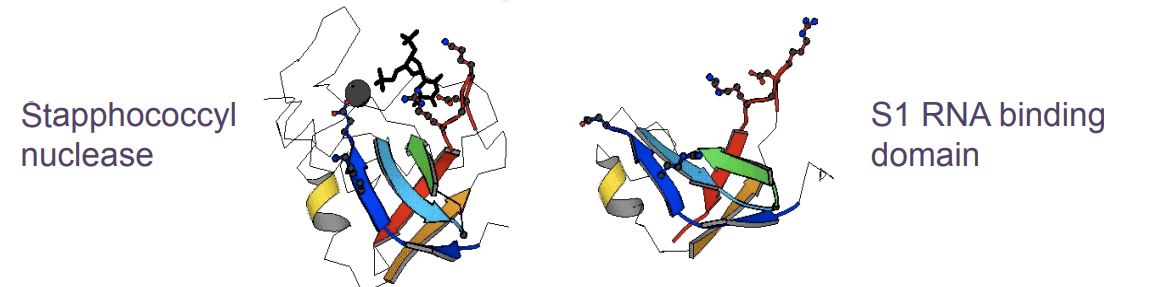
Regular secondary structures are connected by loops.

Loops are highly flexible and produce change of the polypeptide chain orientation.

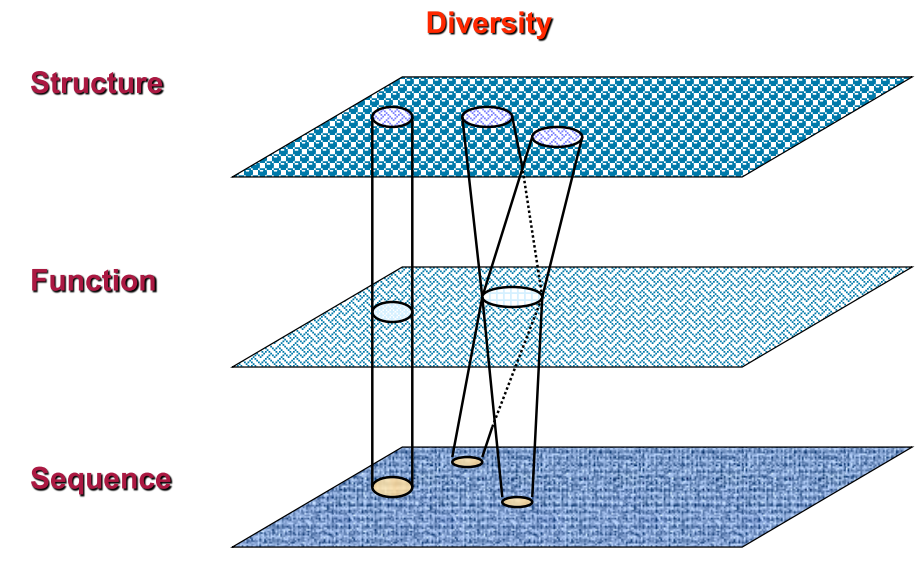
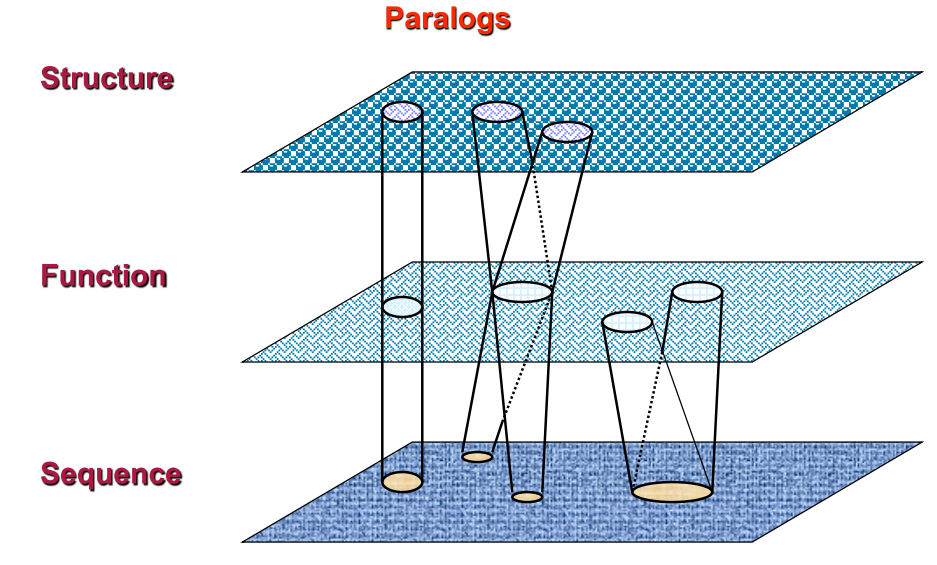
Length of loops varies from 1 to more than 40 residues

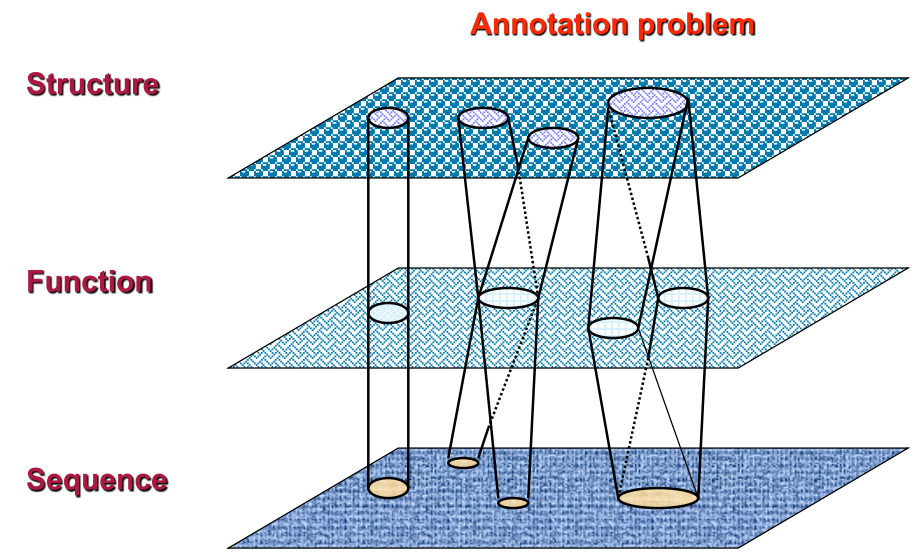
Some super-secondary structures are stabilized with a hydrophobic core



The structural alignment shows the best conservation of residues and a possible relationship in evolution.

Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor.



**Structural Classification of Proteins (SCOP)**

Class: It groups the folds according to the percentage and 3D disposition of the regular secondary structures.

Family: It groups proteins clearly related by homology. In general, we assume they are homologs if the alignment has >30% ID, they have the similar structure and similar function

Superfamily: Proteins which sequences align with very few %ID but showing similar structural patterns and similar function.

Fold: Proteins with very similar disposition of the regular secondary structures

**DATABASE CATH**

Class(C)

Class is determined according to the secondary structure composition and packing within the structure. It is assigned automatically

• Architecture(A)

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually

• Topology(T)

Structures are grouped into fold groups at this level depending on both the overall shape and connectivity of the secondary structures. This is done using an automated structure comparison algorithm.

• Homologous superfamily (H).

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are automatically identified either by high sequence identity or structure comparison.

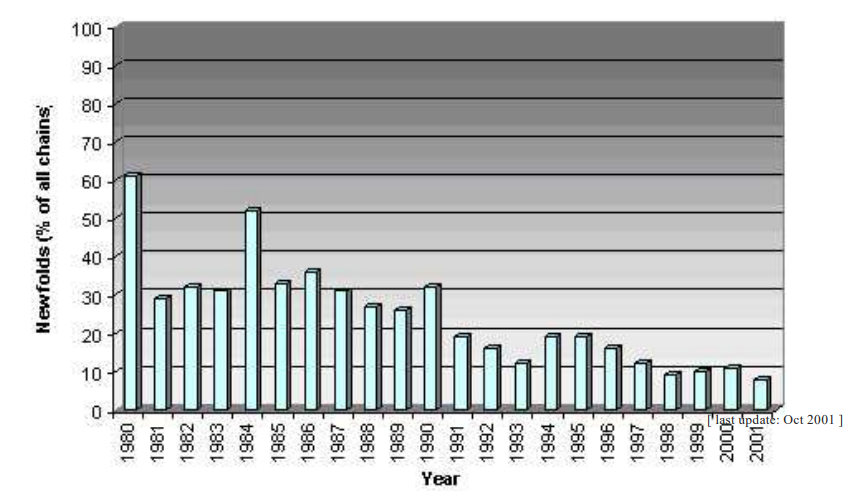
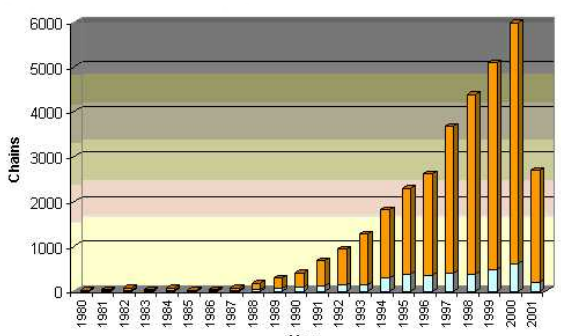
• Sequence Family Levels: (S,O,L,I, D)

Domains within each H-level are subclustered into sequence families using multi-linkage clustering S(35%), O(60%, L (90%), I (100%)

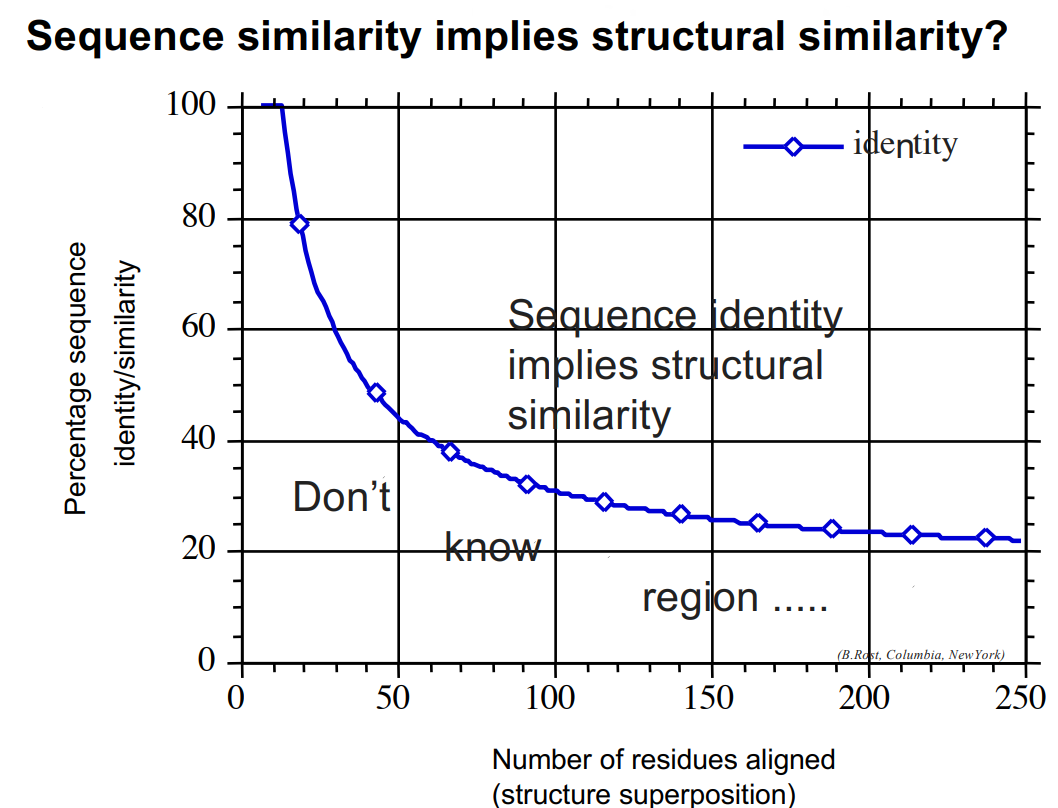
**PRINCIPLE OF FOLDS**

* Structures are more conserved than sequences, in order to preserve the function, they preserve the structure.
* Folds have a compact structure formed by the organization of regular secondary structures
* Folds have a hydrophobic core that stabilizes its conformation
* The hydrophobic core is formed by non-polar residues, while polar residues are mostly in the surface and active sites
* Structure is conserved among proteins with similar sequence
* The total number of folds is finite
* The structure of a protein a be built by joining several folds (this can be produced by gene fusion).

Definition of homology modelling: extrapolation of the structure for a new (target) sequence from the known 3D-structures of related family members (templates).

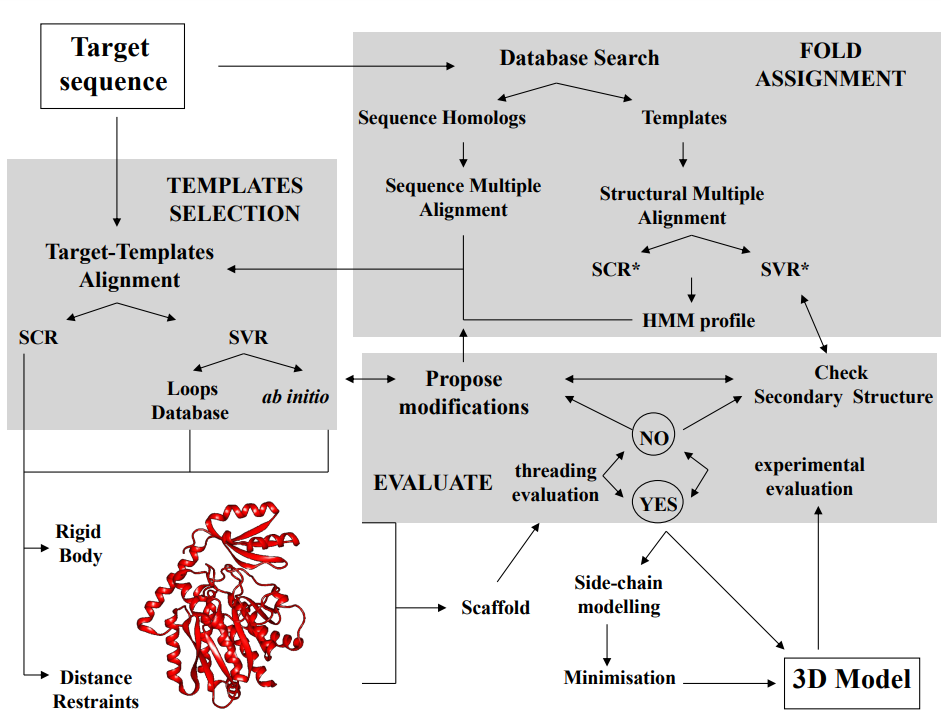


New structures found were with folds already seen

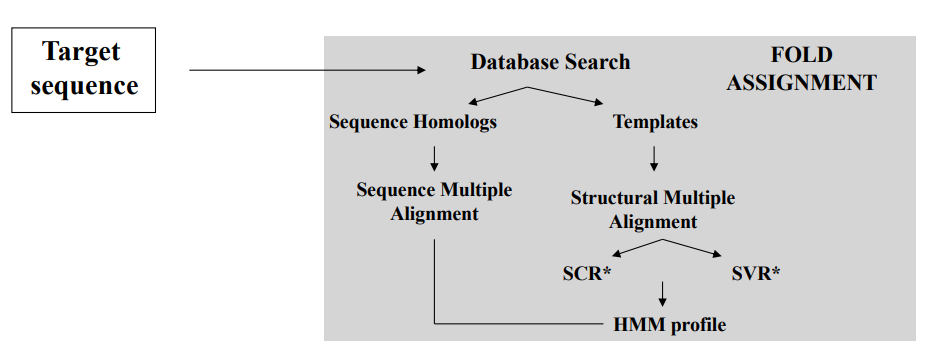
The dots are the number of identical residues

So, fold is more conserved than sequences (because of hydrogen bonds), especially secondary ones (alpha helix and beta strands are the most conserved) . Inside structures the loops are the ones that can variate more.

In order to have a reliable and accurate structures we need to have different sources of other structures. The process to complete it all is going to be with this schema:



**FOLD ASSIGMENT**

****

SCR conserved and SVR are the variables one.

1. Compares the sequence of the target with a set of sequences with known structure

2. Ranking the comparisons by scores.

3. Scores are related to P-values or E-values (high score implies low Pvalue). P-value is the probability of obtaining the same alignment by chance.

4. Scores are calculated using a residue-substitution matrix:

1. PAM: based on the alignment of sequences of homologs

2. BLOSUM: based on the alignment of blocs of similar sequences

5. One sequence can have more than one domain; therefore, we can obtain the best scores for partial parts of the target.

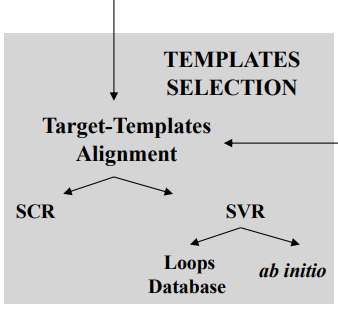
6. Methods (see practice)

1. BLAST algorithm, matches words from a pre-calculated and indexed set and joints them into sentences (forming the sequence)

2. FastA: Smith & Waterman algorithm

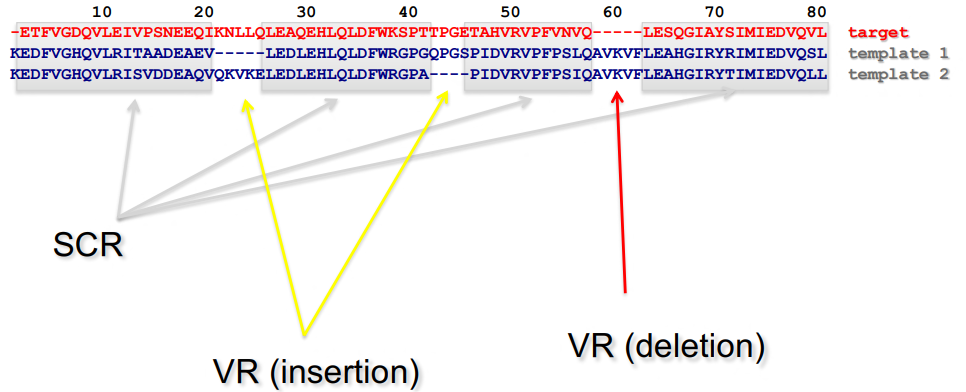
3. Scanning PFAM: algorithm of Hidden Markov Models

**TEMPLATE SELECTION**

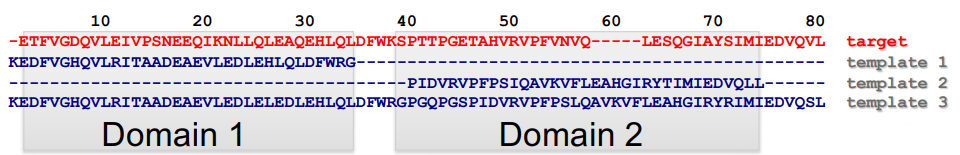


After the fold we need to get our template in order to discover/know our structure.

For some cases we may need more than one template, this can be done by ClustalW.



We also can combine different template in order to mix knowledge



We can build our model by Rigid Body Assembly, for the side chains we will use another one.

First, we assembly the core of the proteins the main chain, loops from variable region and we also want to minimize the energy of the folding.

Model the probability of density function and the model of the loops

Finally, we go for the side chains and it rotations of the amino acids and the energetic criteria.

**RIGID BODY ASSEMBLY**

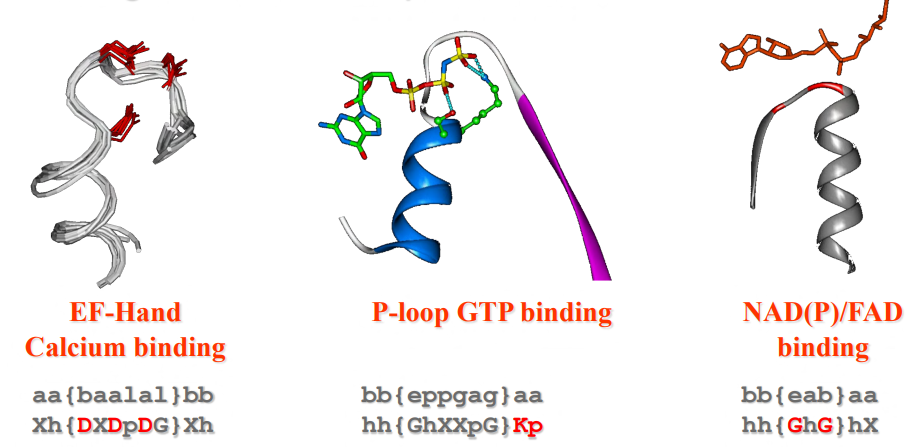
• Averaging core template backbone atoms (weighted by local sequence similarity with the target sequence)

• Leave non-conserved regions (loops) for later …

Two ways to approach the loop part:

Use the “spare part” algorithm to find compatible fragments in a **Loop-Database**

“ab-initio” rebuilding of loops (Monte Carlo, molecular dynamics, **genetic algorithms**, etc.)

Loop-Database

**Model building: Rigid Body Assembling**

(Energy minimization)

• Modelling will produce unfavourable contacts and bonds: idealization of local bond and angle geometry

• extensive energy minimization will move coordinates away: keep it to a minimum

• Methods: Newton Rapson; Steepest Descent; Conjugate Gradient

We also use Ramachandran plot

**Feature properties can be associated with:**

• a protein (e.g. X-ray resolution)

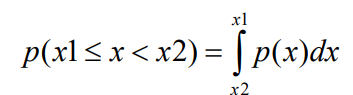
• residues (e.g. solvent accessibility)

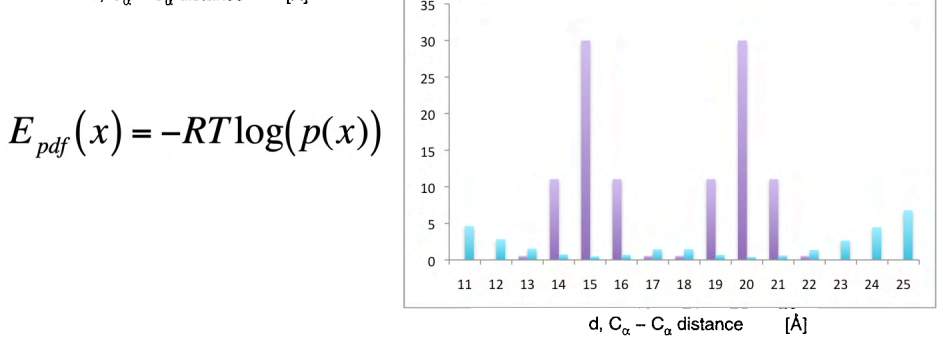
• pairs of residues (e.g. Ca - Ca distance)

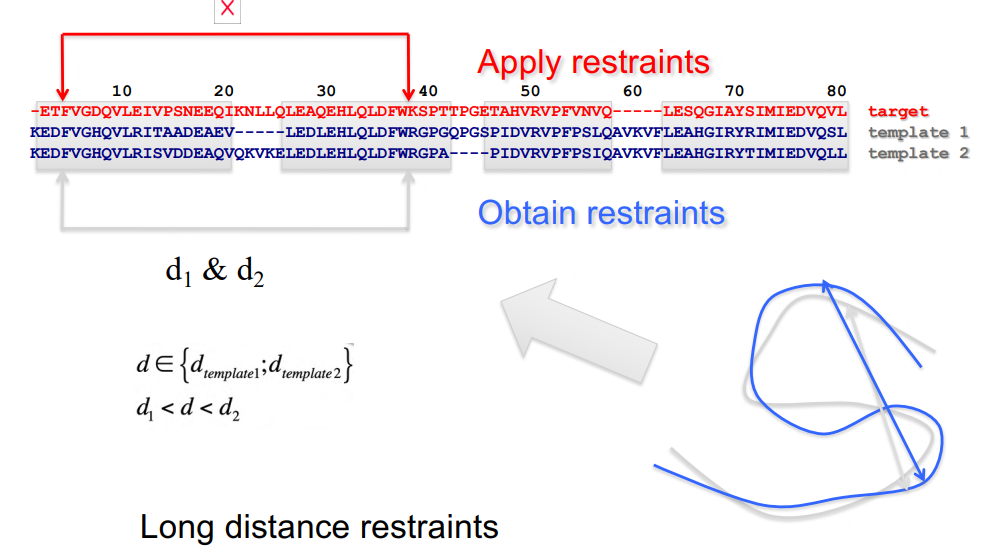
• other features (e.g. main chain classes)

How we transform this data into data

A restraint is defined as probability density function (pdf), p(x)







Distance restraints between Aa in SCR & VR (required to locate the conformation of the VR)

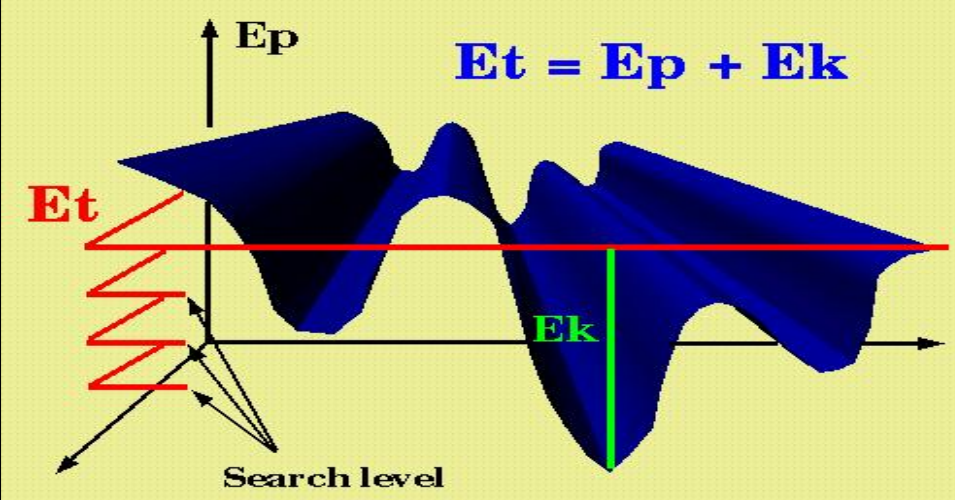
**Spatial restraints (Simulated annealing)**

Optimizing a target function:

1. Start with e.g., a random conformation model and use only local restraints.

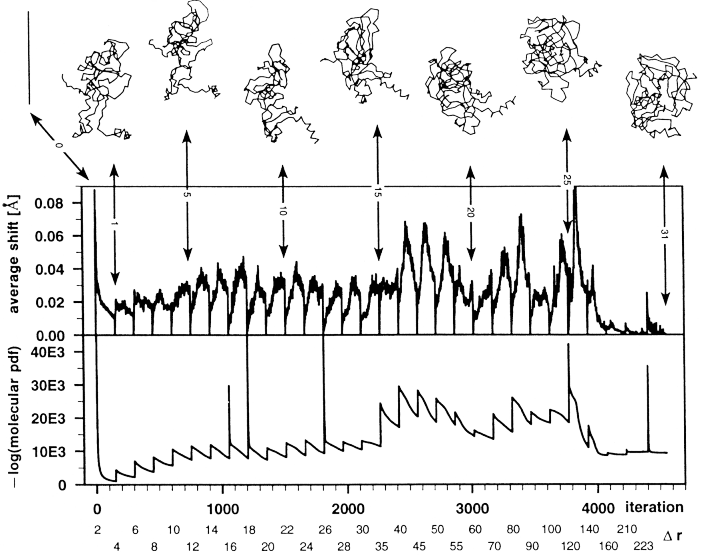
2. Minimize some steps using a conjugate gradient optimization and molecular dynamics steps.

3. Repeat, introducing more and more long-range restraints until all restraints are used.



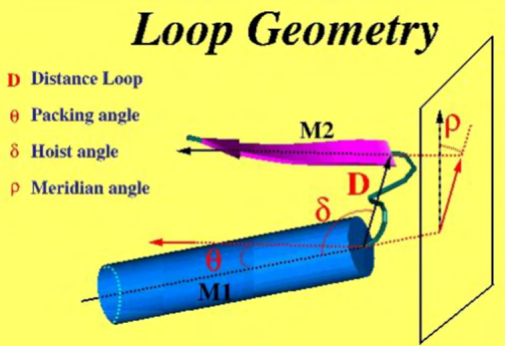
The total energy remains constant but for the kinetic part depends on the molecules of the system.

When simulation we can increase or remove temperature in order to see where the system stabilizes.



Here we see all the possible conformations made by the model with its correspondence energies as we variate the temperature . From here we can see and pick the one (best). When we reach the maximum number of restraints, we stop the iteration.

**LOOPS MODELLING**



Using the structure of a known loop:

1. The C-tail and N-tail of the loop (template 2) when superposed with the core of the main template (template 1) produce a low RMSD.

2. The selection of the loop follow two criteria: similar sequence profile with the target and similar anchoring geometry of the loop with the main template.

If we don’t have any template, we can use pdf’s.

**Using PDF of loops and minimization methods:**

1. Calculate specific PDF residue properties of loops

2. Minimize by simulated annealing the loops

3. Extract main motion from normal modes on templates and apply them as restrictions on the conformational changes of the model

4. Methods:

1. Loop-model from MODELLER

2. ArchPred

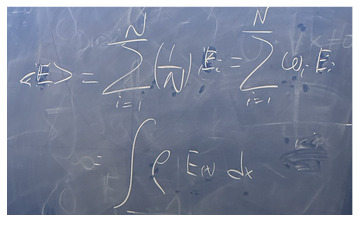
3. Rosetta

**SIDE-CHAINS**

Let be a rotamer library, we define the probability of side-chain “i” in conformation “k” as CM(i,k). Initially CM(i,k)=1/Ki, where Ki is the total of rotamers of residue “i”.

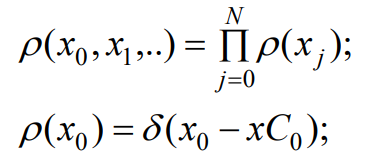
W equals to weight

1/n the probability

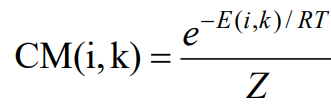


The probability of the whole scaffold comes from the product of all the probabilities of the conformations.

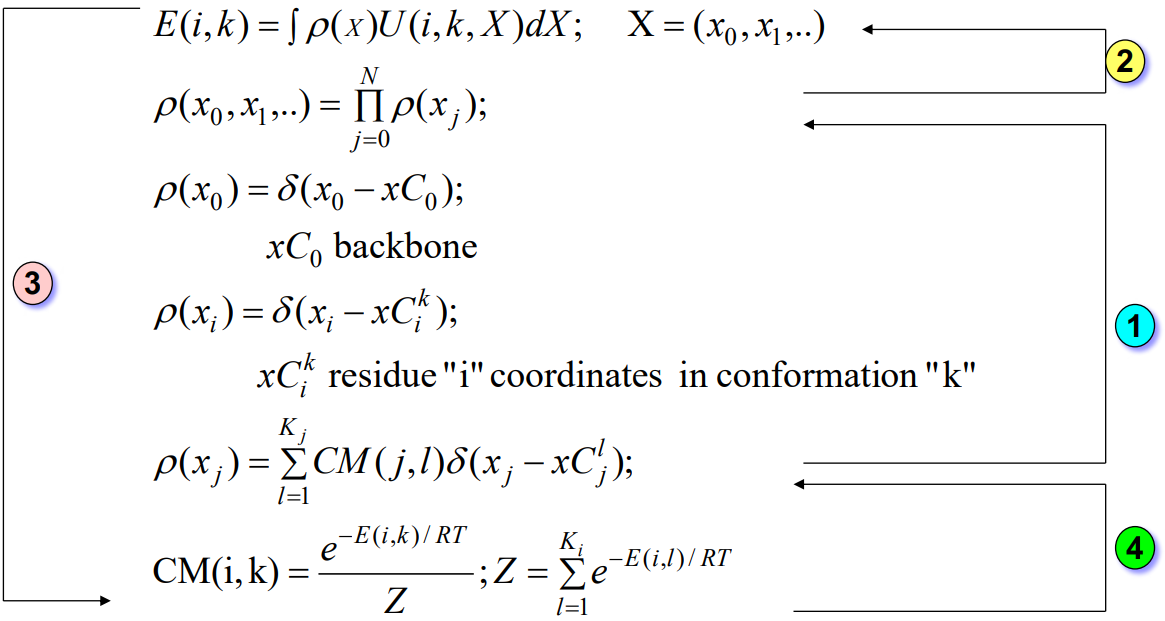
For the side chains we may have more than one result.



Once we have the energy we can recalculate the probabilities, taking into account the average energy for each position.

 **RECALCULATING EQUATION**

The process is repeated until a convergence is reached:



We also need to take into the equation the waters that are in the structure, but we do not.

**TYPE OF ERRORS**

1. Errors in side-chain packing .

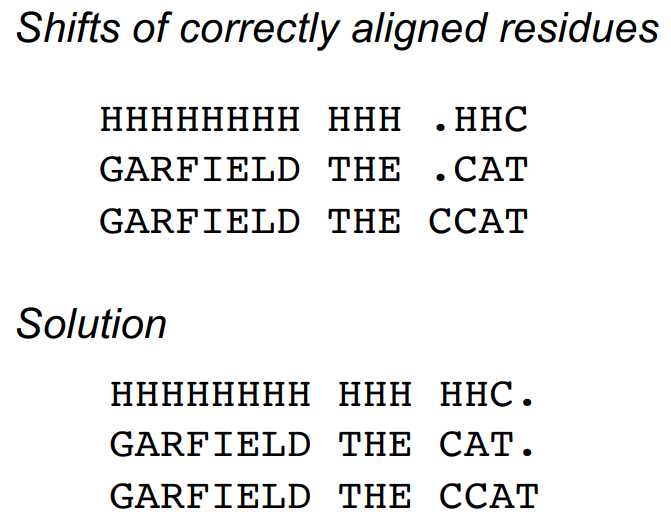
2. Shifts of correctly aligned residues .

3. Regions without template .

4. Errors due to misalignments .

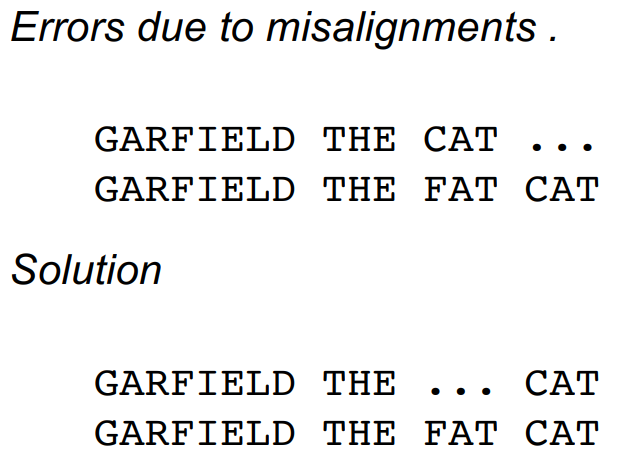
5. Errors produced by incorrect templates (biggest one).

**Second error**

****Dot = gap

As we break the align, we reduce the number of hydrogen bonds and our enthalpy goes shitty.

Fourth error



As CAT and FAT resemble but they are not the same we could get a misalignment, with tea coffee is solved but there are still problems with ClustalW.

**How to test the model?**

1. Compare the RMSD (root mean square deviation of atomic position) between the model and the real structure

2. Check that secondary structures are correctly aligned

3. Calculate the percentage of residues that are closer than a threshold after superposing the model and the real structure

4. Calculate the percentage of identical residues aligned when superposing the real structure and the model.

5. Check the energy of threading to compare the real structure and the model (see next chapter)**.**

**How to detect possible errors in the model if we don’t know the solution?**

1. Compare the model and all the templates

2. Check that secondary structures are not broken

3. Check if the prediction of secondary structure agrees with the secondary structure of the model

4. Check if the loops of the target are similar to some loops in the database of loops and they agree in sequence and anchoring geometry

5. Check the capping of helices 6. Check pseudo-energies of threading and compare the model with the templates.

**How to improve the model?**

1. Decide the changes in the alignment according to the secondary structure prediction or the structure of the templates and recalculate the model

2. Change the main template and recalculate the model

3. Include new templates

4. Calculate the main motion of normal modes from the templates of the homologous family and optimize by molecular dynamics under motion restrictions the conformation

5. Recalculate the pseudo energy profile of the new model and compare with the original model to test the improvement

**FOLD PREDICTION**

We have 3 ways:

Fold recognition (threading)

2. ab initio fold prediction

3.Protein folding (MD (molecular dynamics) with explicit solvent)

Threading

Idea: Find the optimal structure for a new (target) sequence in the set of known 3D-structures (templates) by threading the target sequence.

**ENERGY CONSIDERATION**

Entropic : conformational and hydrophobic

Enthalpic : Vdw (shape and contacts), Hbond (define geometry, must be complete), electrostatic (severe solvation penalty).

Complex formation implies to bury new interactions: unstable (hydrophobic) surfaces in water may indicate binding regions.

Protein – protein complexes: permanent associations very easy to identify them (“quaternary structure”), enzyme – substrate (transient), regulatory associations, multi protein clusters/groups (Nuclear porus, cytochromes, …).

Protein – NA complexes: transcription factors, replication, splicing transcription machineries.

**Docking: Prediction of the structure of complexes:** Ligand-protein docking, protein docking.

**Receptor and ligand are the actors (normally the bigger is the receptor)**

**Pose:** how the ligand positions on the receptor site

**Interface:** Contact region in a complex

**Scoring**: How the binding is evaluated

**(Virtual | Reverse) screening**: test for the feasibility of binding among a high number of ligands/receptors.

**Docking evaluation**: How to test the success of the docking

**PROTEIN DOCKING TERMS**

Interaction prediction: if two proteins interact to each other

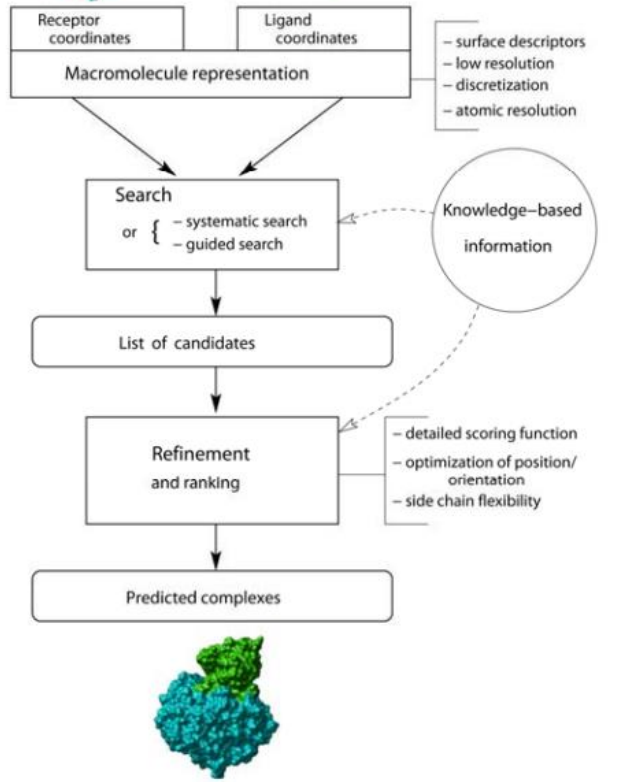
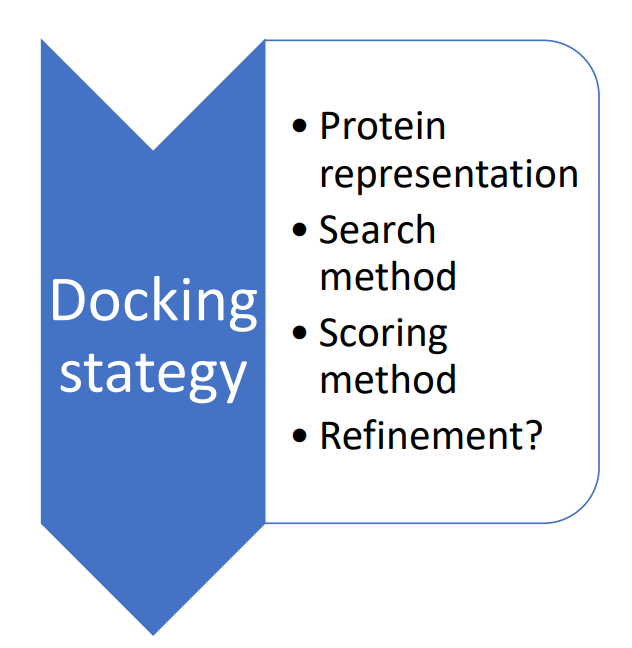
Interface prediction: which regions interact with each other

Protein docking: predict the position (poses, structure of the complex)

Bound vs unbound docking : Docking using conformations in the complex (bound) or free (unbound)

Flexible vs Rigid docking : Protein flexibility is taken into account

Local vs global docking : whether binding site is roughly known



Systematic search means to try everything, every position every orientation.

**LIGAND-PROTEIN DOCKING**

**3 main levels**

Molecular docking (first level)

* Prediction of 3D structure of ligand-protein or protein-protein complexes.
* One receptor to **one or few** ligands
* We centre in the quality of the structure; we need to resemble to a structure of the PDB. Realistic binding energies
* Usually combine with other techniques like MD.
* We could consider experimental information

Virtual screening (second level)

* Identification of possible ligands from compound databases.
* One receptor – multiple ligands (> 106 )
* Calculation should be fast >10000 ligand-receptor dockings / day / proc.)
* The main objective is to select “some” ligands, that can be optimized with other methods

Reverse screening (third level)

* Identification of possible receptors for a known ligand.
* One ligand multiple receptors.
* Points to possible side effects.

Active site prediction

* Identification of binding regions

**LIGAND-PROTEIN DOCKING**

• Complete atomic representation: large cost, high resolution

• Simplified representations: quick and robust. Low resolution.

• 3D Grid representations (receptor): easier energy calculation. Definition of “pharmacophores” (MIPs)

•Flexibility (Ligand and receptor): ensembl docking. Ligand has less flexibility than a protein.

**SCORING**

When to score : score associated to the search process, **scoring a posteriori.**

Structural complementarity: robust and low resolution.

Classical force-fields, statistical pot : high resolution and easy to transfer.

**BASIC STRATEGIES**

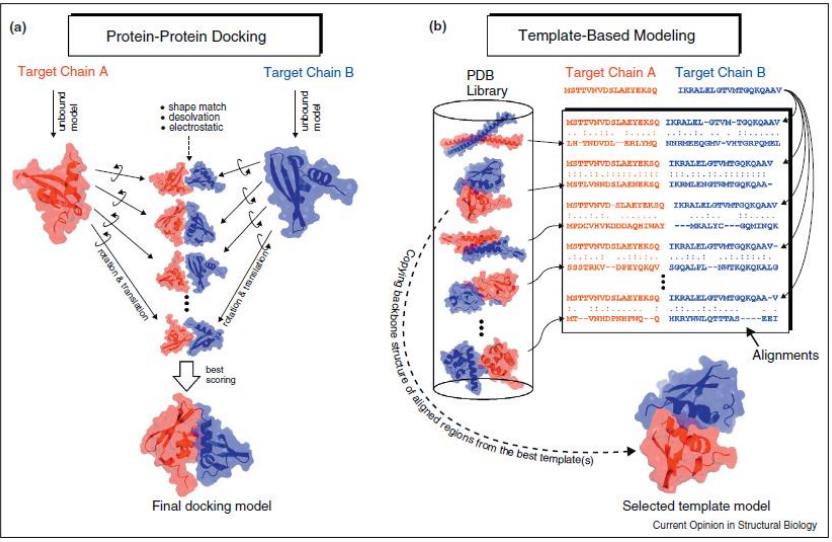
“Pure” Ab Initio docking

We only have info about the ligand and the receptor structures is known:

* Pseudo random approaches
* Direct search (geometric hashing)
* Brute force approaches

Data driven docking (template based)

* Experimental, homology data is used
* Machine-learning methods
* Co-evolution methods
* Can be combined to help ab initio approaches



Ab initio Rigid-Body docking

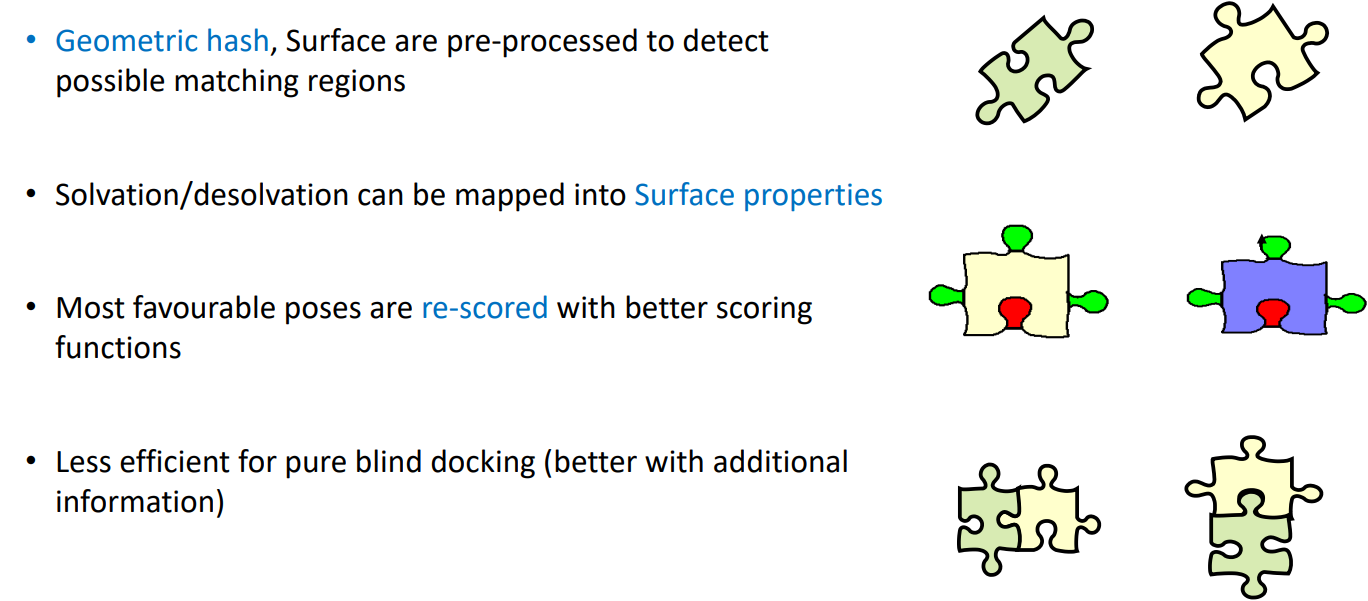
Proteins are mapped onto 3D grids. Each grid point is evaluated as inner (0 for ligand), Surface (1), outer (0).

Blind 6Dim (3 translations x 3 rotations) search.

Score is based in 3D complementarity: i.e. matches among the “Surface” points (calculated with from the product of grid points).

**Fast Fourier Transforms** to speed up translational (Fast Fourier Transform, FFT) or rotational (Spherical Polar Fourier, SPF) searches.

Computational cost can decrease by >104 (from N6 to N3 lnN3 ).



**ADDITIONAL TRICJS**

**All this tricks are to allow some flexibility in the structure.**

Flexible docking

* Traditional MD: expensive, used to refine structures after docking.
* Conformational sampling: rigid docking with set of possible. Conformational search added to position and orientation.
* Combined cycles of docking and simulation
* RosettaDock combines rigid body MonteCarlo for orientation/translation + MonteCarlo among rotamer libraries (very expensive)

Soft docking

* Backbone still rigid
* Sidechain flexibility is mimicked using soft Vdw potentials, and or coarser FFT grids.

**DATA-DRIVEN METHODS**

Homology/threading based methods: Template based, use data from homologues (PDB).

Coevolution methods: Growing popularity in protein structure prediction (Alpha fold). Uses data from “massive” multiple sequences alignment

Interface prediction: To reveal interfaces without structure prediction