Prediccion de la estructura de Proteinas

Major Algorithmic Tasks :

- Structural Alignment of Proteins and their Classification.
- Functional Annotation.
- Protein Structure Modelling
- Prediction of Protein Interactions and the Structure of Complexes.
- Computer Assisted Drug Design.
- Protein Design.
- Alignment and modeling of RNA structures.
- Modeling of DNA 3D structure (HiC).

Prediccion de la estructura de Proteinas

Protein Structure Prediction-Folding

- Given only the amino-acid sequence of a protein, deduce its native tertiary structure.

Target sequence

M	A	A	G	Y	A	V	L	S

→

structural model

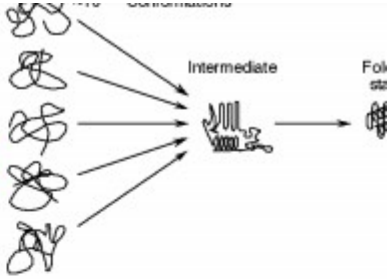
Prediccion de la estructura de Proteinas

Protein structure

- Most proteins will fold spontaneously in water
  - amino acid sequence should be enough to determine protein structure
- However, the physics are daunting:
  - 20,000+ protein atoms, plus equal amounts of water
  - Many non-local interactions
  - Can take seconds (most chemical reactions take place ~10^{12} --1,000,000,000,000x faster)

Prediccion de la estructura de Proteinas

Levinthal Paradox



- Cyrus Levinthal, Columbia University, 1968
- Levinthal's paradox
  - *If we have only 3 rotamers (α,β,λ) per residue a 100 residue protein has 3^{100} possible conformations.*
  - *To search all these takes longer than the time of the universe, however, proteins fold in less than a second.*

- Resolution: Proteins have to fold through some directed process
- Goal - to understand the dynamics of this process

Prediccion de la estructura de Proteinas

Protein Folding vs Structure Prediction

- Protein folding investigates the ***process*** of the protein acquisition of its three-dimensional shape.
  - The role of statistics is to support or discredit some hypotheses based on physical principles.
- Protein structure prediction is solely concerned with the final ***3D structure*** of the protein
  - use theoretical and empirical means to get to the end result.

Prediccion de la estructura de Proteinas

Methods of Structure Prediction

- Homology modeling
  - Easy cases
  - high seq. identity to known structures
- Fold recognition
  - No discernable sequence identity to a known structure
  - a similar fold is (probably) known but hard to identify
- Ab initio (de novo) methods
  - Most difficult
  - No similar folds are known

Prediccion de la estructura de Proteinas

Fold Recognition – Threading

The **RAPTOR Algorithm**

- Jinbo Xu's Ph.D. thesis work.
- J. Xu, M. Li, D. Kim, Y. Xu, *Journal of Bioinformatics and Computational Biology*, 1:1(2003), 95-118.

Prediccion de la estructura de Proteinas

There are not too many candidates!

- There are only about 1000 – 1500 topologically different domain structures. Fold recognition methods aim to assign the correct fold to a given sequence and to align the sequence to the chosen fold.

Prediccion de la estructura de Proteinas

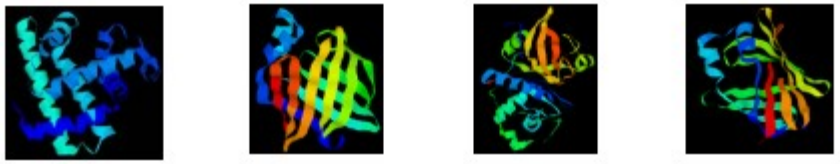
Protein Threading

- Make a structure prediction through finding an optimal placement (threading) of a protein sequence onto each known structure (structural template)
  - "placement" quality is measured by some statistics-based energy function
  - best overall "placement" among all templates may give a structure prediction

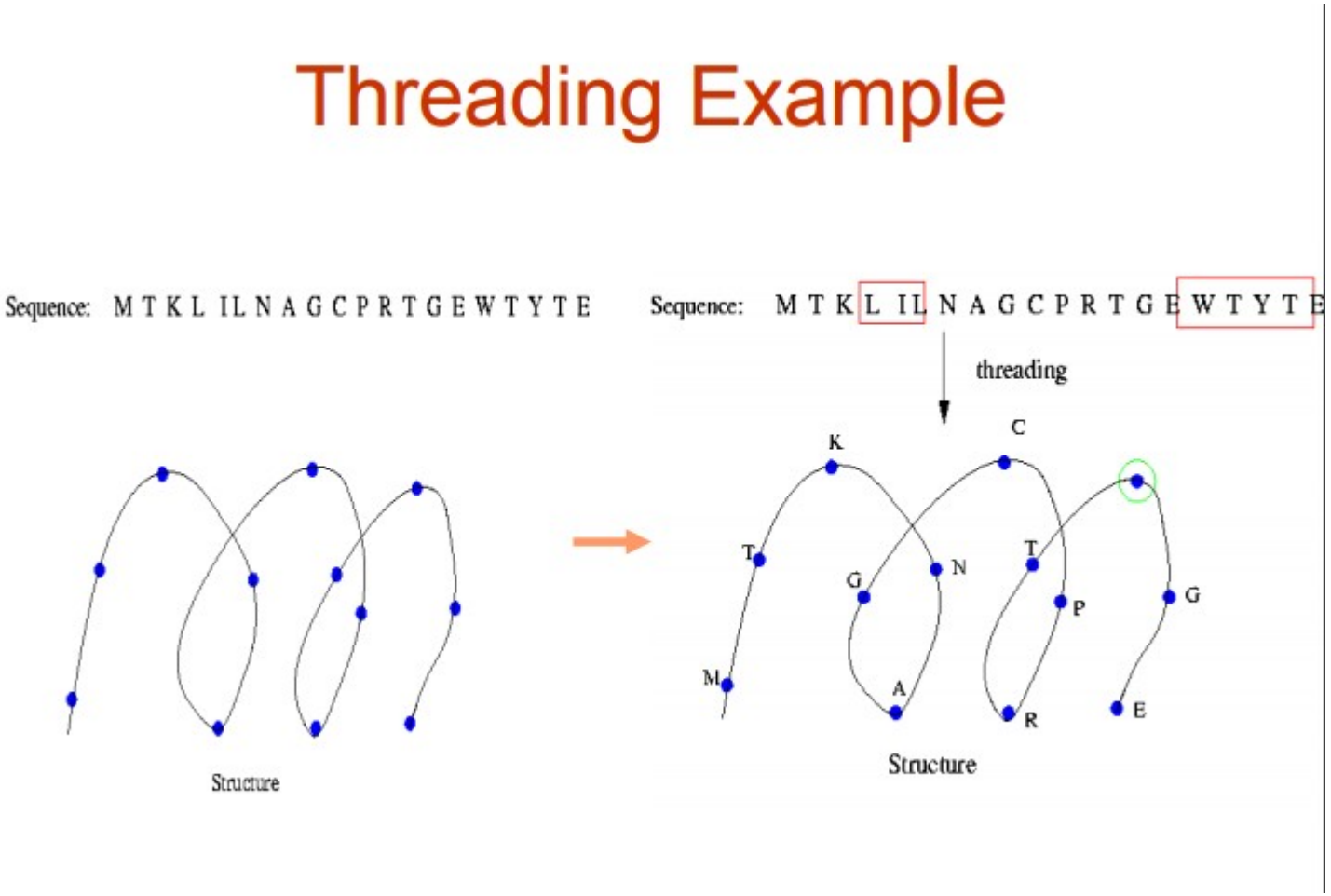
target sequence

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

template library



Prediccion de la estructura de Proteinas



Prediccion de la estructura de Proteinas

Formulating Protein Threading by LP

- Protein Threading Needs:
  1. Construction of a Structure Template Library
  2. Design of an Energy Function
  3. Sequence-Structure Alignment algorithm
  4. Template Selection and Model Construction

Prediccion de la estructura de Proteinas

Assumptions :

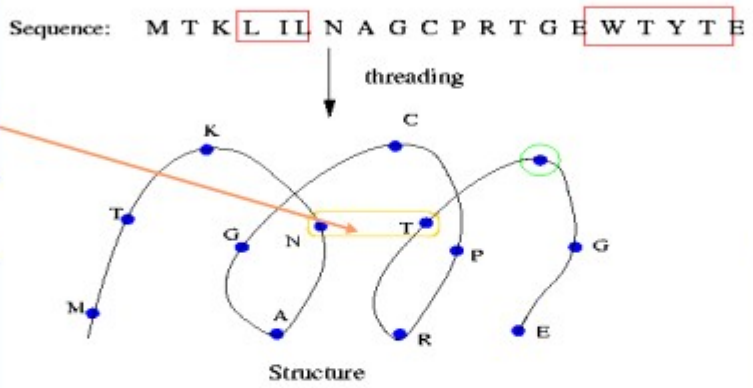
1. Each template sequence is parsed a linear series of (conserved) cores connected by (variable) loops. Each core is a conserved part of an α–helix or β-sheet.
2. Alignment gaps are confined to loops.
3. Only interactions between residues in cores are considered. An interaction is defined btwn two residues, if they are at least 4 positions apart in the sequence and the distance btwn their Cβ atoms is less than 7A.
4. An interaction is defined btwn two cores if there is at least one residue-residue interaction btwn the cores.

Prediccion de la estructura de Proteinas

Threading Energy Function

how preferable to put two particular residues nearby: E_p

(Pairwise potential)



how well a residue fits a structural environment: E_s

(Fitness score)

alignment gap penalty: E_g

(gap score)

sequence similarity between query and template proteins: E_m

(Mutation score)

Consistency with the secondary structures: E_{ss}

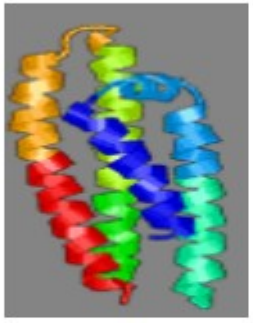
$$E = E_p + E_s + E_m + E_g + E_{ss}$$

Minimize E to find a sequence-structure alignment

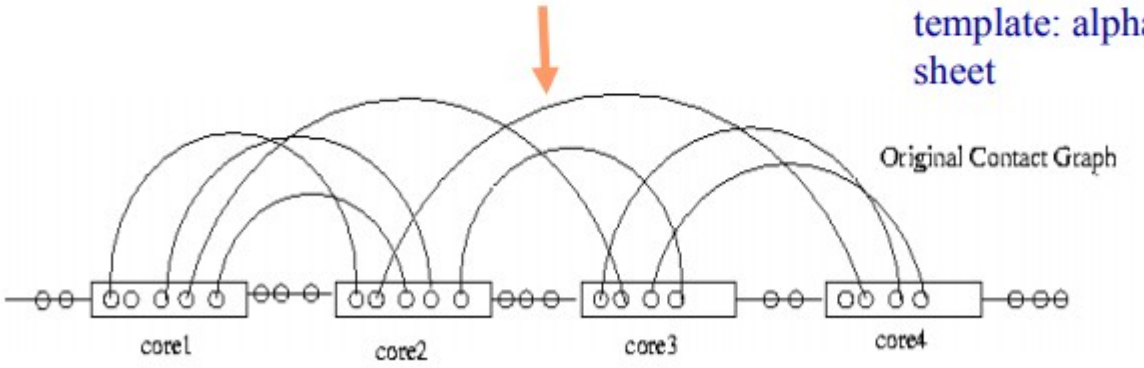
Prediccion de la estructura de Proteinas

Contact Graph

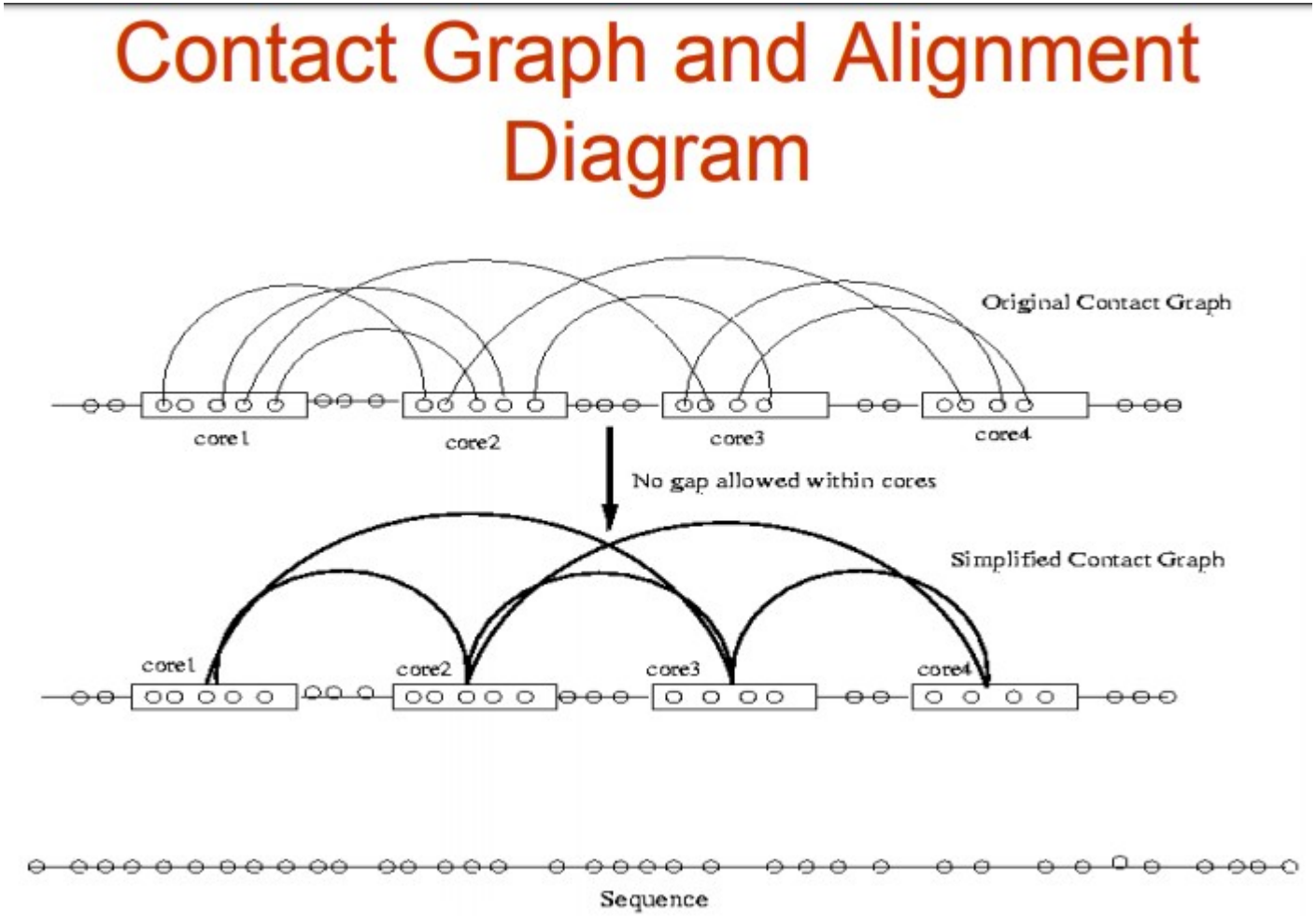
template



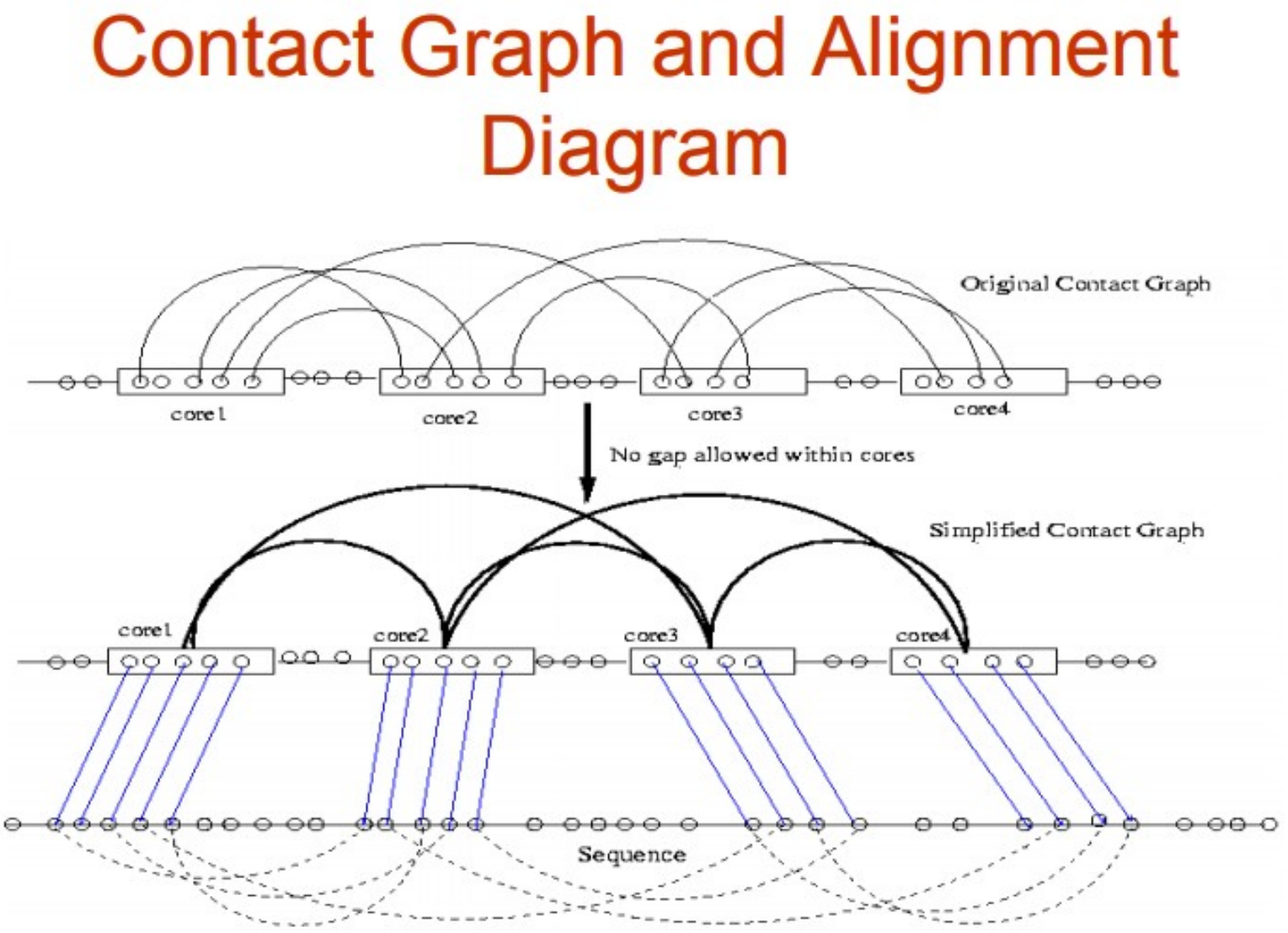
1. Each residue as a vertex
2. One edge between two residues if their spatial distance is within a given cutoff.
3. Cores are the most conserved segments in the template: alpha-helix, beta-sheet



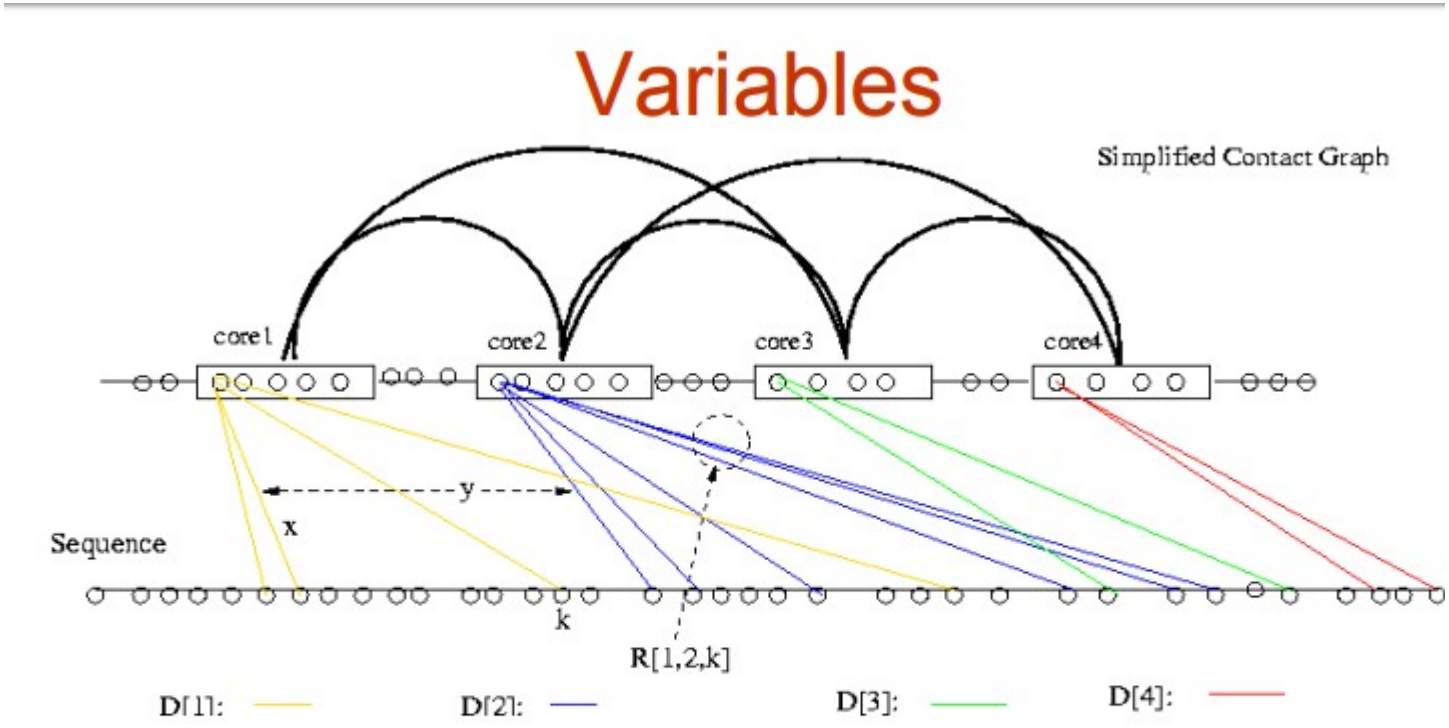
Prediccion de la estructura de Proteinas



Prediccion de la estructura de Proteinas



Prediccion de la estructura de Proteinas



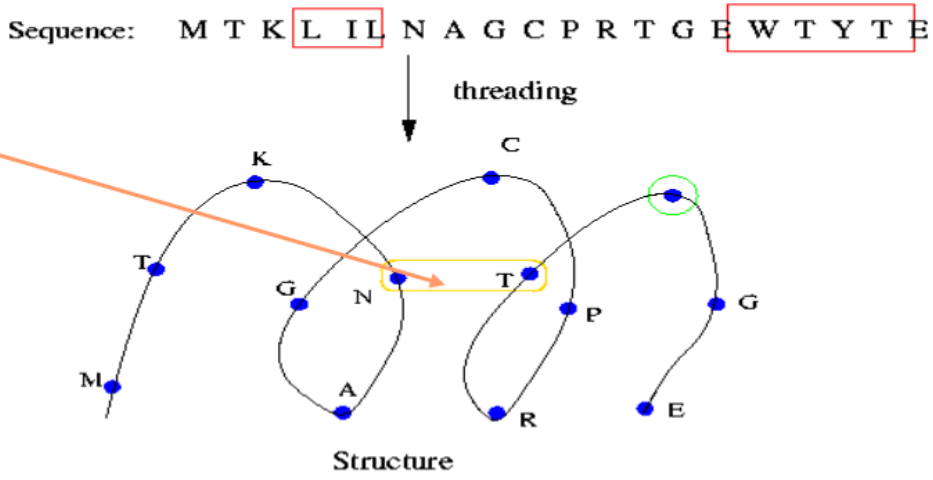
- x(i,l) denotes core i is aligned to sequence position l
- y(i,l,j,k) denotes that core i is aligned to position l and core j is aligned to position k at the same time.
- D[i] – valid alignment positions for c(i).
- R[i,j,l] – valid pos. of c(j) given that c(i) is aligned to s(l).

Prediccion de la estructura de Proteinas

Threading Energy Function

how preferable to put two particular residues nearby: E_p

(Pairwise potential)



how well a residue fits a structural environment: E_s

(Fitness score)

alignment gap penalty: E_g

(gap score)

sequence similarity between query and template proteins: E_m

(Mutation score)

Consistency with the secondary structures: E_{ss}

$$E = E_p + E_s + E_m + E_g + E_{ss}$$

Prediccion de la estructura de Proteinas

Formulation used in RAPTOR

$$\textit{Minimize}$$

$$E = \sum a_{i,l} x_{i,l} + \sum b_{(i,l)(j,k)} y_{(i,l)(j,k)}$$

$$\textit{s.t.}$$

$$x_{i,l} = \sum_{k \in R[i,j,l]} y_{(i,l)(j,k)}, \forall l \in D[i]$$

$$x_{j,k} = \sum_{l \in R[j,k,i]} y_{(i,l)(j,k)}, \forall k \in D[j]$$

$$\sum_{l \in D[i]} x_{i,l} = 1$$

$$x_{i,l}, y_{(i,l)(j,k)} \in \{0,1\}$$

E_g, E_p

E_s, E_{ss}, E_n

Encodes scoring system

Encodes interaction structures

Ver: RAPTOR

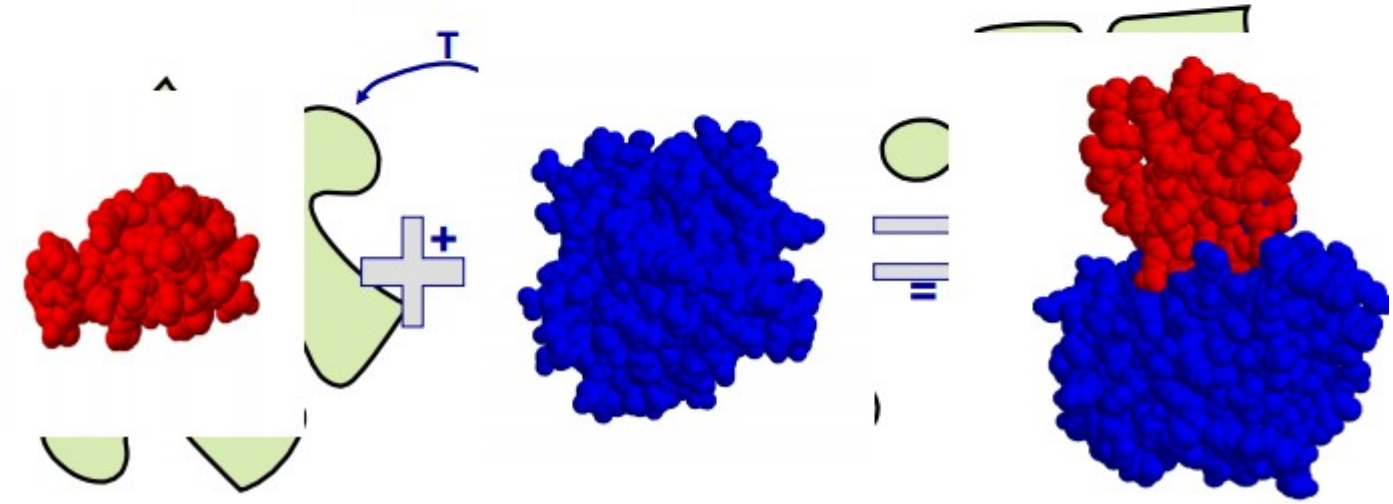
Solving the Problem Practically

1. More than 99% threading instances can be solved directly by linear programming, the rest can be solved by branch-and-bound with only several branch nodes
2. Relatively efficient
3. Easy to extend to incorporate other constraints

Prediccion de la estructura de Proteinas

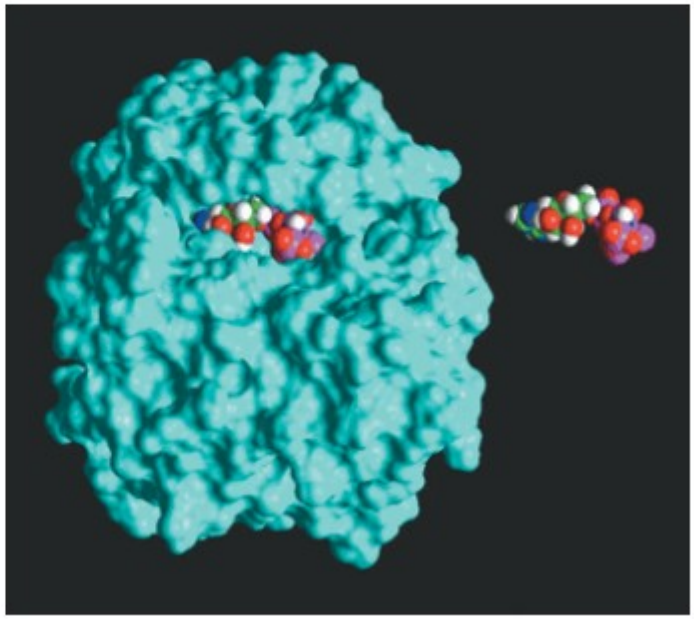
Docking Problem

Given 2 input molecules in their native conformation, the goal is to find their correct association as it appears in nature.



Prediccion de la estructura de Proteinas

Detection of a Lead Drug Compound : The Key-in-Lock Principle



Prediccion de la estructura de Proteinas

Docking - Motivation

- Computer aided drug design – a new drug should fit the active site of a specific receptor.
- Understanding of the biochemical pathways - many reactions in the cell occur through interactions between the molecules.
- Crystallizing large complexes and finding their structure is difficult.

Prediccion de la estructura de Proteinas

The Docking Problem

- Input: A pair of molecules represented by their 3D structures.
- Tasks :
  - Decide whether the molecules will form a complex (interact / bind).
  - Determine the binding affinity.
  - **Predict the 3D structure of the complex.**
  - Deduce function.

Prediccion de la estructura de Proteinas

Forces Governing Biomolecular Recognition

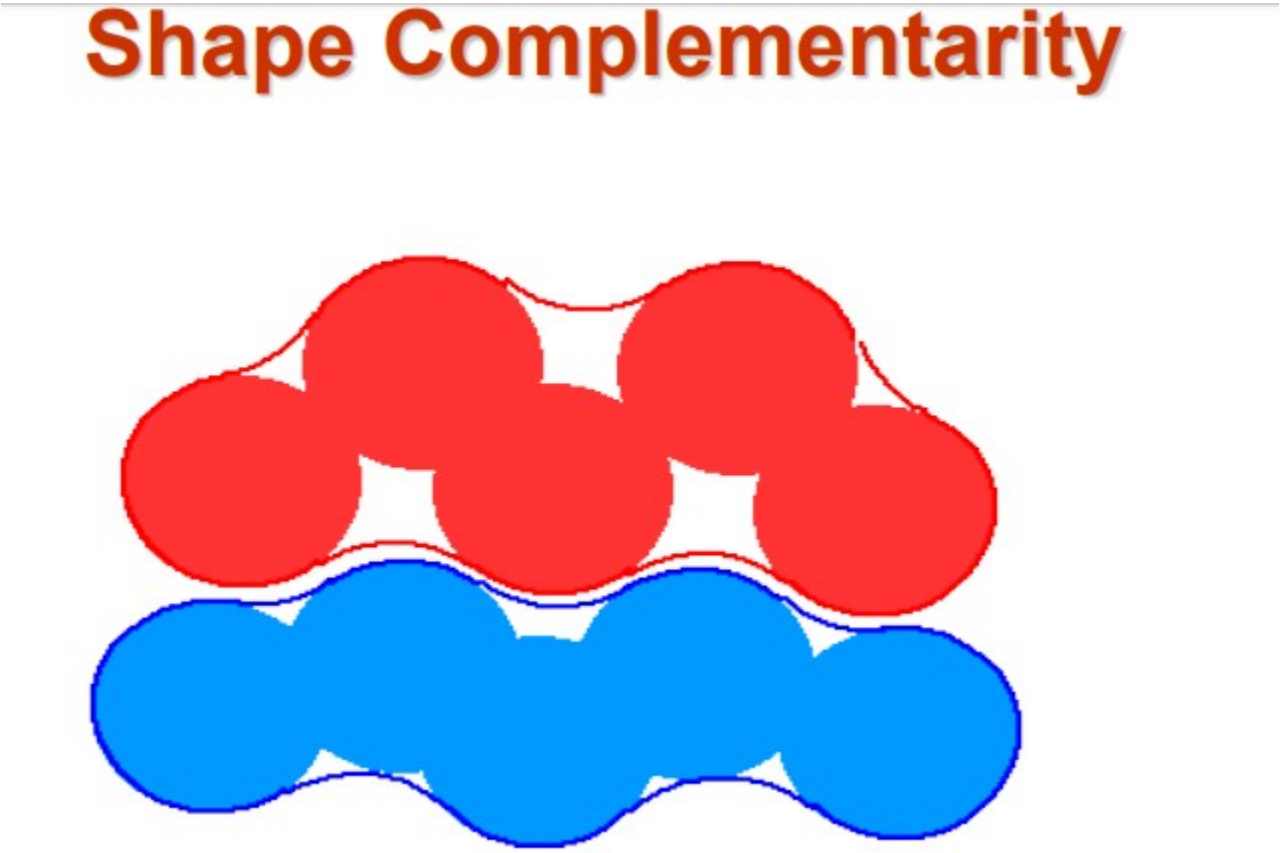
Depend on the molecules and the solvent.

- Van der Waals.
- Electrostatics.
- Hydrophobic contacts.
- Hydrogen bonds
- Salt bridges .. etc.

All interactions act at short ranges.

Implies that a necessary condition for tight binding is surface complementarity.

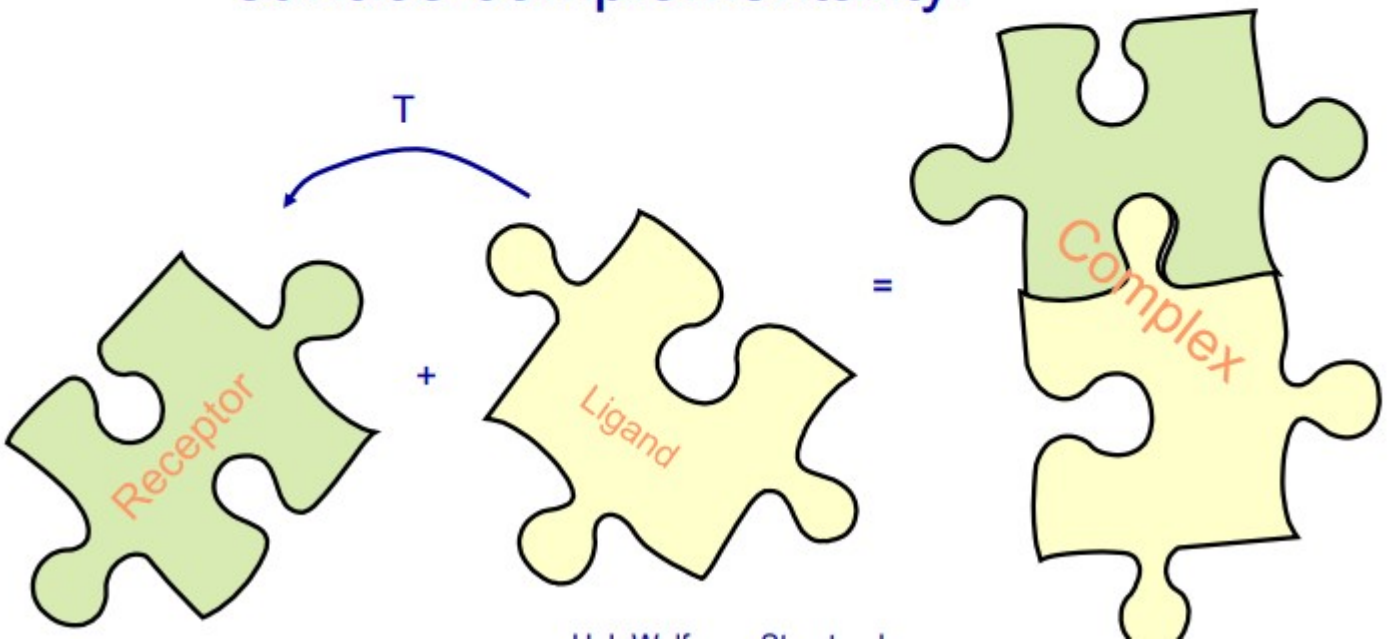
Prediccion de la estructura de Proteinas



Prediccion de la estructura de Proteinas

Necessary Condition for Docking

- Given two molecules find significant surface complementarity.



Prediccion de la estructura de Proteinas

Geometric Docking Algorithms

- Based on the assumption of shape complementarity between the participating molecules.
- Molecular surface complementarity – protein-protein, protein-drug.

Remark : usually “protein” here can be replaced by “DNA” or “RNA” as well.

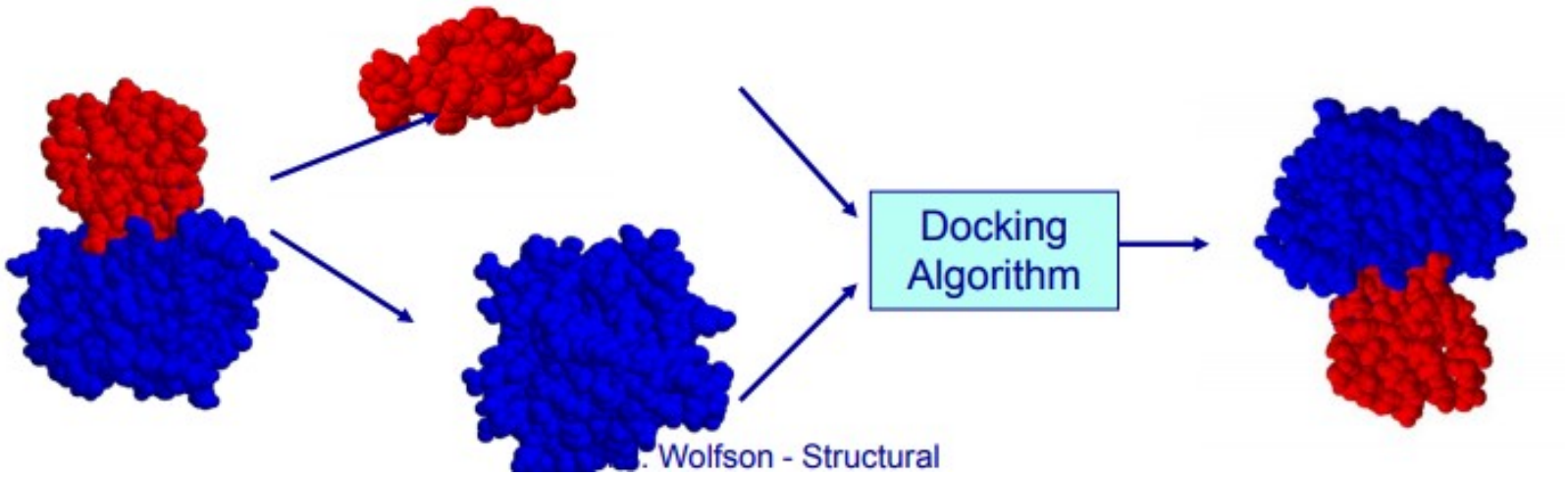
Issues to be examined when evaluating docking methods

- **Rigid** docking vs. **Flexible** docking :
  - If the method allows flexibility:
    - Is flexibility allowed for ligand only, receptor only or both ?
    - Number of flexible bonds allowed and the cost of adding additional flexibility.
- Does the method require prior knowledge of the **active site**?
- **Speed** - ability to explore large libraries.
- Performance in "**unbound**" docking experiments.

Prediccion de la estructura de Proteinas

Bound Docking

- In the bound docking we are given a complex of 2 molecules.
- After artificial separation the goal is to reconstruct the native complex.
- No conformational changes are involved.
- Used as a first test of the validity of an algorithm.

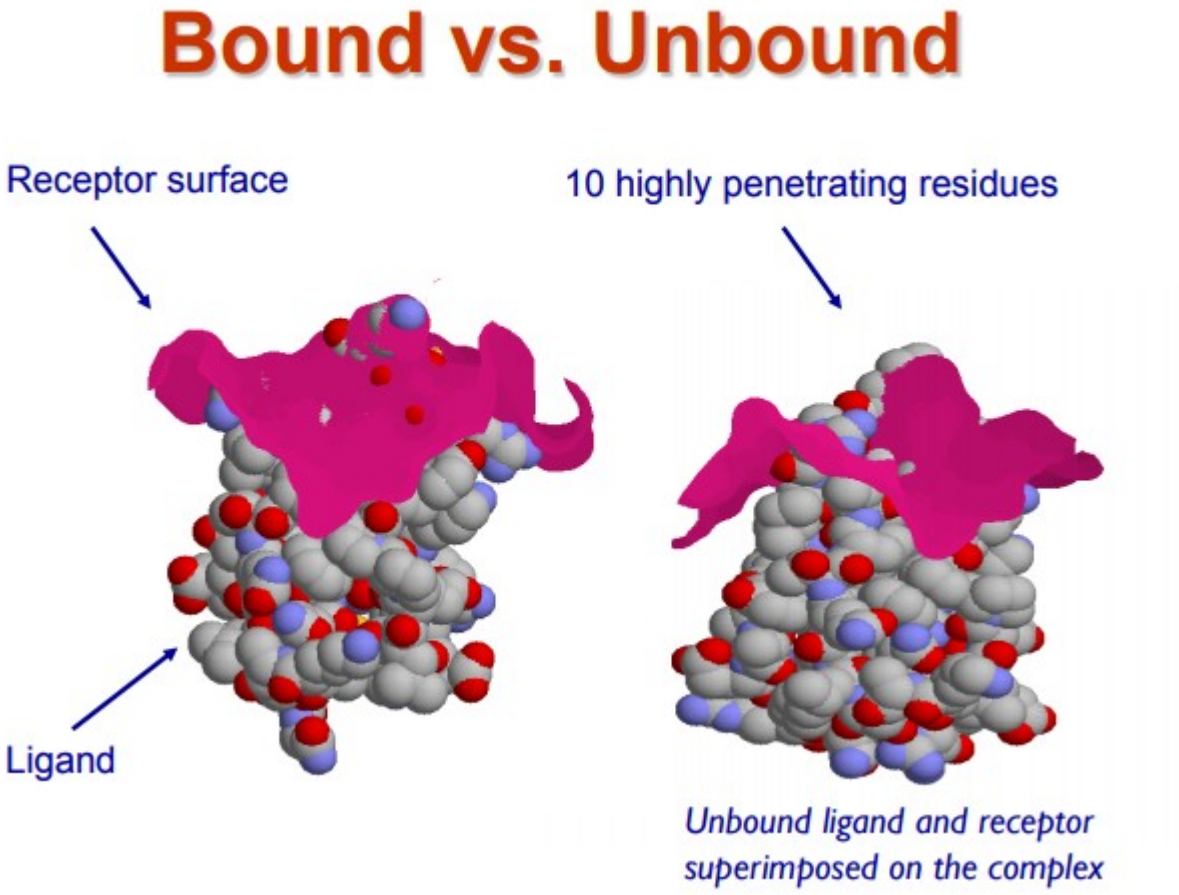


Unbound Docking

- In the unbound docking we are given 2 molecules in their native conformation.
- The goal is to find the correct association.

- **Problems:** conformational changes (side-chain and backbone movements), experimental errors in the structures.

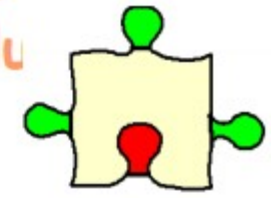
Prediccion de la estructura de Proteinas



Prediccion de la estructura de Proteinas

The PatchDock Algorithm

- Based on local shape feature matching.
- Focuses on local surface patches divided into three shape types: concave, convex and flat.
- The geometric surface complementarity scoring employs advanced data structures for molecular representation: Distance Transform Grid and Multi-Resolu Surface.



Prediccion de la estructura de Proteinas

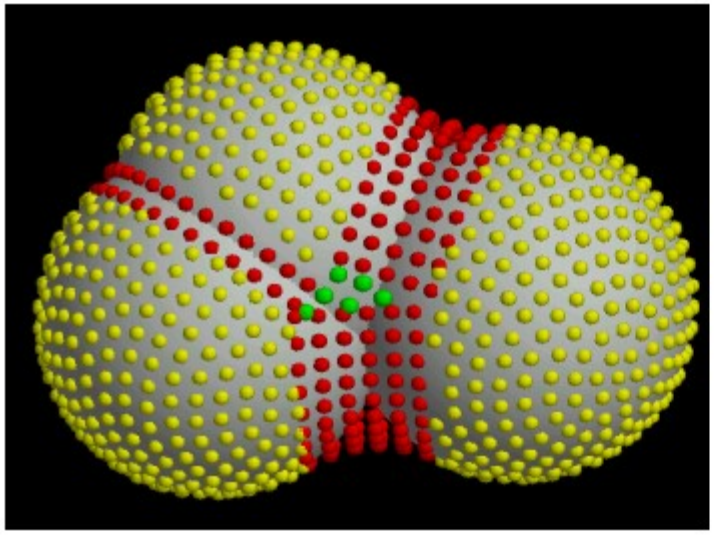
Docking Algorithm Scheme

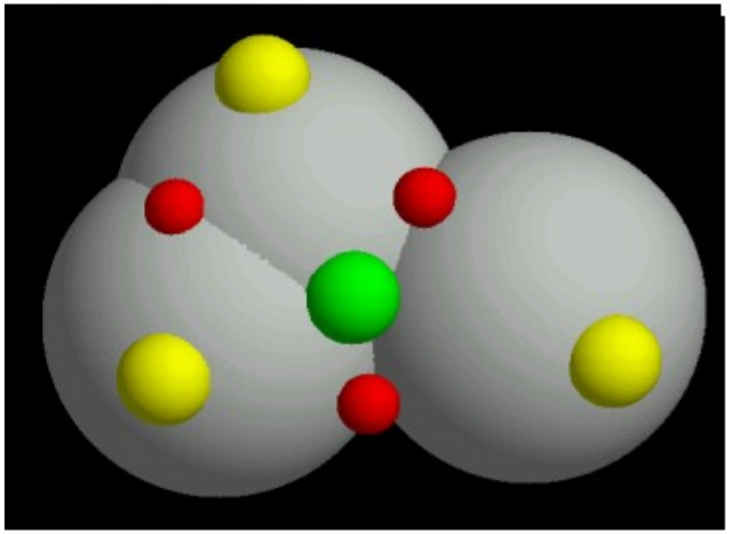
- Part 1: **Molecular surface representation**
- Part 2: Feature selection
- Part 3: Matching of critical features
- Part 4: Filtering and scoring of candidate transformations

Prediccion de la estructura de Proteinas

1. Surface Representation

- Dense MS surface (Connolly)
- Sparse surface (Lin et al.)

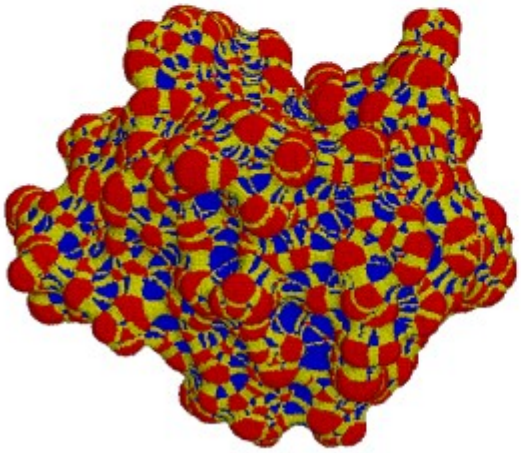




Prediccion de la estructura de Proteinas

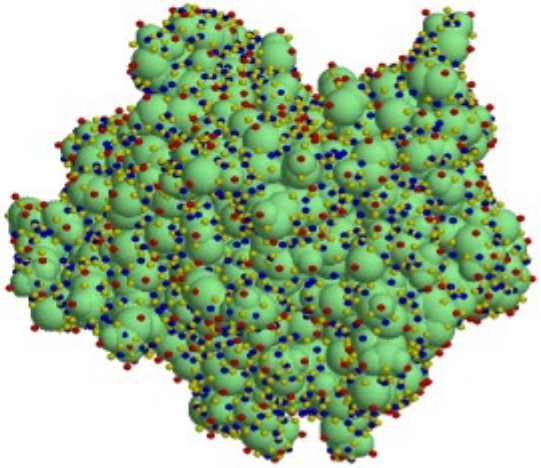
1. Surface Representation

- Dense MS surface (Connolly)



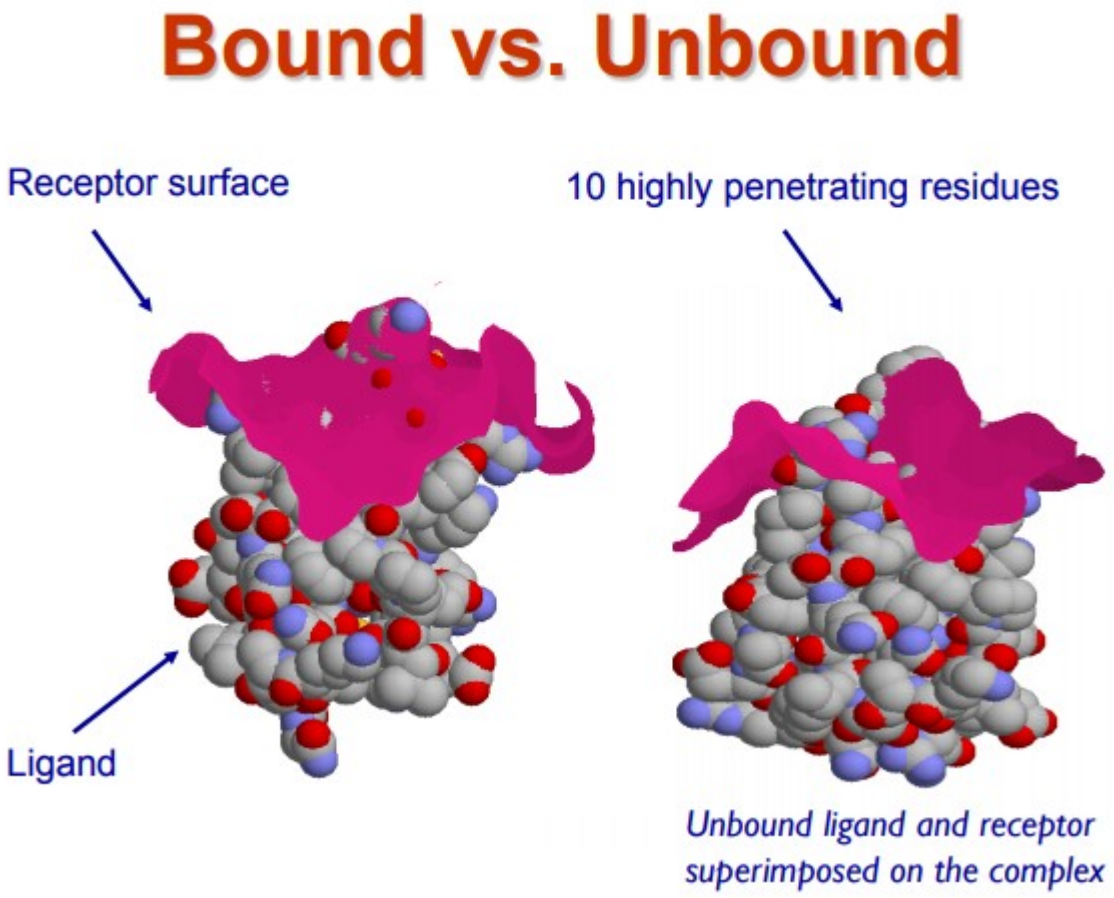
82,500 points

- Sparse surface (Lin et al.)



4,100 points

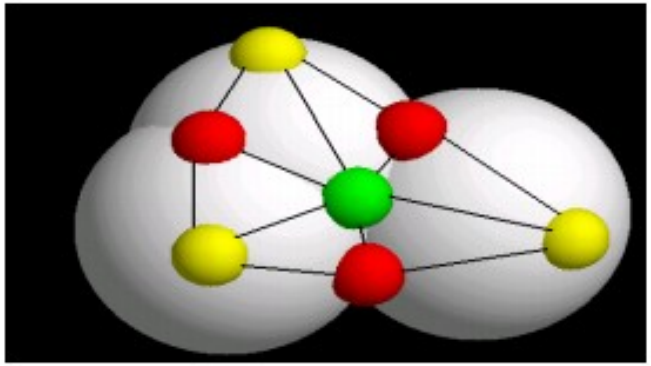
Prediccion de la estructura de Proteinas



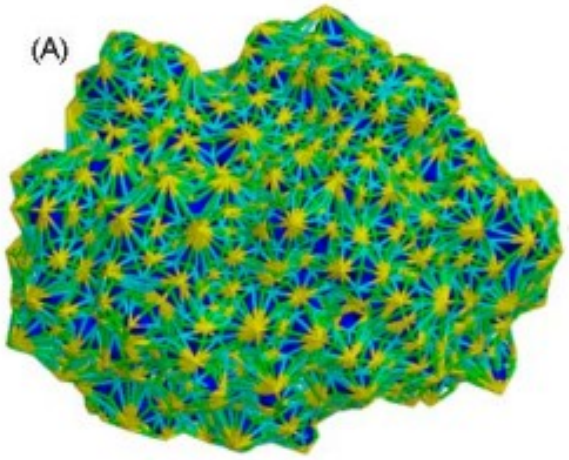
Prediccion de la estructura de Proteinas

Sparse Surface Graph - G_{top}

- Caps (yellow), pits (green), belts (red):



- G_{top} – Surface topology graph:



V = *surface points*

E = *{(u,v)| u,v belong to the same atom}*

H.J. Wolfson - Structural Bioinformatics

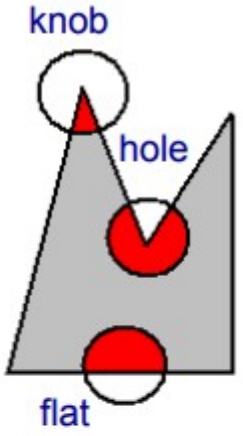
Prediccion de la estructura de Proteinas

Docking Algorithm Scheme

- Part 1: Molecular surface representation
- Part 2: **Feature selection**
  - 2.1 Coarse curvature calculation
  - 2.2 Division to surface patches of similar curvature
- Part 3: Matching of critical features
- Part 4: Filtering and scoring of candidate transformations

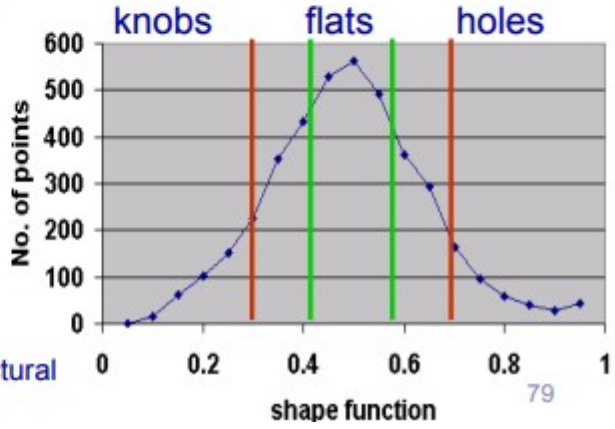
Prediccion de la estructura de Proteinas

2.1 Curvature Calculation



- Shape function is a measure of local curvature.
- 'knobs' and 'holes' are local minima and maxima (<1/3 or >2/3), 'flats' – the rest of the points.

- Problems: sensitivity to molecular movements, 3 sets of points with different sizes.
- Solution: divide the values of the shape function to 3 equal sized sets: 'knobs', 'flats' and 'holes'.



H.J. Wolfson - Structural Bioinformatics

2.2 Patch Detection

Goal: Divide the surface into connected, non-intersecting, equal sized patches of critical points with similar curvature.

- **connected** – the points of the patch correspond to a connected sub-graph of G_{top}.
- **similar curvature** – all the points of the patch correspond to only one type: knobs, flats or holes.
- **equal sized** – to assure better matching we want shape features of almost the same size.

Patch Detection by Segmentation Technique

- Construct a sub-graph for each type of points: knobs, holes, flats.
  Example: G_{knob} will include all surface points that are knobs and an edge between two 'knobs' if they belong to the same atom.
- Compute connected components of every sub-graph.
- Problem: the sizes of the connected components can vary.
- Solution: apply 'split' and 'merge' routines.

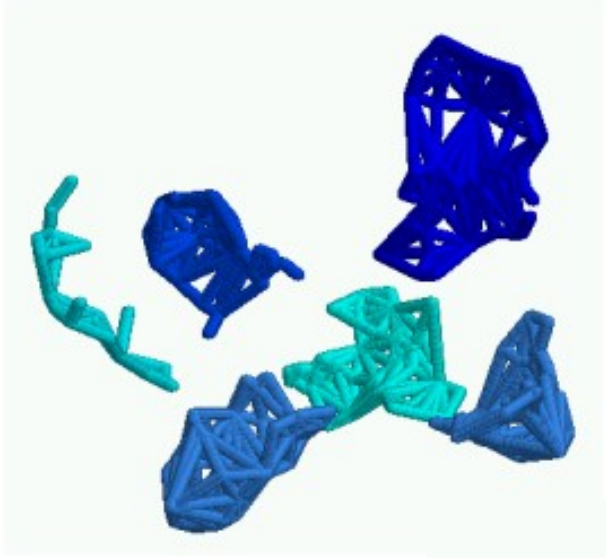
Split and Merge

- **Geodesic distance** between two nodes is a weight of the shortest path between them in surface topology graph. The weight of each edge is equal to the Euclidean distance between the corresponding surface points.

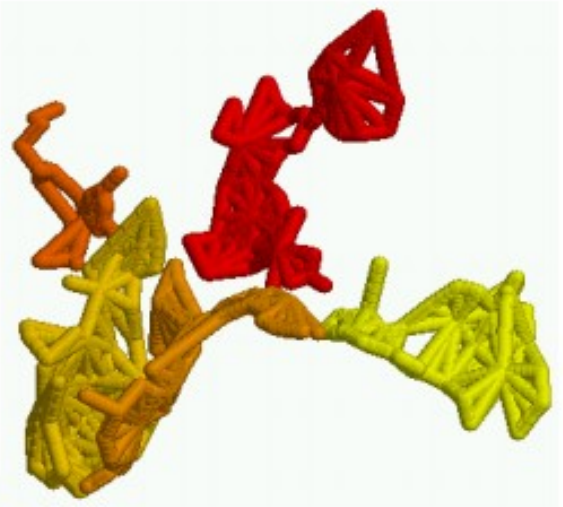
- **Diameter of the component** – is the largest geodesic distance between the nodes of the component. Nodes *s* and *t* that give the diameter are called *diameter nodes*.

Prediccion de la estructura de Proteinas

Complementarity of the Patches:



Interface knob patches of the ligand



Interface hole patches of the receptor

H.J. Wolfson - Structural Bioinformatics

Split and Merge (cont.)

- The diameter of every connected component is computed using the APSP (All pairs shortest paths) algorithm ($O(n^3)$).

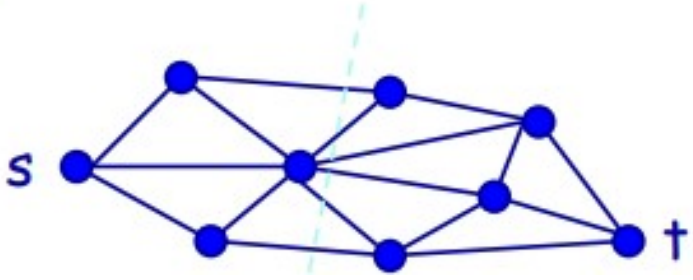
1. *low_patch_thr* ≤ *diam* ≤ *high_patch_thr* → **valid patch**
2. *diam* > *high_patch_thr* → **split**
3. *diam* < *low_patch_thr* → **merge**

- *low_patch_thr = 10Å*
- *high_patch_thr = 20Å*

Split and Merge (cont.)

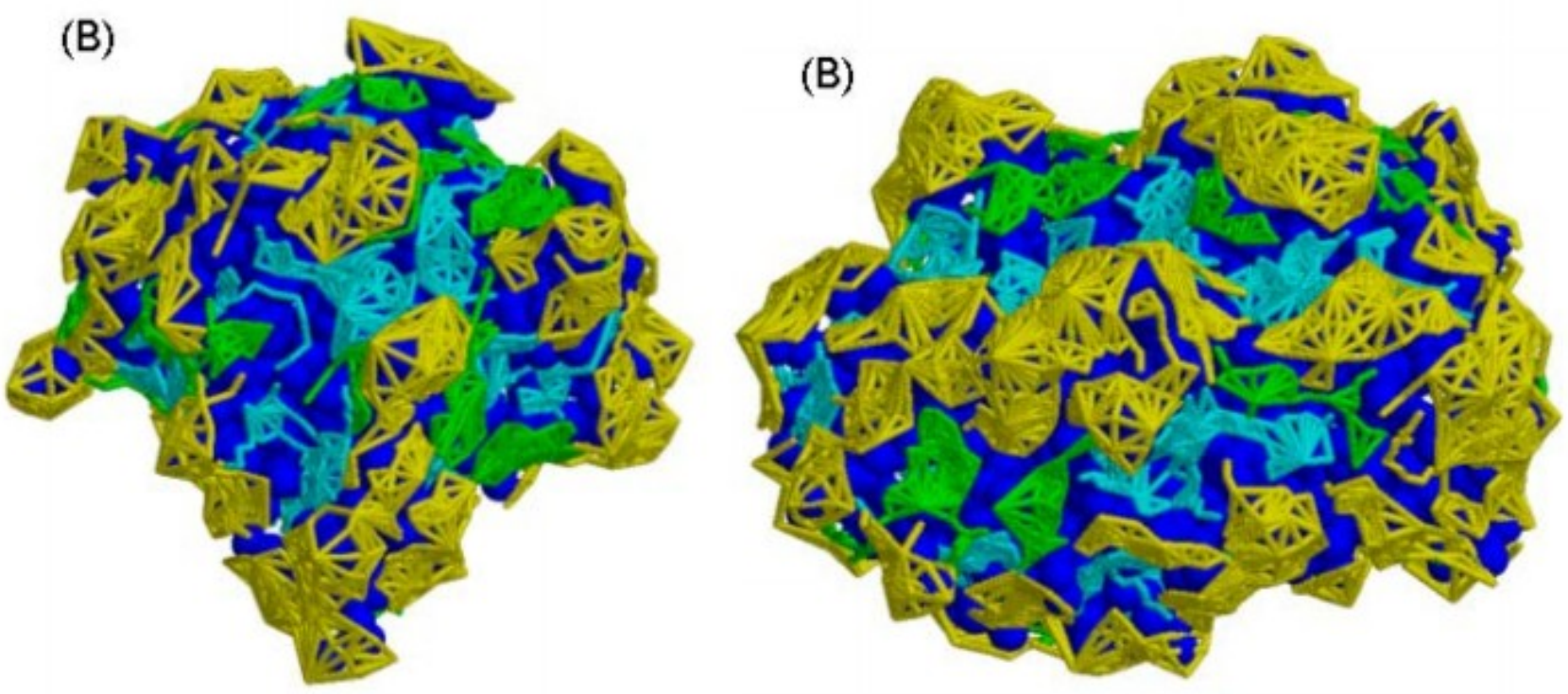
- Split routine: compute Voronoi cells of the diameter nodes *s,t*. Points **closer to *s*** belong to new component *S*, points **closer to *t*** belong to new component *T*. The split is applied until the new component has a valid diameter.
- Merge routine: compute the geodesic distance of every component point to all the patches. Merge with the patch with closest distance.

- *Note*: the merge routine may merge point with patch of different curvature type.



Prediccion de la estructura de Proteinas

Examples of Patches for trypsin and trypsin inhibitor



Yellow – knob patches, cyan – hole patches, green – flat patches, the proteins are in blue.

Prediccion de la estructura de Proteinas

Focusing on Active Site

There are major differences in the interactions of different types of molecules (enzyme-inhibitor, antibody-antigen, protein drug). Studies have shown the presence of energetic *hot spots* in the active sites of the molecules.

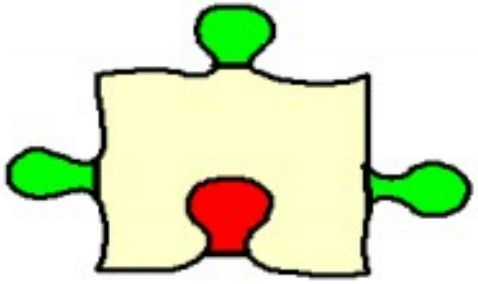
Enzyme/inhibitor –
Select patches with high enrichment of hot spot residues
(Ser, Gly, Asp and His for the enzyme;
Arg, Lys, Leu, Cys and Pro for the inhibitor).

Antibody/antigen –
1. Detect CDRs of the antibody.
2. Select hot spot patches
(Tyr, Asp, Asn, Glu, Ser and Trp for antibody;
and Arg, Lys, Asn and Asp for antigen)

Protein/drug – Select large protein cavities

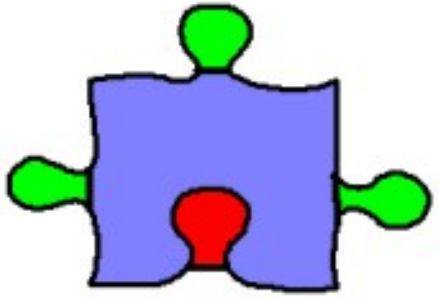
Docking Algorithm Scheme

- Part 1: Molecular surface representation



- Part 2: Feature selection

- Part 3: Matching of critical features



- Part 4: Filtering and scoring of candidate transformations

3. Matching of patches

The aim is to align knob patches with hole patches, and flat patches with any patch. We use two types of matching:

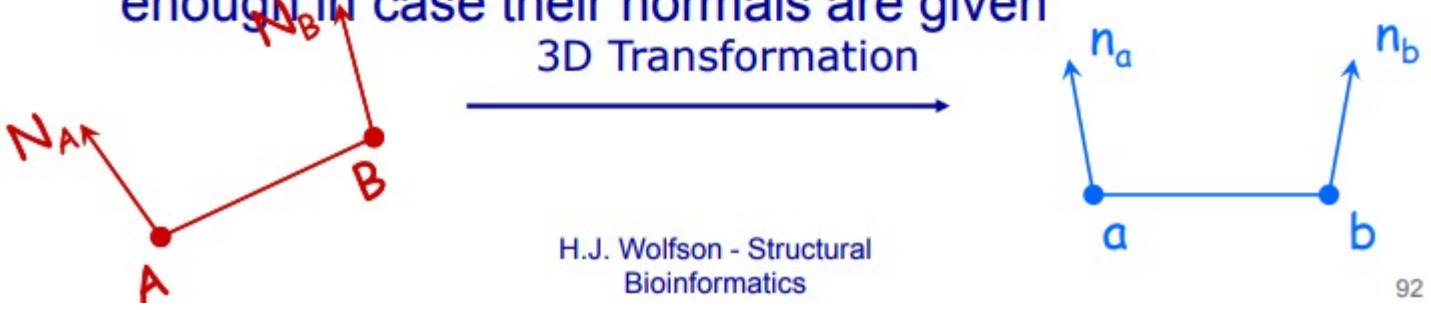
- Single Patch Matching – one patch from the receptor is matched with one patch from the ligand. Used in protein-drug cases.
- Patch-Pair Matching – two patches from the receptor are matched with two patches from the ligand. Used in protein-protein cases.

Prediccion de la estructura de Proteinas

Creating Transformations in 3D Space

- A correspondence between a pair of 3 points is necessary to compute a 3D transformation

- A correspondence between a pair of 2 points is enough in case their normals are given

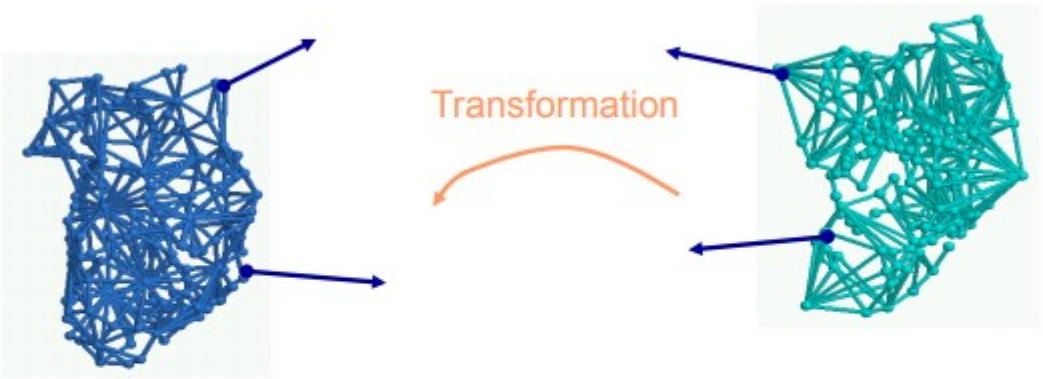


Prediccion de la estructura de Proteinas

Single Patch Matching

Receptor hole patch

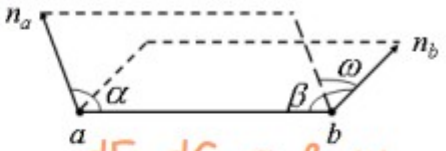
Ligand knob patch



- Base: a pair of critical points with their normals from one patch.
- Match every base from a receptor patch with all the bases from complementary ligand patches.
- Compute the transformation for each pair of matched bases.

Prediccion de la estructura de Proteinas

Base Compatibility



dE, dG, α, β, ω

The signature of the base is defined as follows:

1. Euclidean and geodesic distances between the points: dE, dG
2. The angles α, β between the [a,b] segment and the normals
3. The torsion angle ω between the planes

Two bases are compatible if their signatures match.

Prediccion de la estructura de Proteinas

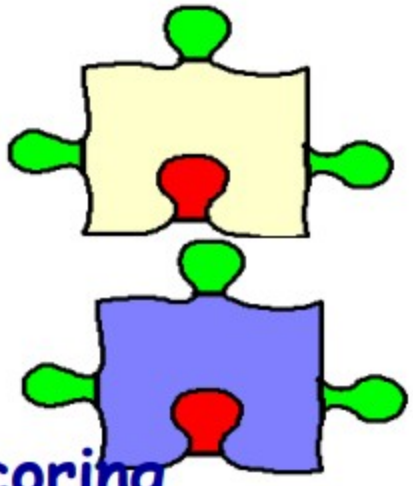
Patch Matching

- **Preprocessing:** the bases are built for all ligand patches (single or pairs) and stored in hash table according to base signature.
- **Recognition:** for each receptor base access the hash-table with base signature. The transformations set is computed for all compatible bases.

Prediccion de la estructura de Proteinas

Docking Algorithm Scheme

- Part 1: Molecular surface representation
- Part 2: Feature selection
- Part 3: Matching of critical features
- Part 4: **Filtering and scoring of candidate transformations**



Filtering Transformations with Steric Clashes

- Since the transformations were computed by local shape features matching they may include unacceptable steric clashes.
- Candidate complexes with slight penetrations are retained due to molecular flexibility.

Steric clash test:

```
For each candidate ligand transformation
   transform ligand surface points
   For each transformed point
      access Distance Transform Grid and check distance value
    If it is more than max_penetration
     Disqualify transformation
```

Prediccion de la estructura de Proteinas

Scoring Shape Complementarity

- The scoring is necessary to rank the remaining solutions.
- The surface of the receptor is divided into five shells according to the distance function: S1-S5

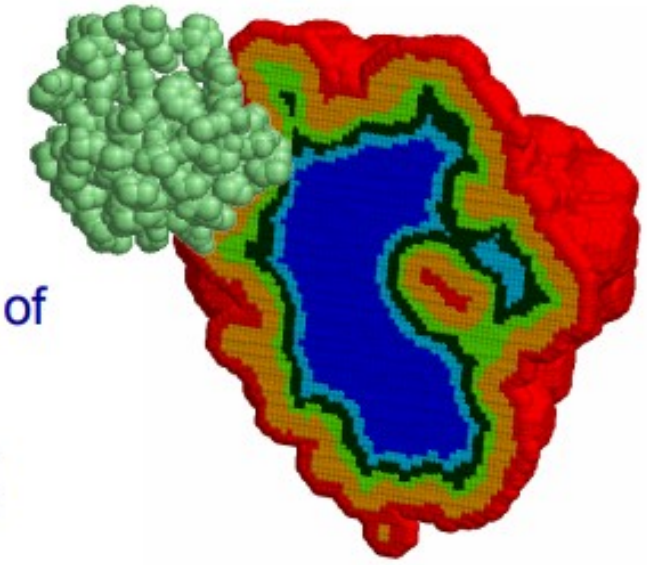
[-5.0,-3.6), [-3.6,-2.2), [-2.2, -1.0), [-1.0,1.0), [1.0→).

- The number of ligand surface points in every shell is counted.
- Each shell is given a weight: W1-W5

-10, -6, -2, 1, 0.

- The geometric score is a weighted sum of the number of ligand surface points N inside every shell:

$$score = \sum_i N_{S_i} W_i$$



Prediccion de la estructura de Proteinas

Docking Algorithm Scheme

- Part 1: Molecular surface Representation
- Part 2: Features selection
- Part 3: Matching of critical features

The correct solution is found in 90% of the cases with RMSD under 5A.

The rank of the correct solution can be in the range of 1 – 1000.

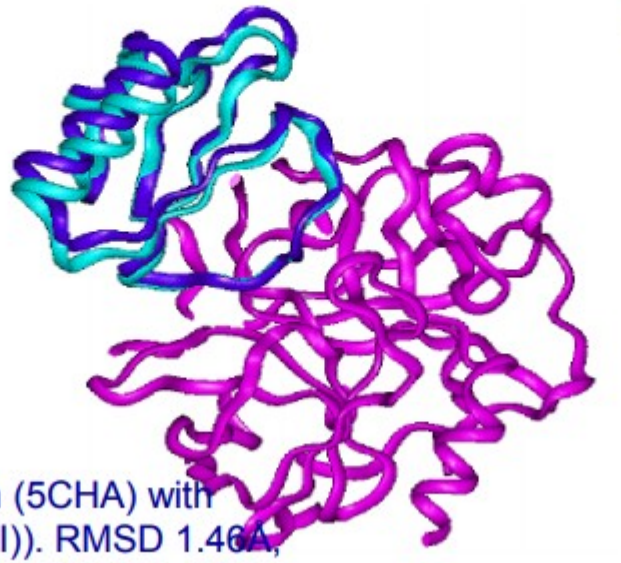
- Part 4: Filtering and scoring of candidate transformations

→ Refinement and Rescoring minimizing an Energy Function !

Prediccion de la estructura de Proteinas

Example 1: Enzyme-inhibitor docking (unbound case)

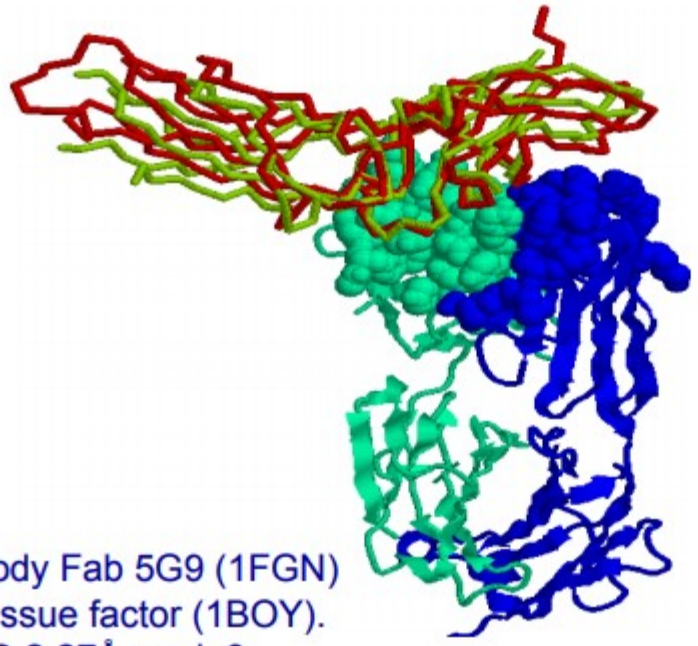
- trypsin
- inhibitor from complex
- docking solution



A-chymotrypsin (5CHA) with Eglin C (1CSE(I)). RMSD 1.46A, rank 10

Prediccion de la estructura de Proteinas

Example 2: Antibody-antigen docking (unbound case)



antibody

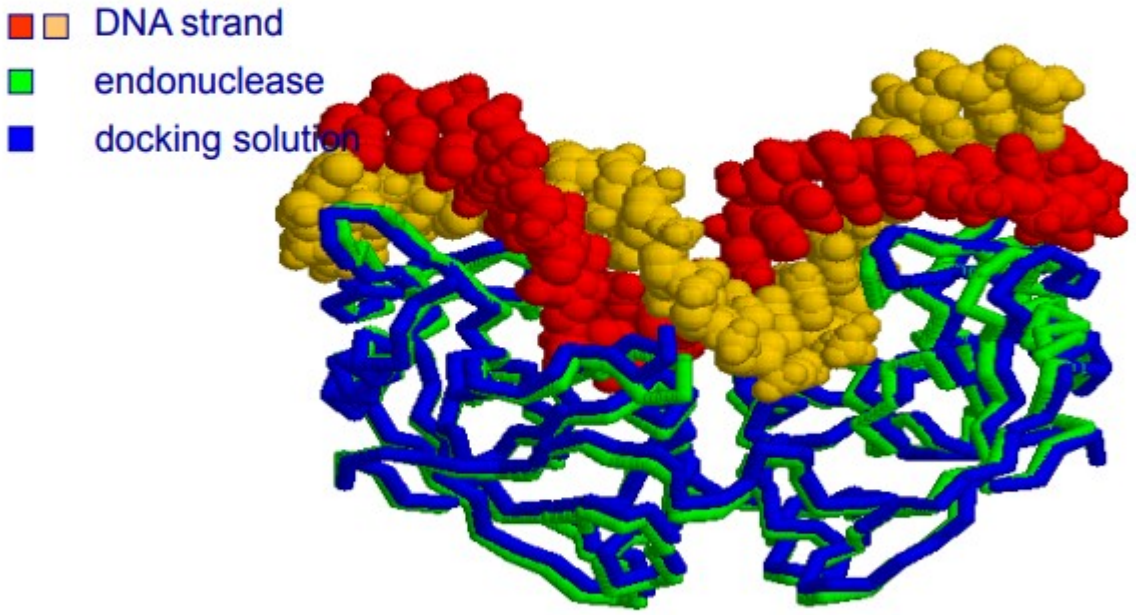
tissue factor from complex

docking solution

Antibody Fab 5G9 (1FGN) with tissue factor (1BOY). RMSD 2.27Å, rank 8

Prediccion de la estructura de Proteinas

Example 3: Protein-DNA docking (semi-unbound case)



Endonuclease I-PpoI (1EVX)
with DNA (1A73).
RMSD 0.87Å, rank 2

Prediccion de la estructura de Proteinas

References (PatchDock):

- **D. Duhovny, R. Nussinov, H.J. Wolfson**, *Efficient Unbound Docking of Rigid Molecules,* 2'nd Workshop on Algorithms in Bioinformatics (WABI'02 as part of ALGO'02), 2002, Lecture Notes in Computer Science 2452, pp. 185-200, Springer Verlag.
- **D. Schneidman-Duhovny, Y. Inbar, R. Nussinov and H. J. Wolfson,** *PatchDock and SymmDock: servers for rigid and symmetric docking,* Nuc. Acids Res., 33, W363—W367, (2005).

- **SERVER URL** : http://bioinfo3d.cs.tau.ac.il/PatchDock/

Acceda a:
PatchDock