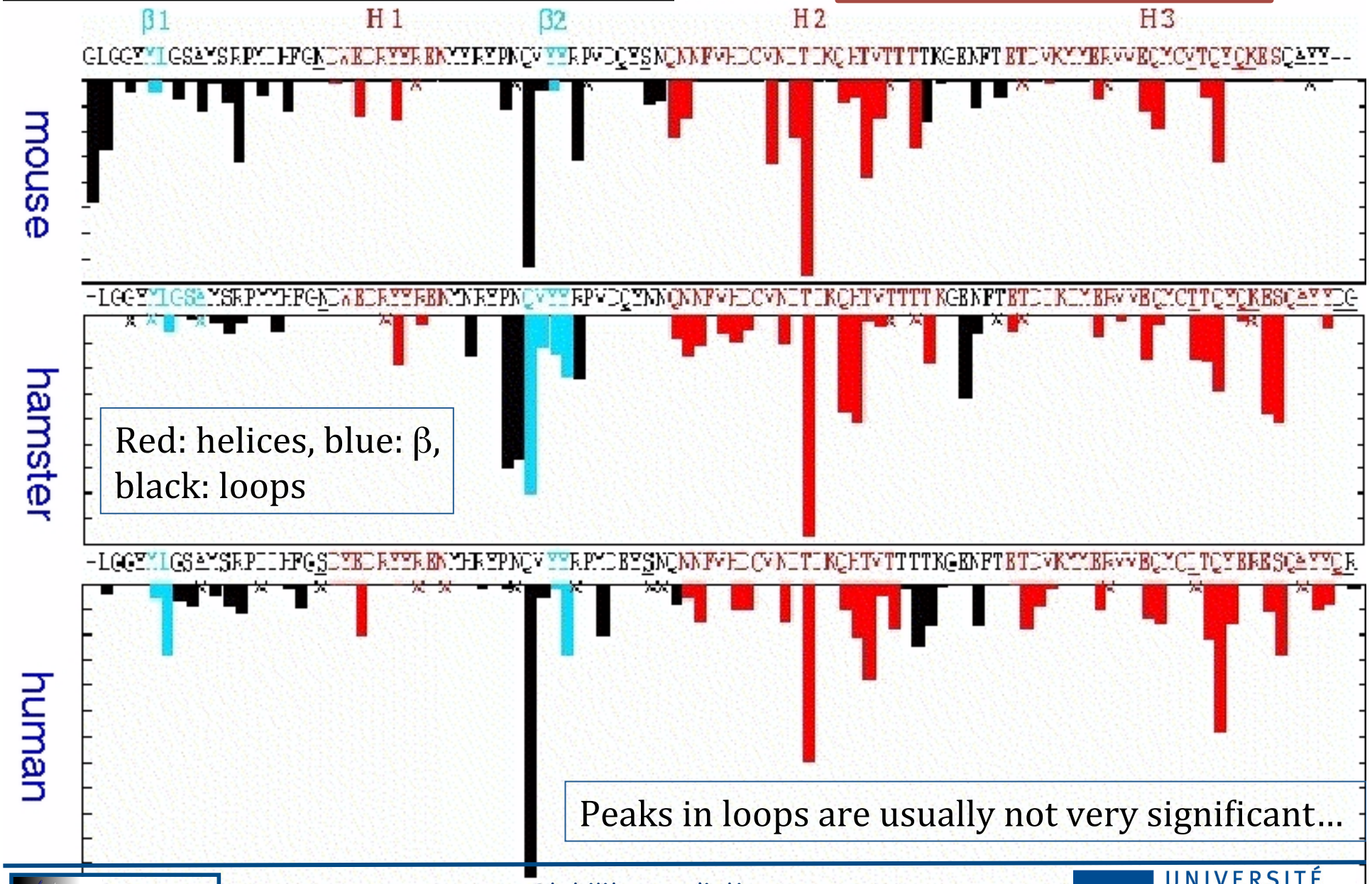
Positive peaks of O_i : Positions for which the sequence is relatively optimal

Negative peaks of N_i : Positions for which the sequence is relatively nonoptimal

=> good sites for introducing stabilizing mutations

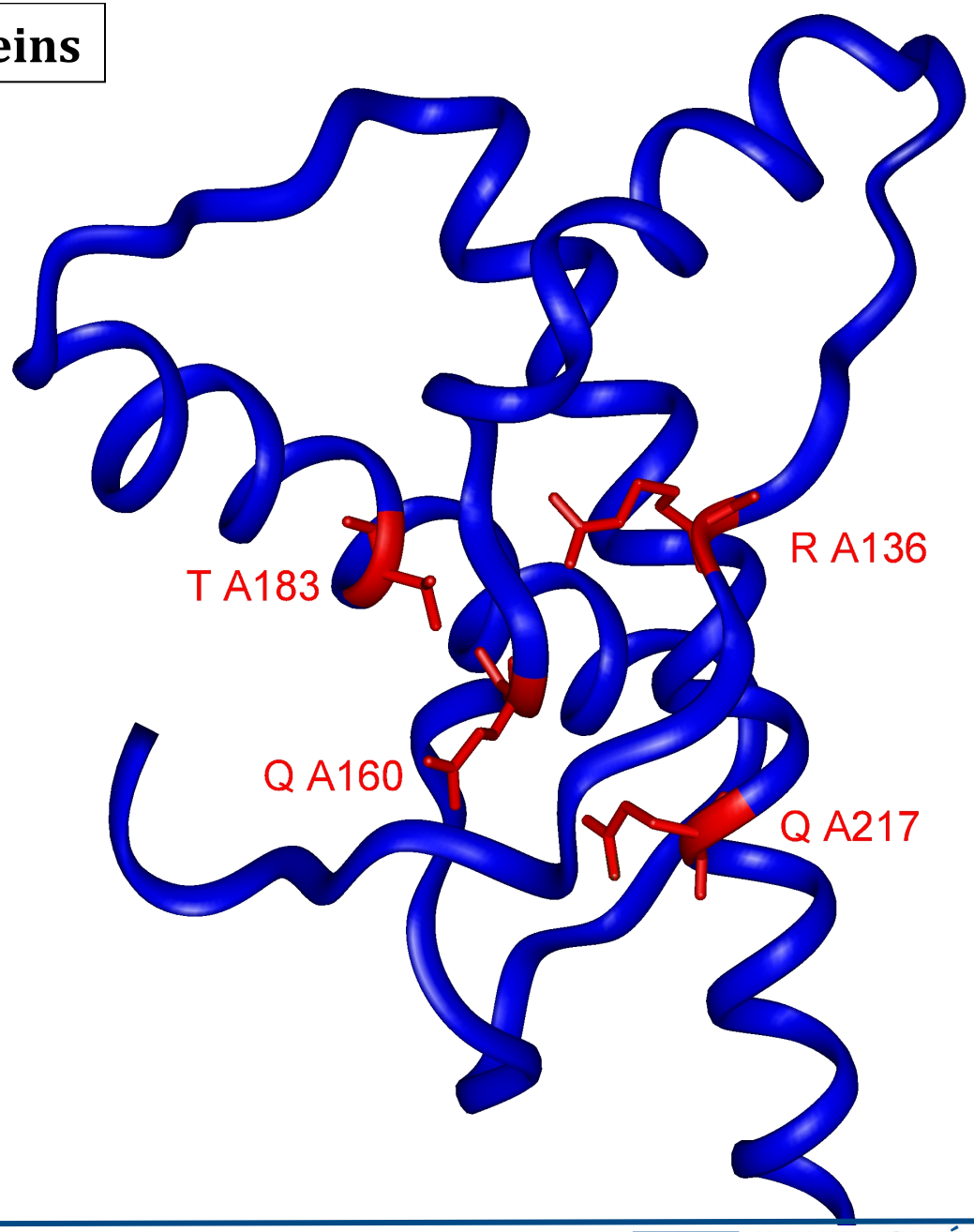
Rational design of modified proteins

Example, prion protein



Rational design of modified proteins

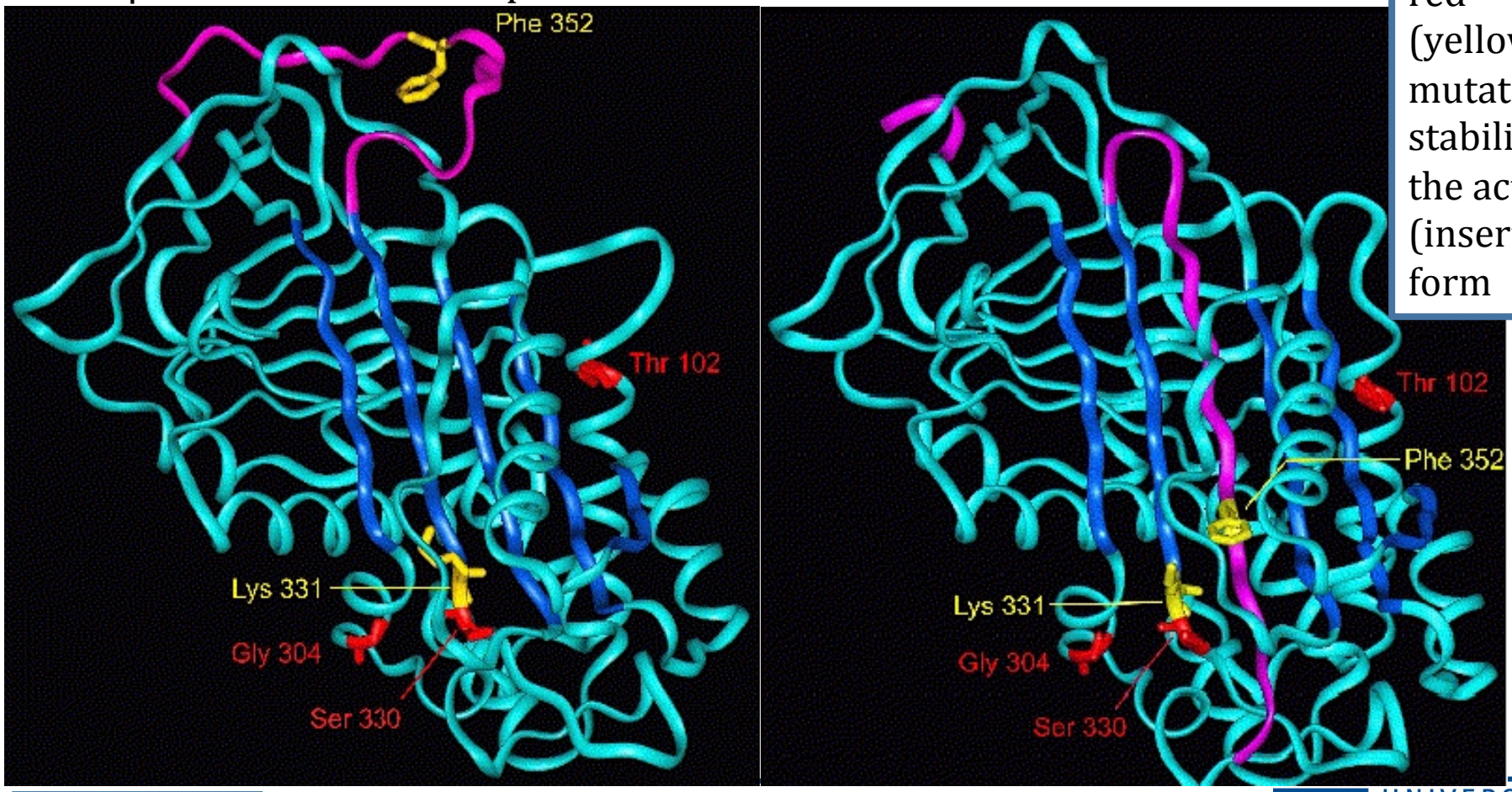
Example, prion protein



Rational design of modified proteins

Example, α1 antitrypsin

= Serine protease inhibitor = serpin. Loop (left, purple) recognized and cleaved by serine proteases => conformational change of the inhibitor => loop inserts into the β-sheet (right) => irreversible trapping of serine protease at the "bottom right" of the inhibitor.
**Potential problem: in the absence of serine protease, polymerization by insertion of loop in the β-sheet of another serpin => conformational diseases



red (yellow): mutations stabilizing the active (inserted) form

Rational design of modified proteins

Other examples:

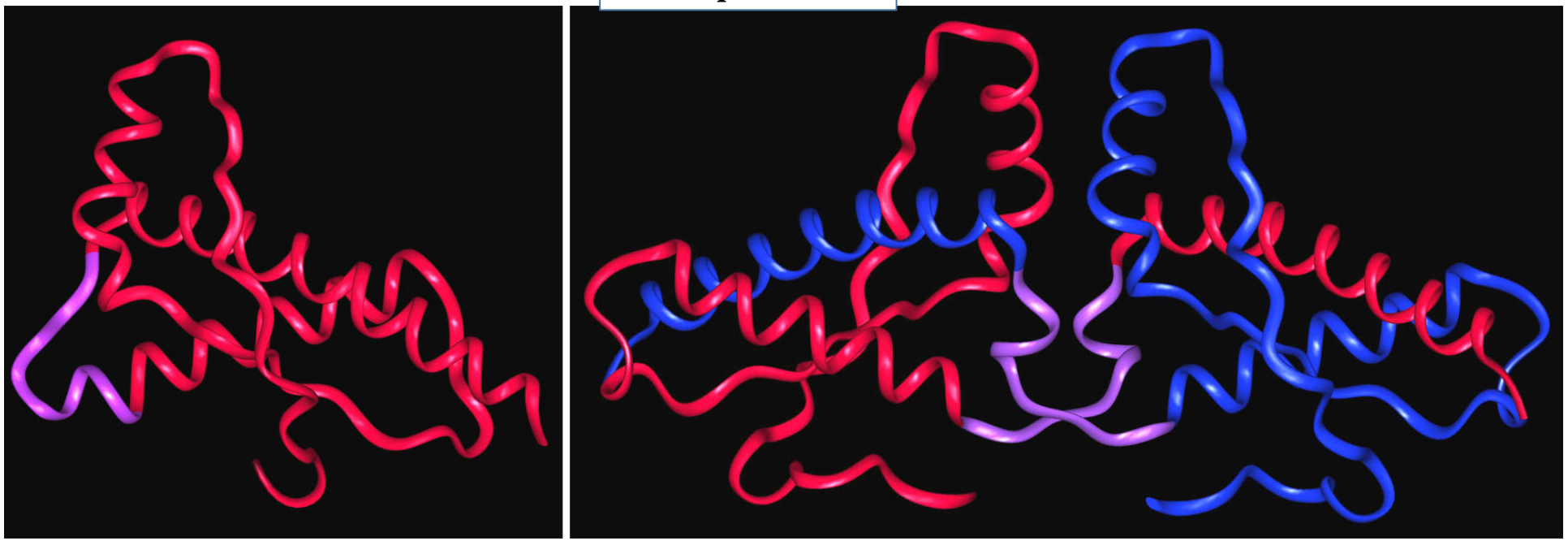
3D domain swapping

= exchange of an identical structural element between 2 monomers to generate an oligomeric unit

Purple: hinge loop

Blue and red: 2 chains

Prion protein



First protein observed in 3D domain swapped form: ribonuclease A in 1962

Now: ~50-100 proteins known in 3D domain swapped form

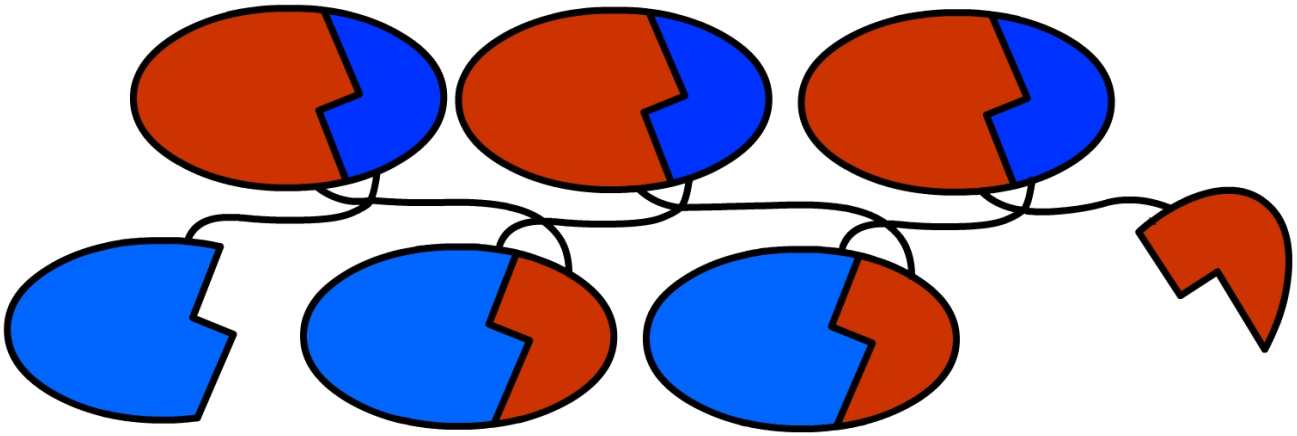
Some in physiological conditions, others not

These proteins have various sequences and structures

Biological importance

- Role in the regulation of the biological activity
- Possible evolutionary mechanism to build oligomeric structures
- Possible relation with conformational diseases: some proteins (prion protein, cystatin C ...) are known in aggregated form and in 3D domain swapped form

Model



Hypothesis :

3D domain swapping is due or facilitated by structural weaknesses

Definition of structural weaknesses

2 types

- regions of the sequence that show marked intrinsic preferences for non-native conformations in the absence of tertiary interactions

Predicted by programs that predict local structure, indicating regions of local structure that are intrinsically preferred – and that are different from the native structure (e.g. Prelude & Fugue)

Likely to slowing the folding process

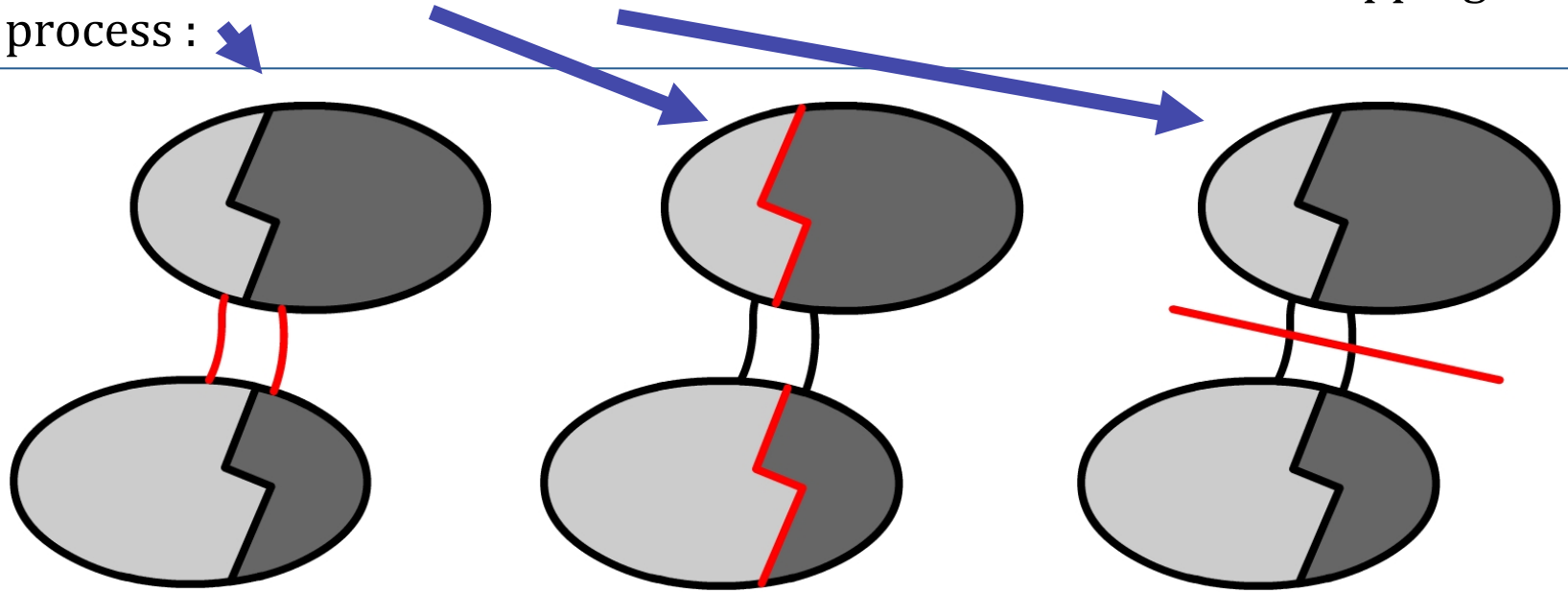
- sequence regions that are not optimal for the stability of the native structure

Predicted by programs predicting the change in stability upon mutations (e.g. PoPMuSic)

Regions of very negative N_i

Structural weaknesses are more often found in proteins that undergo 3D domain swapping than in other proteins of the same length.

In addition, these weaknesses are found in areas crucial to the swapping process :



Hinge loop | Closed interface | Open interface

These weaknesses are likely to influence the swapping process at a kinetic level (folding rate) or thermodynamic level (stability)

-> consistent with the starting hypothesis (but does not prove it ...)

Specific examples

Prion protein

Purple: hinge loop

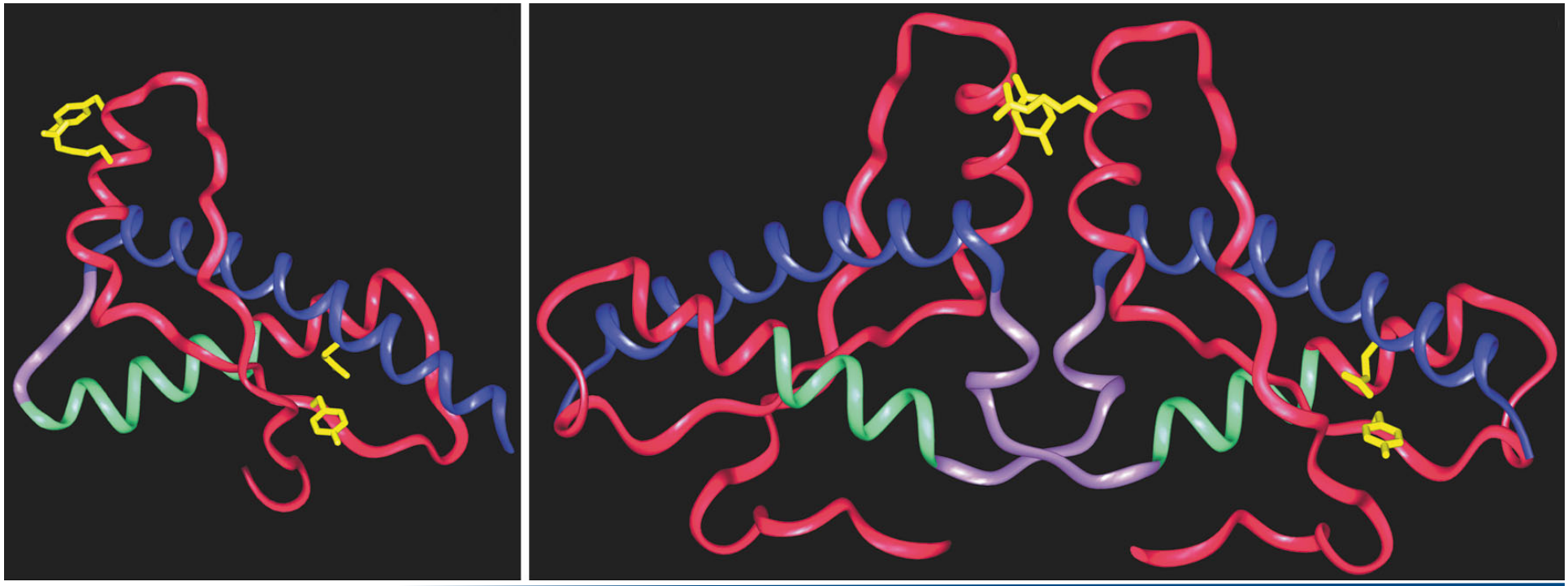
Blue: swapped part

Green: structural weaknesses

Yellow : cation-π interactions

Structural weakness at the hinge loop

Cation-π interactions at closed and open interfaces



Other examples: 3D domain swapping

Specific examples

Barnase -> 3D swapped trimer

Purple: hinge loop
Blue: swapped part
Green: structural weaknesses
Yellow : cation-π interactions

Structural weakness at the closed interface
Cation-π interactions at the closed interface