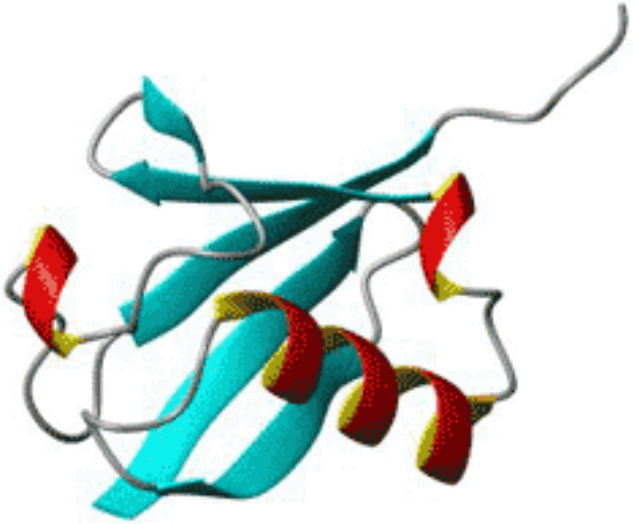


Structural Bioinformatics

Haim Wolfson

Lecture overview

- **Introduction and Motivation.**
- **Protein Folding – the RAPTOR threading algorithm.**
- **Modeling of protein-protein interactions – the PatchDock docking algorithm.**

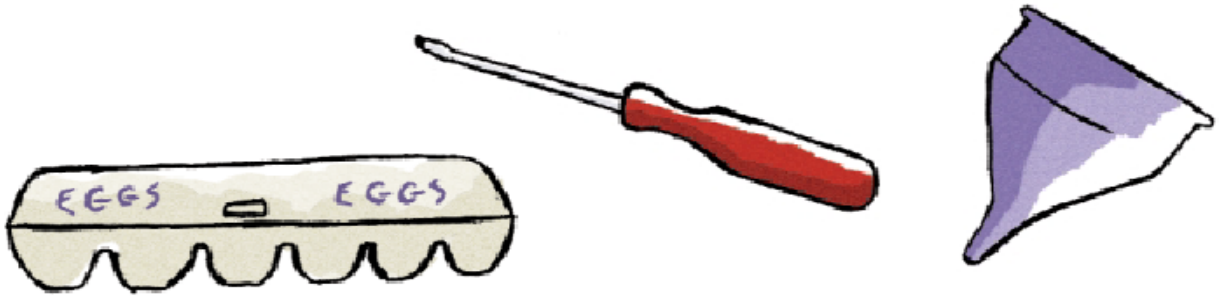
Why 3D Structures?

1. 3D Structure (shape) is better preserved than sequence (text).
2. Structural motifs may predict similar biological function.
3. Drug Design.

- Example, identification of a p
  description via a picture.

  Mid-aged man, black hear
  eyes and moustache…

Shape to function



Macromolecules, like many everyday objects,
have been shaped (by evolution) to get their job done.

Elucidation of macromolecular shape can supply insight on the function of the molecules involved.

"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

Structural Bioinformatics aka Computational Structural Biology

- Deals with Structural data of molecules.

- Exploits (and develops) algorithms for interpretation and handling of **3D** (spatial data) – Geometric Computing.

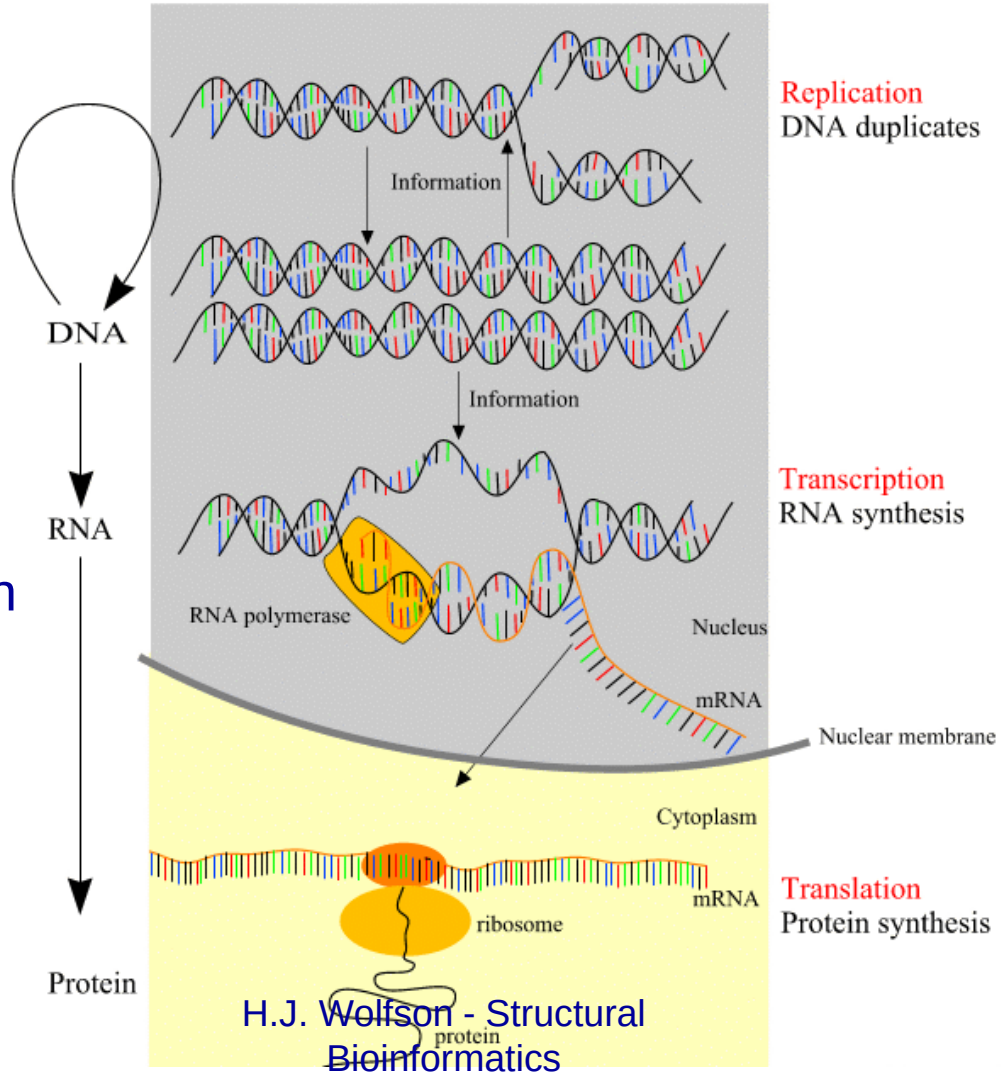
- Sister computational disciplines – Computational Geometry, Computer Vision, Computer Graphics, Medical Image Interpretation, Pattern Recognition.

Recommended Web Sites:

- **Proteopedia** http://proteopedia.org/
- **Protein Data Bank (PDB)** http://www.rcsb.org/pdb/

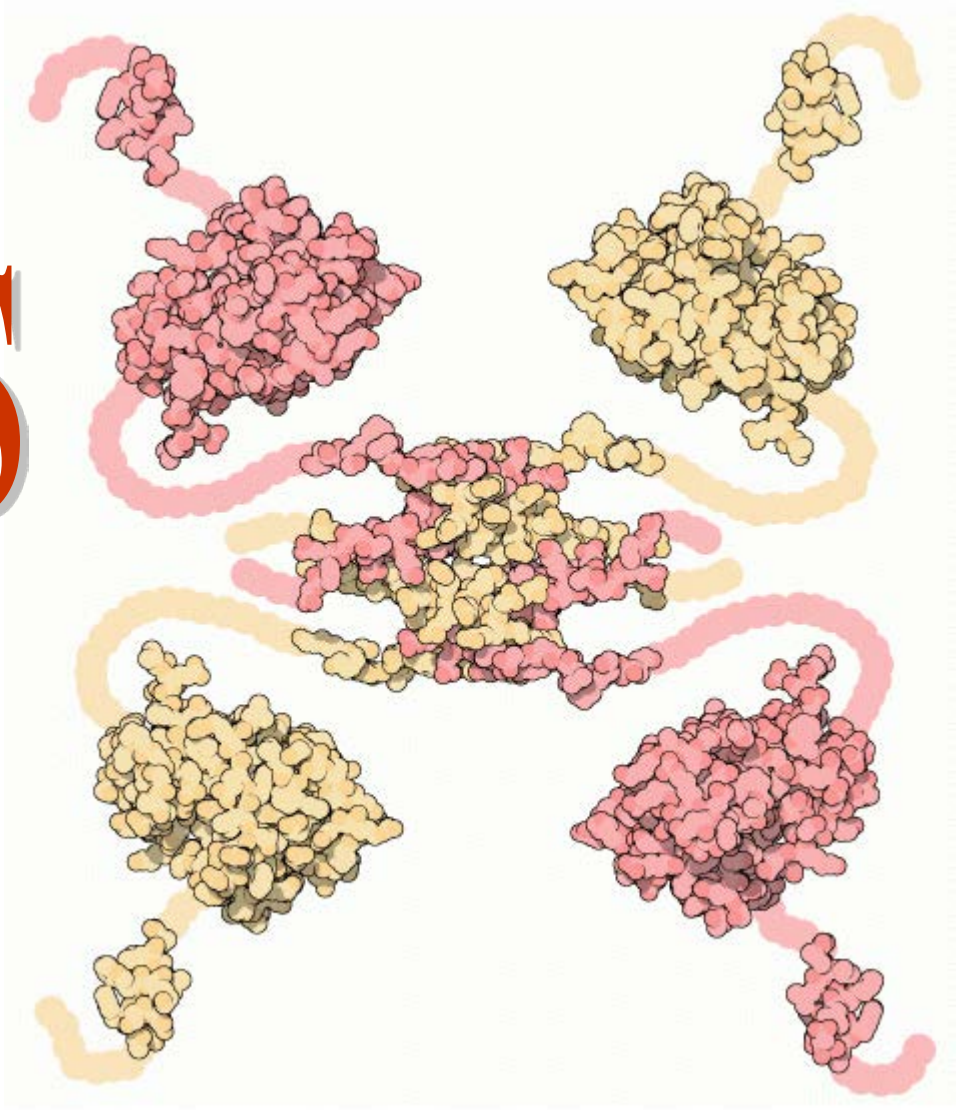
The Central Dogma

RNA is an information carrier from DNA to Protein



H.J. Wolfson - Structural Bioinformatics

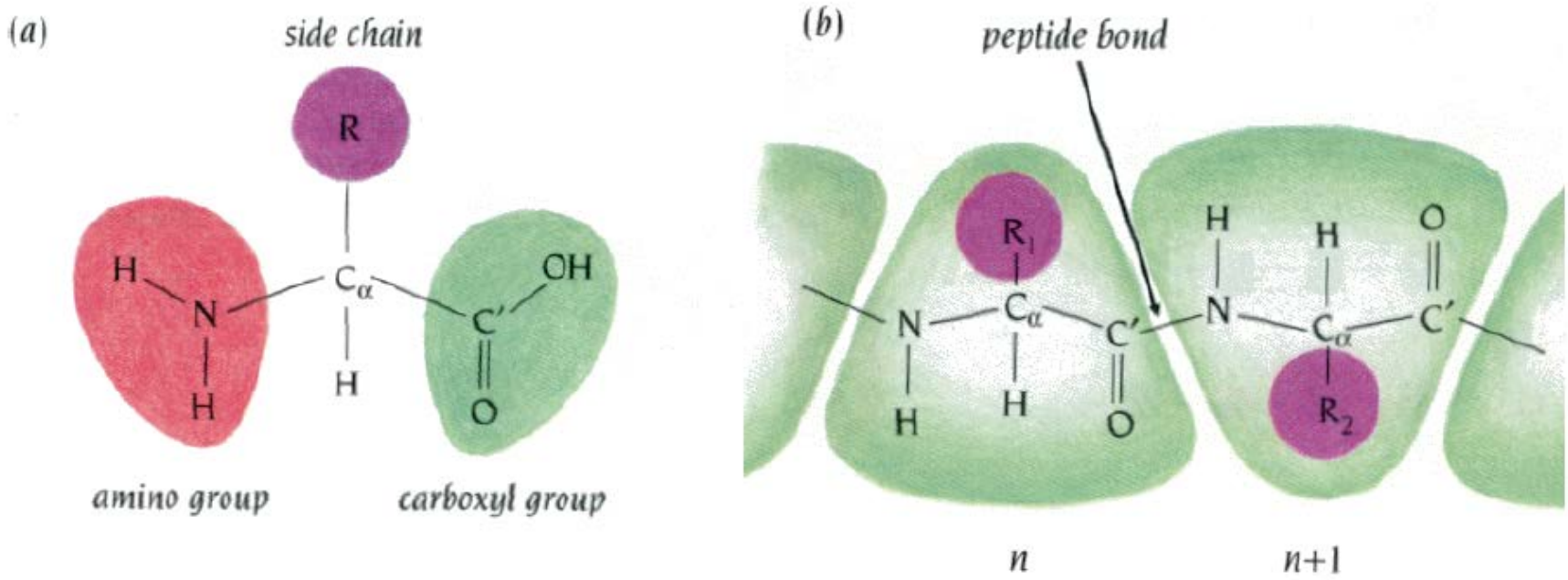
Proteins

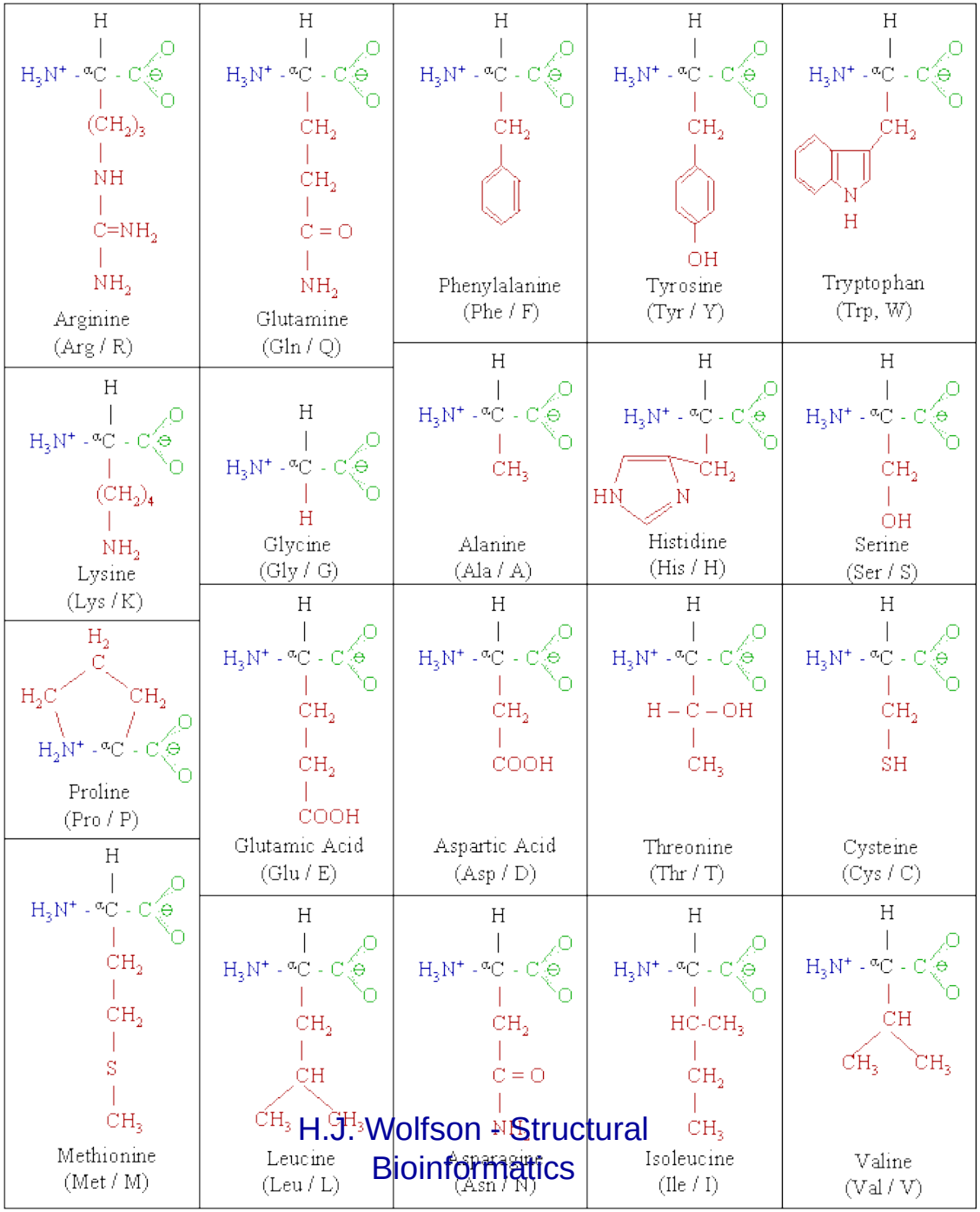


The Biological Role
(Robots of the Cell)

1. Catalysis (enzymes).
2. Signal propagation:
   - transmit nerve impulses
   - control cell growth and differentiation.
3. Transport (of electrons or macromolecules).
4. Immune system (e.g. antibodies which bind to specific foreign particles such as bacteria and viruses).
5. Structural proteins (hair, skin, nails).

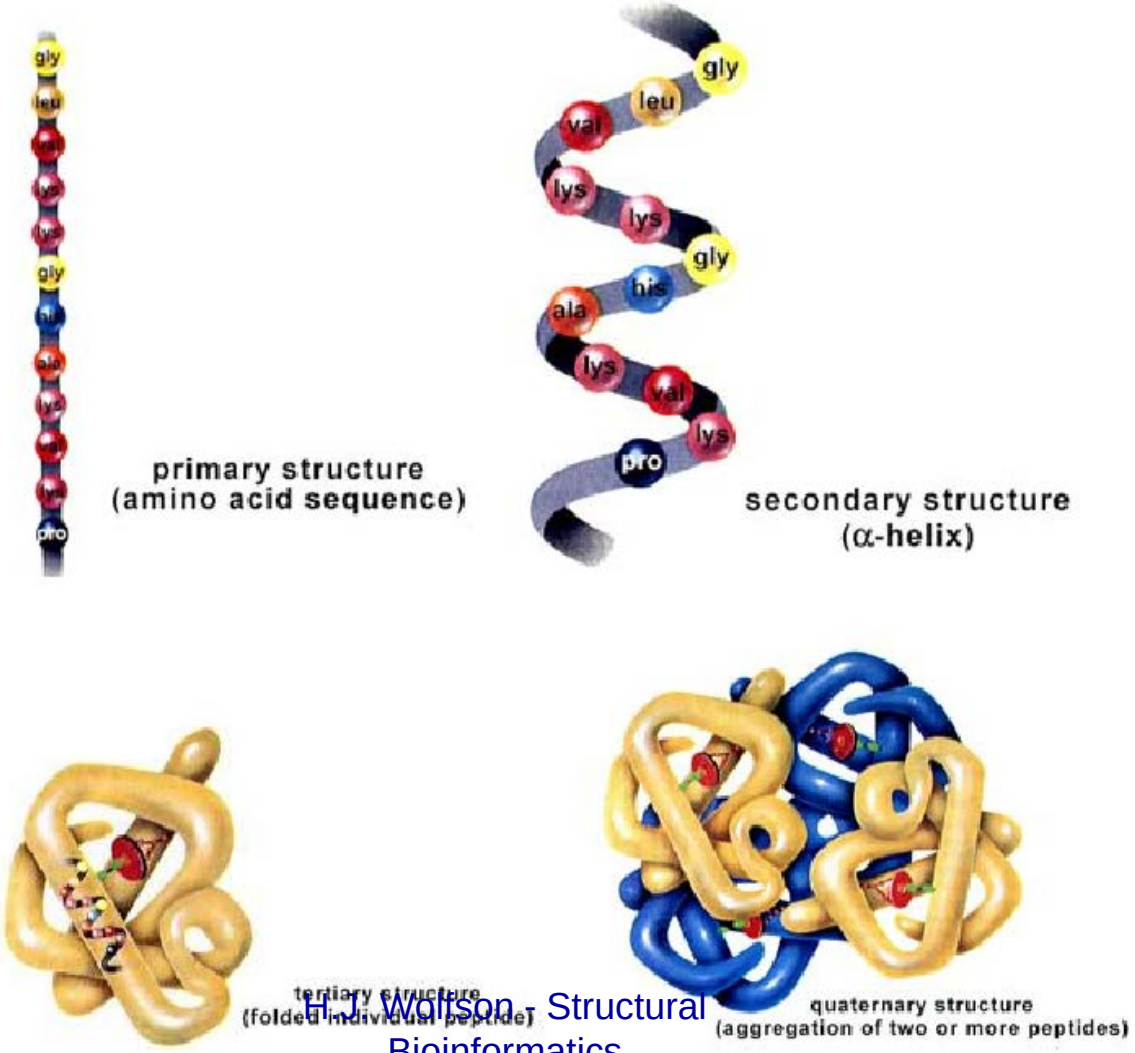
Amino Acids and the Peptide Bond





H.J. Wolfson - Structural Bioinformatics

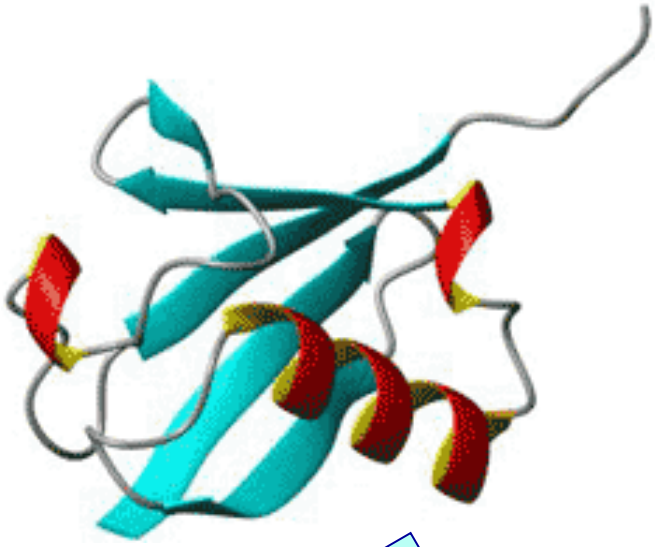
Protein Structure



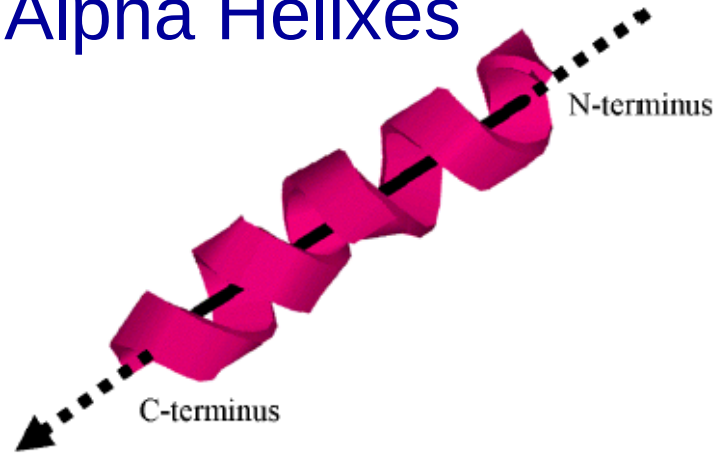
Primary Structure

- Primary structure: The order of the amino acids composing the protein.
- AASGDXSLVEVHXXVFIVPPXIL…..

Secondary Structure

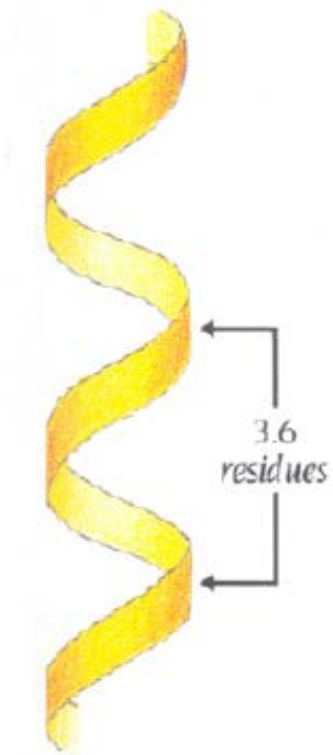


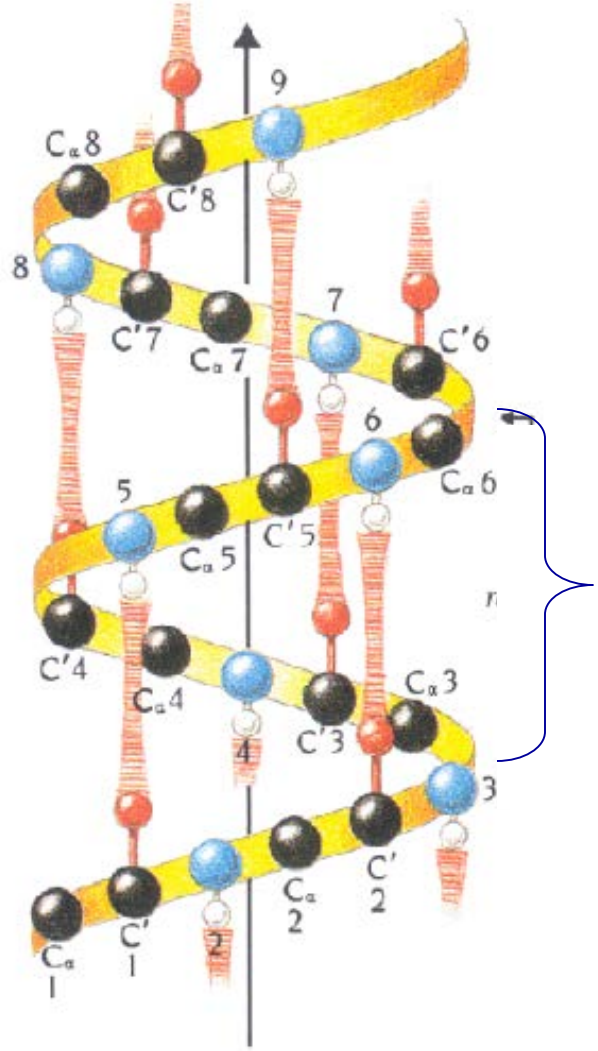
Alpha Helixes

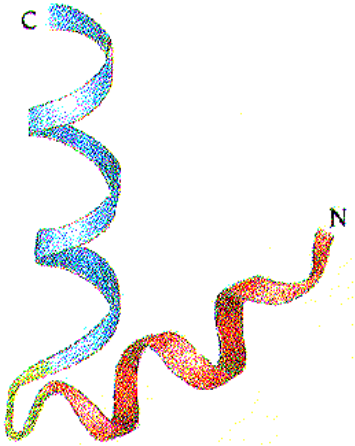


Beta Strand

Alpha-Helix



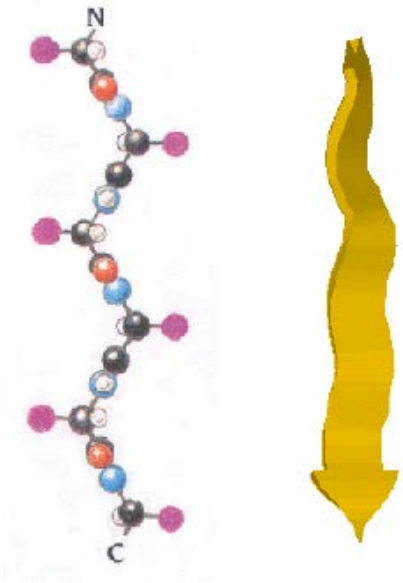


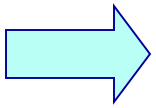


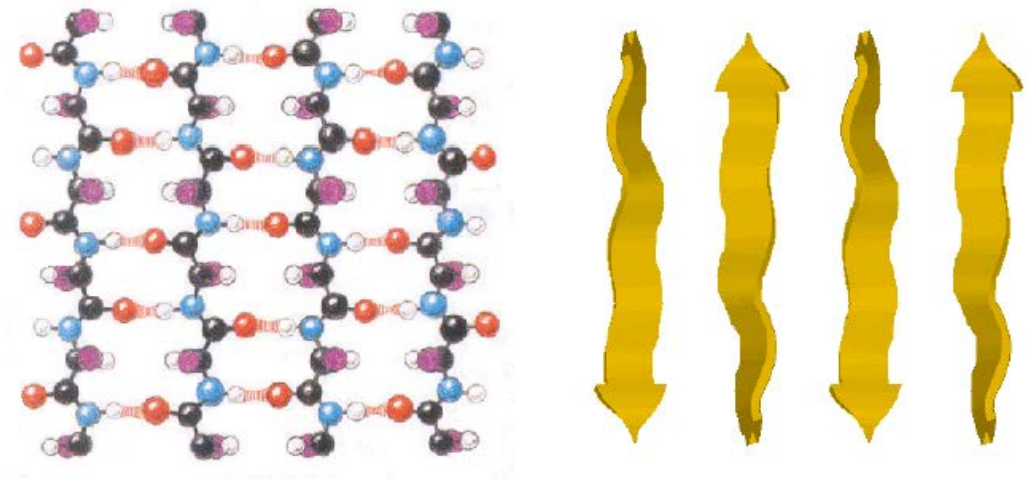
Length

Main-chain atoms N and O are colored red and blue respectively.
The hydrogen bonds between them are red and striated.

Beta Strands and Beta Sheet





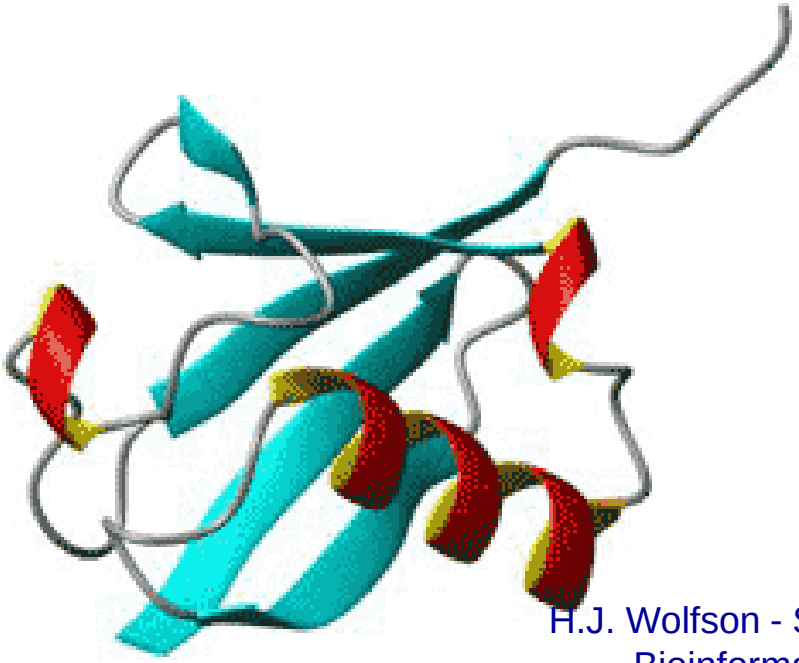


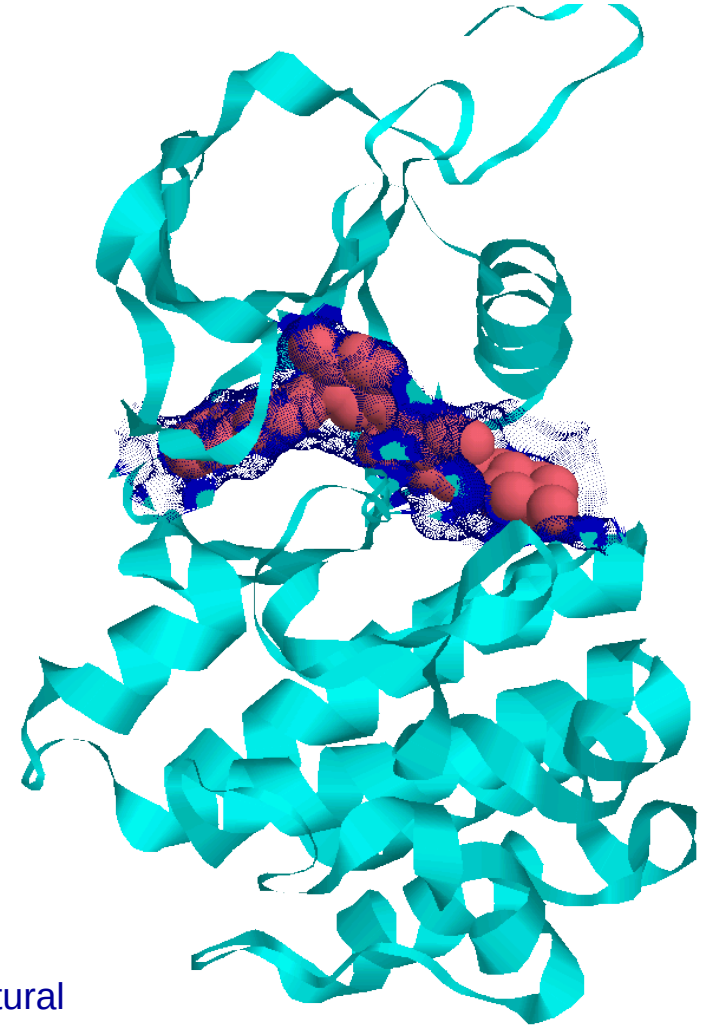
Beta strand. Typical Length 5-10 residues.

Beta sheets. Backbone NH and O atoms hydrogen bonded to each other. O, N, H and C atoms are colored red, blue, white and black respectively. Side chains are shown as purple circles.

Tertiary structure

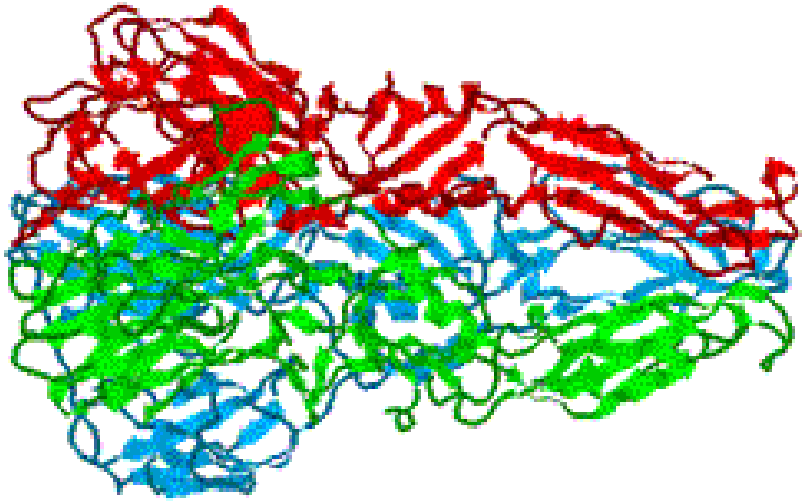
- Full 3D folded structure of the polypeptide chain.

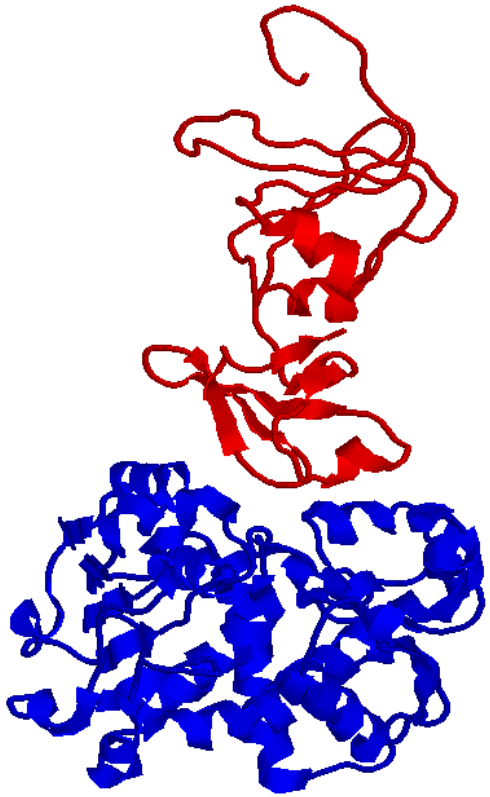




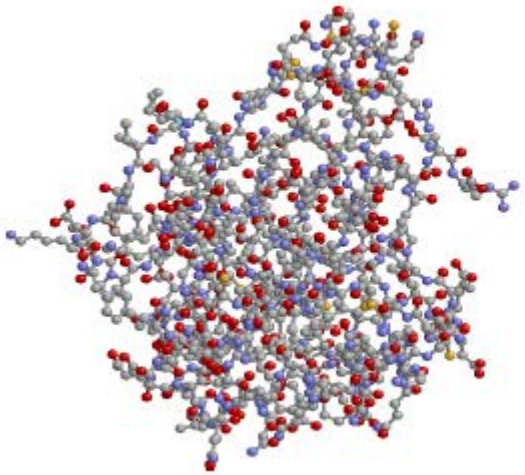
Quaternary structure

- The interconnections and organization of more than one polypeptide chain.

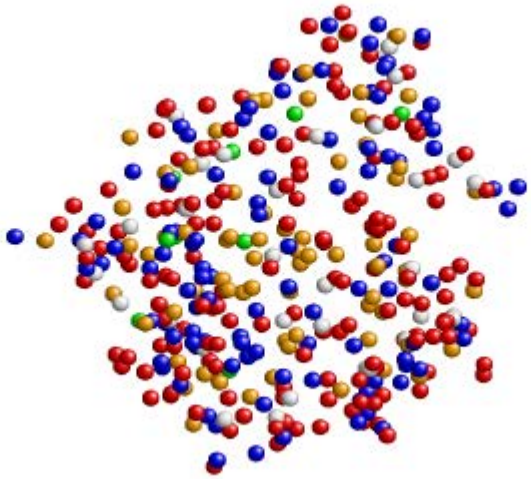




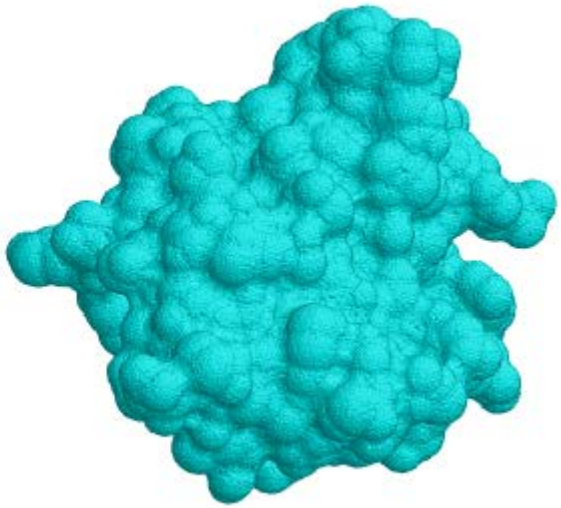
Different Representations



Amino acids



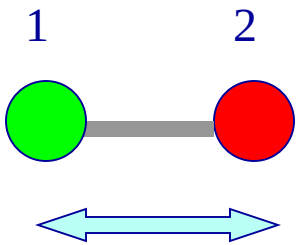
Functional groups



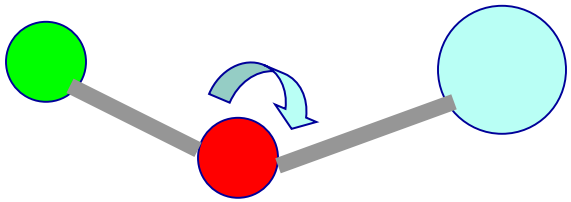
Surface

Degrees of Freedom in Proteins

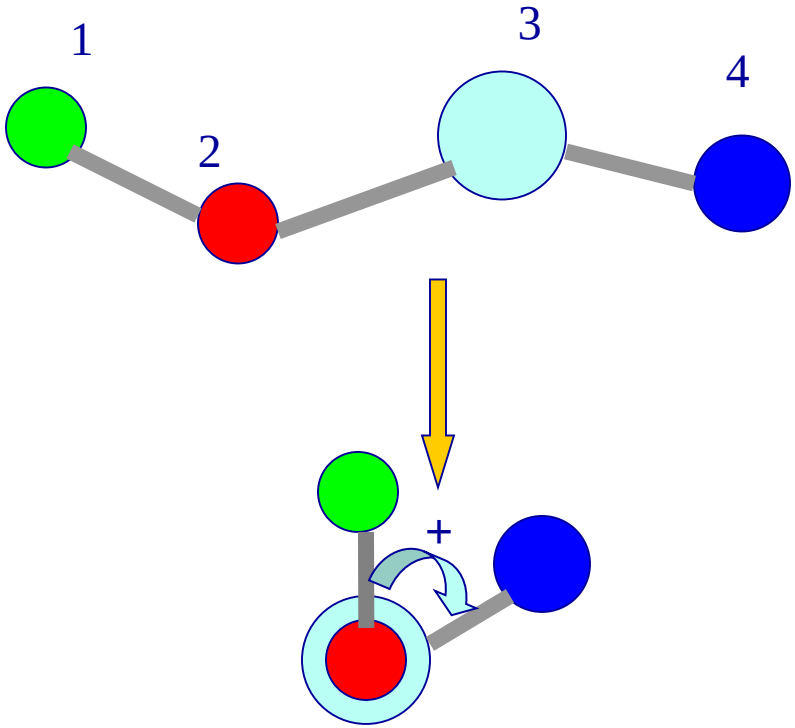
Bond length



Bond angle

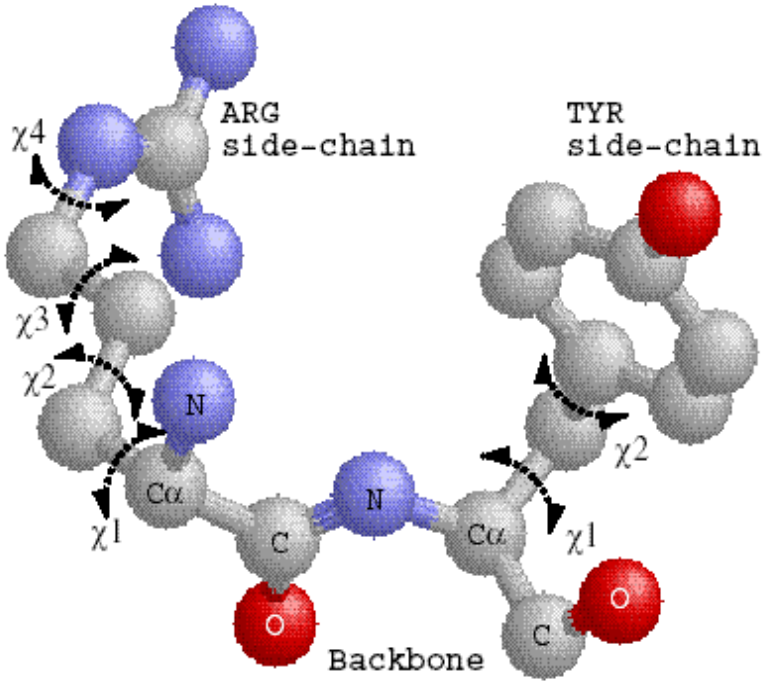


Dihedral angle



Patrice Koehl, koehllab.genomecenter.ucdavis.edu/teaching/ecs229/01/files/ECS229_Lecture2.ppt

Backbone and Side-Chains



Determination of Protein Structure

X-ray crystallography

NMR (Nuclear Magnetic Resonance)

EM (electron microscopy)

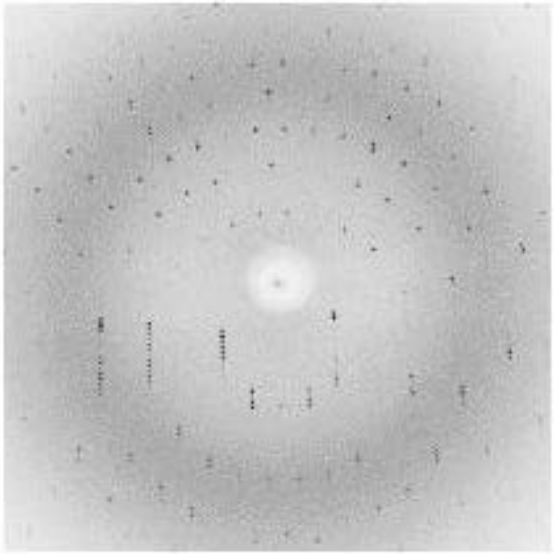
Size of protein molecules (diameter)

- cell ($1x10^{-6}$ **m**) μ microns
- ribosome ($1x10^{-9}$ **m**) nanometers
- protein ($1x10^{-10}$ **m**) angstroms

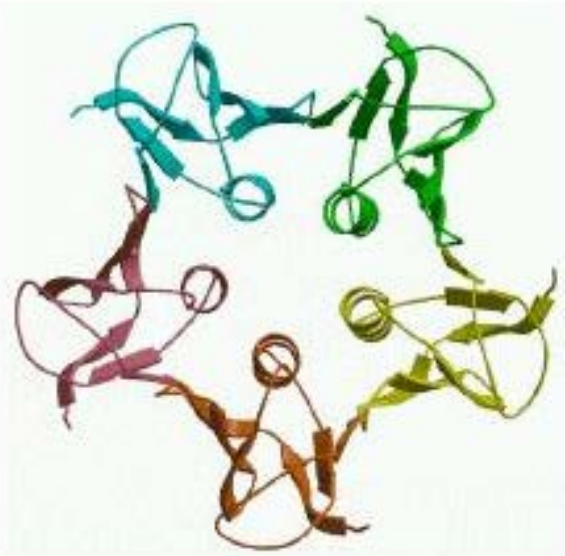
X-ray Crystallography

- Microscope is not suitable for distance smaller than the wavelength of the light you are using.
- X-rays get us in the right wavelength range. Each protein has a unique **X-ray diffraction** pattern**.**

Crystallization



Diffraction



Conversion of Diffraction Data to Electron Density and Image reassemble

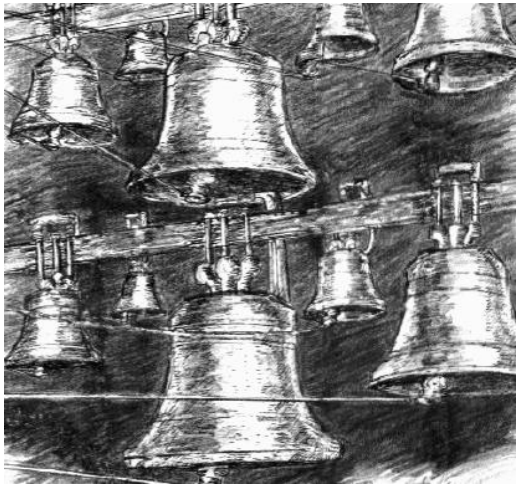
Figure from: http://www-structmed.cimr.cam.ac.uk/Course/Overview/Overview.html

Nuclear Magnetic Resonance (NMR)

- Is based on the quantum mechanical properties of atoms (spin) and it determines information about atoms from the their response to applied magnetic fields.

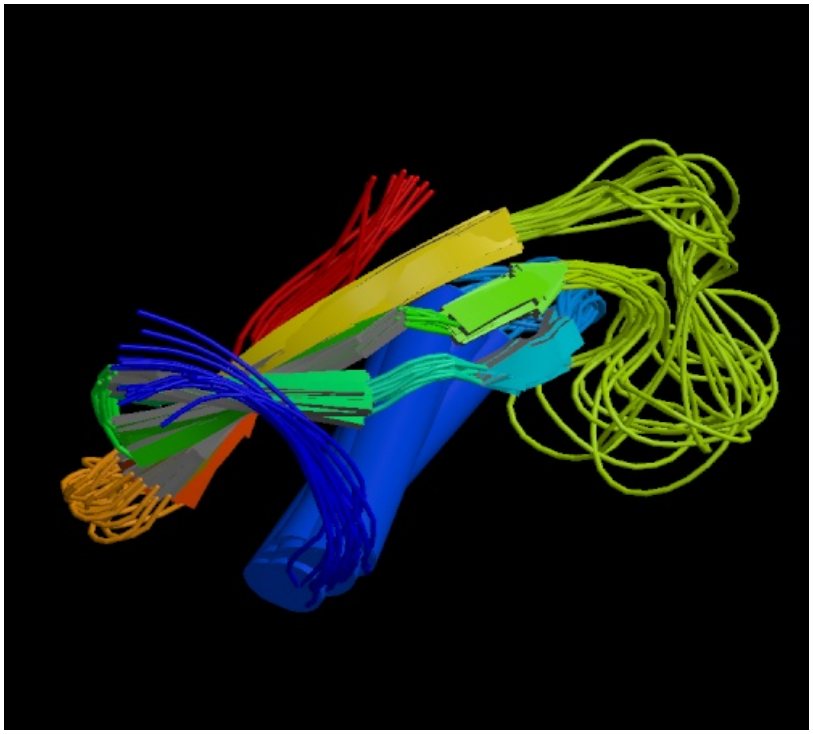
- Provides the interatomic distances, and features of the spectrum that can be interpreted in terms of torsion angles.

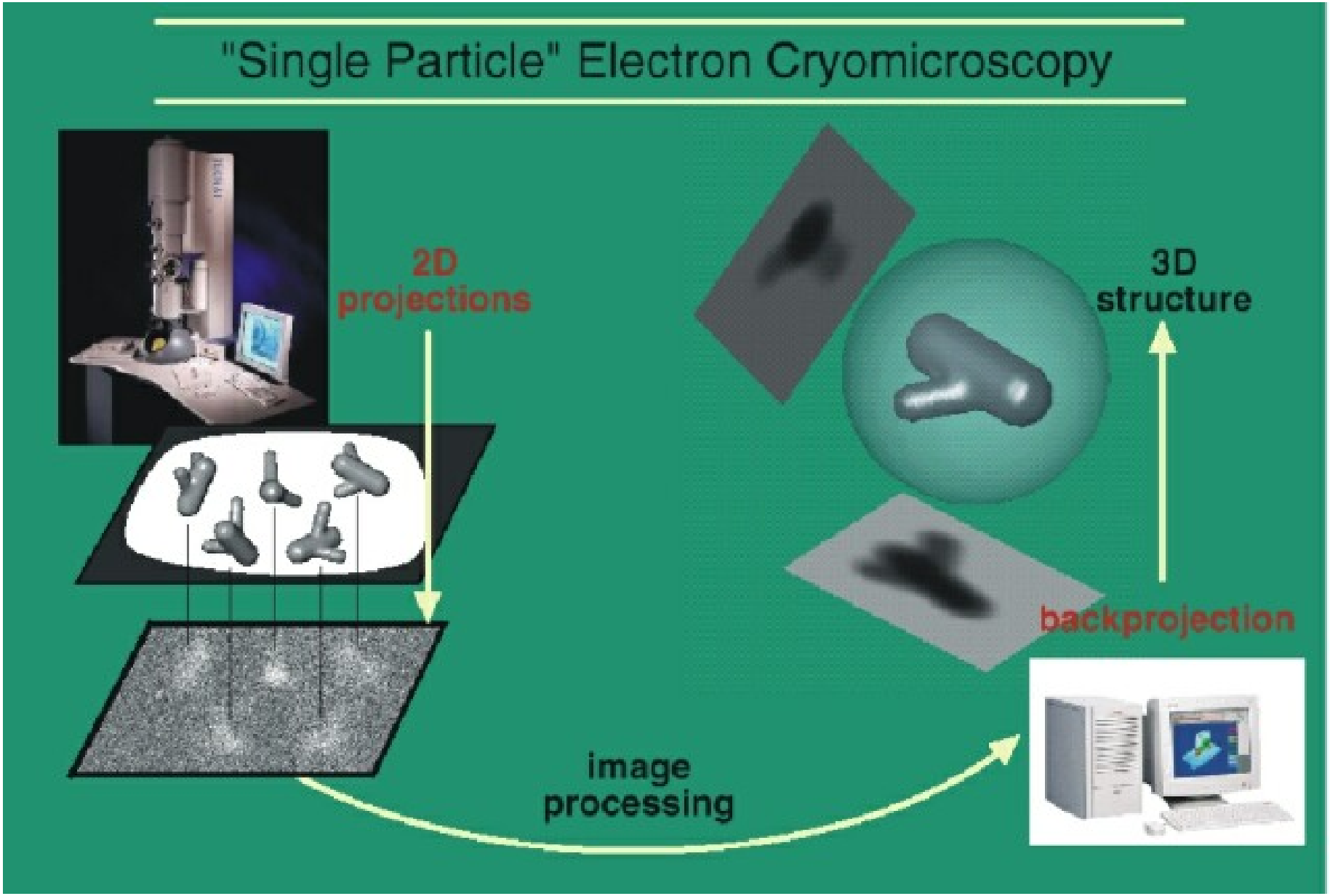
- Solved by Distance Geometry methods.



An NMR result is an ensemble of models

Cystatin (**1a67)**





H.J. Wolfson - Structural Bioinformatics

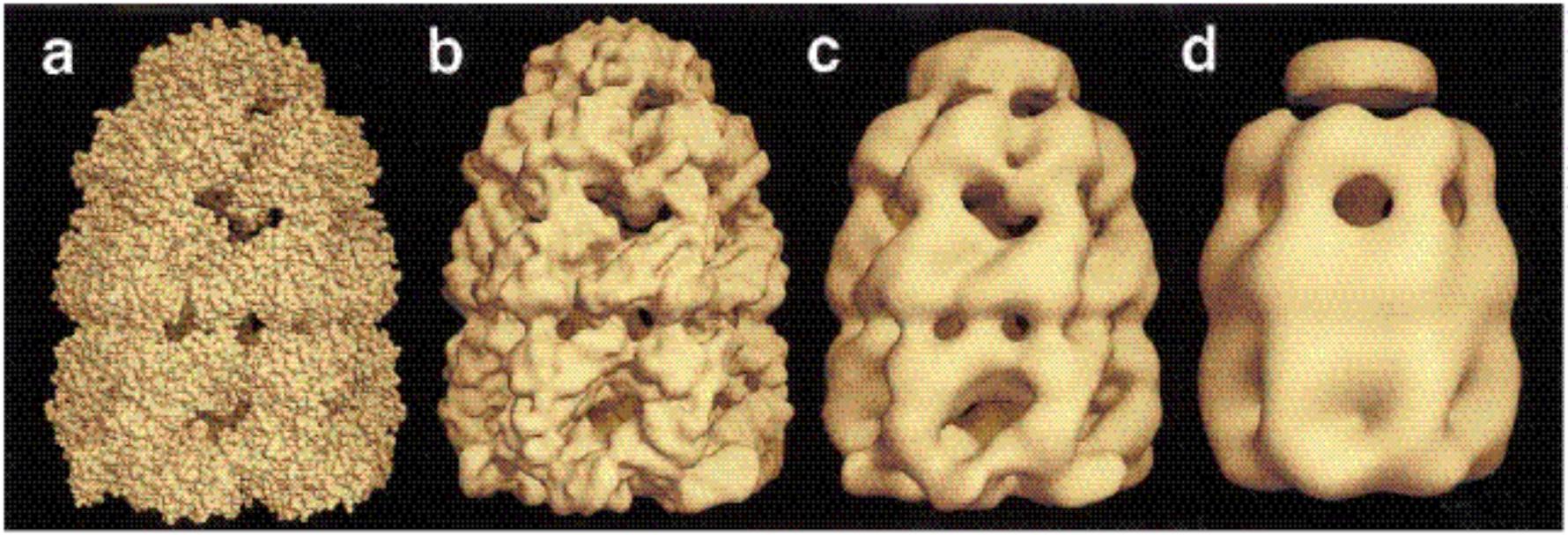
http://www.mpibpc.gwdg.de/abteilungen/103/single_part.htm

EM vs. Crystallography & NMR

	EM	Crystallography	NMR
Physical limits	Frozen (mostly easy)	Crystal (difficult)	In solution (easy) Metal atoms cause problems
Time	Fast	Slow	Medium
Resolution	High to low (3 -30 A)	High (< 3A)	High (< 3A)
Possible Structure Size	Big (structures containing many proteins)	Small	Very small (single proteins) <300aa

High Resolution to Low Resolution

High resolution → Low resolution

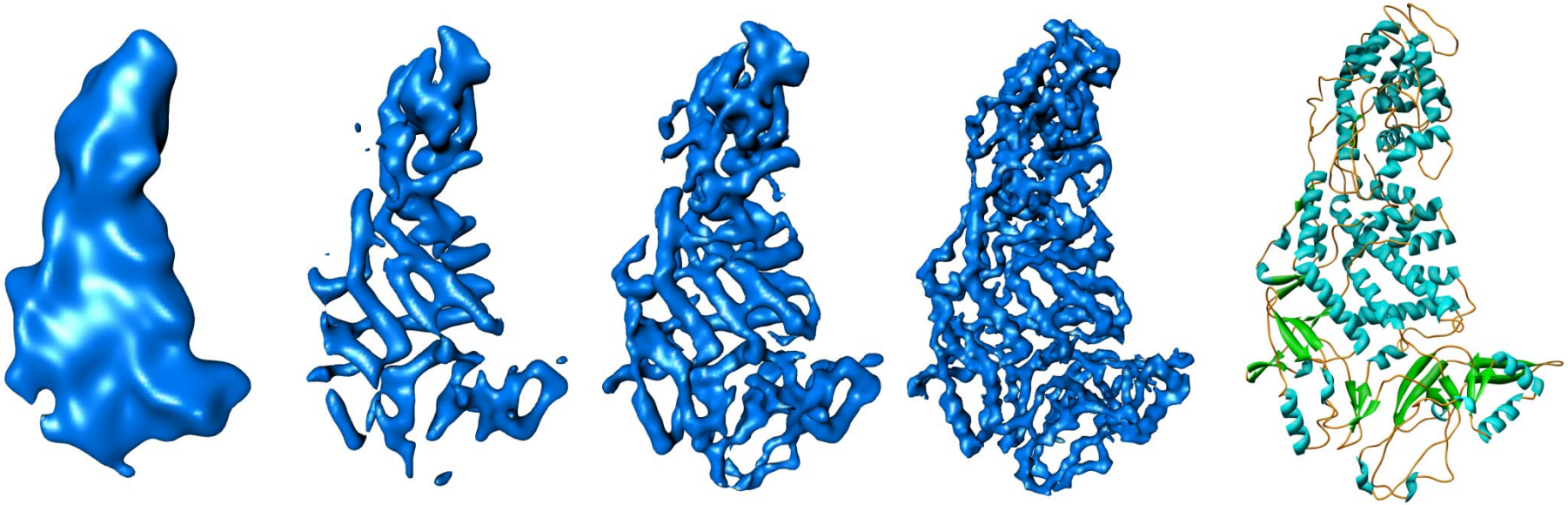


Space-filling model

4Å — 10Å — 20Å

Crystallography — EM

Features as a Function of Resolution



Low Resolution | Intermediate Resolution | High Resolution | 2BTV VP3A

15+ Å | 9 Å | 6 Å | <4 Å

Size
Shape

Domains

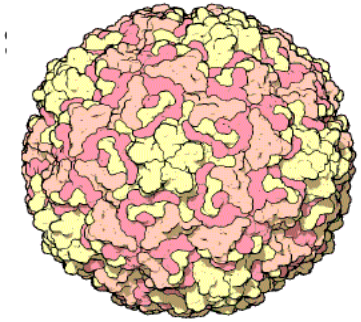
α Helices
β sheets

Strands
Connectivity

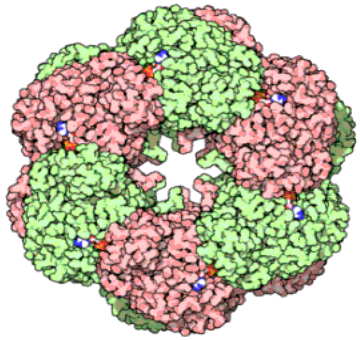
Sidechains

Proteins work together

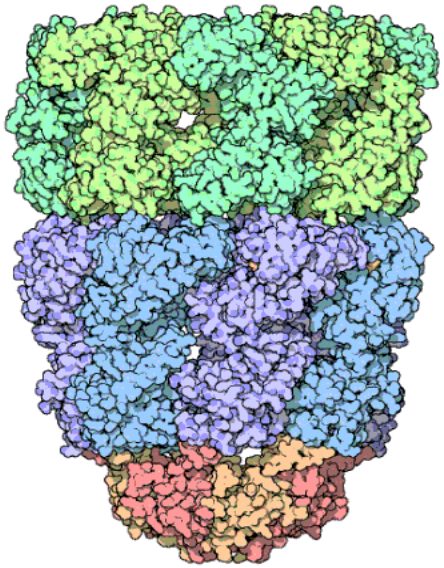
- Vital cellular functions are performed by complexes of proteins.
- Structures of single proteins are usually not informative about function if taken out of context



Rhinovirus



Glutamine Synthetase

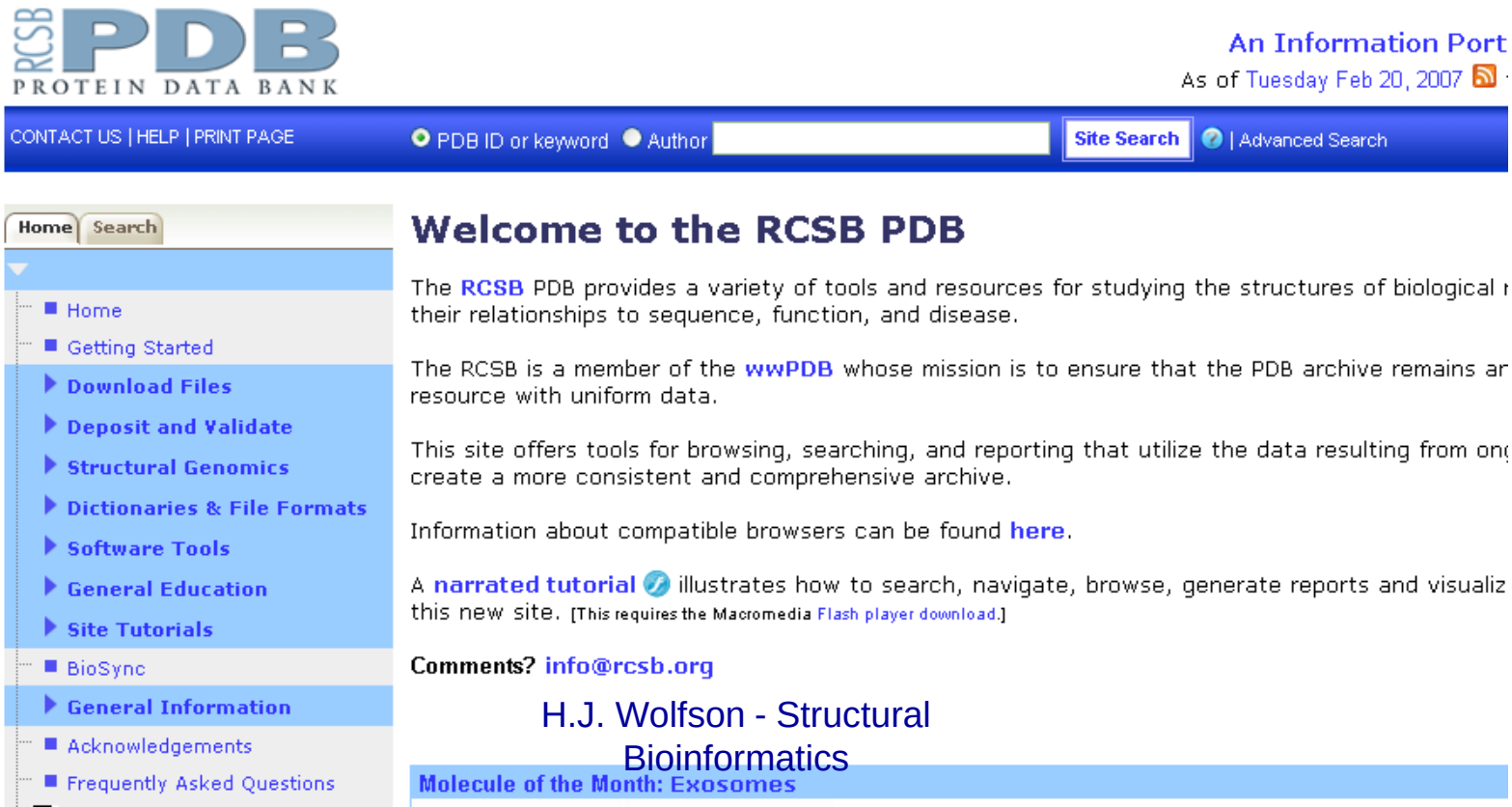


Chaperon

The figures are adapted from ***http://www.rcsb.org/pdb/molecules/molecule_list.html***

The Protein Data Bank (PDB)

- International repository of 3D molecular data.
- Contains x-y-z coordinates of all atoms of the molecule and additional data.



dings.do

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | MyPDB Login

RCSB PDB PROTEIN DATA BANK

An Information Portal to 125526 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands | Go

Advanced Search | Browse by Annotations

PDB-101 | WORLDWIDE PDB PROTEIN DATA BANK | EMDataBank | NUCLEIC ACID DATABASE StructuralBiology Knowledgebase | Worldwide Protein Data Bank Foundation

PDB Current Holdings Breakdown

Jan 8, 2017

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	105028	1796	5389	4	112217
NMR	10239	1187	237	4 8	11671
ELECTRON MICROSCOPY	966	30	335	0	1331
HYBRID	97	3	2	1	103
other	181	4	6	13	204
Total	116511	3020	5969	26	125526

(Click on any number to retrieve the results from that category.)
101837 structures in the PDB have a structure factor file.
8993 structures in the PDB have an NMR restraint file.
2762 structures in the PDB have a chemical shifts file.
1316 structures in the PDB have a 3DEM map file.

H.J. Wolfson - Structural Bioinformatics

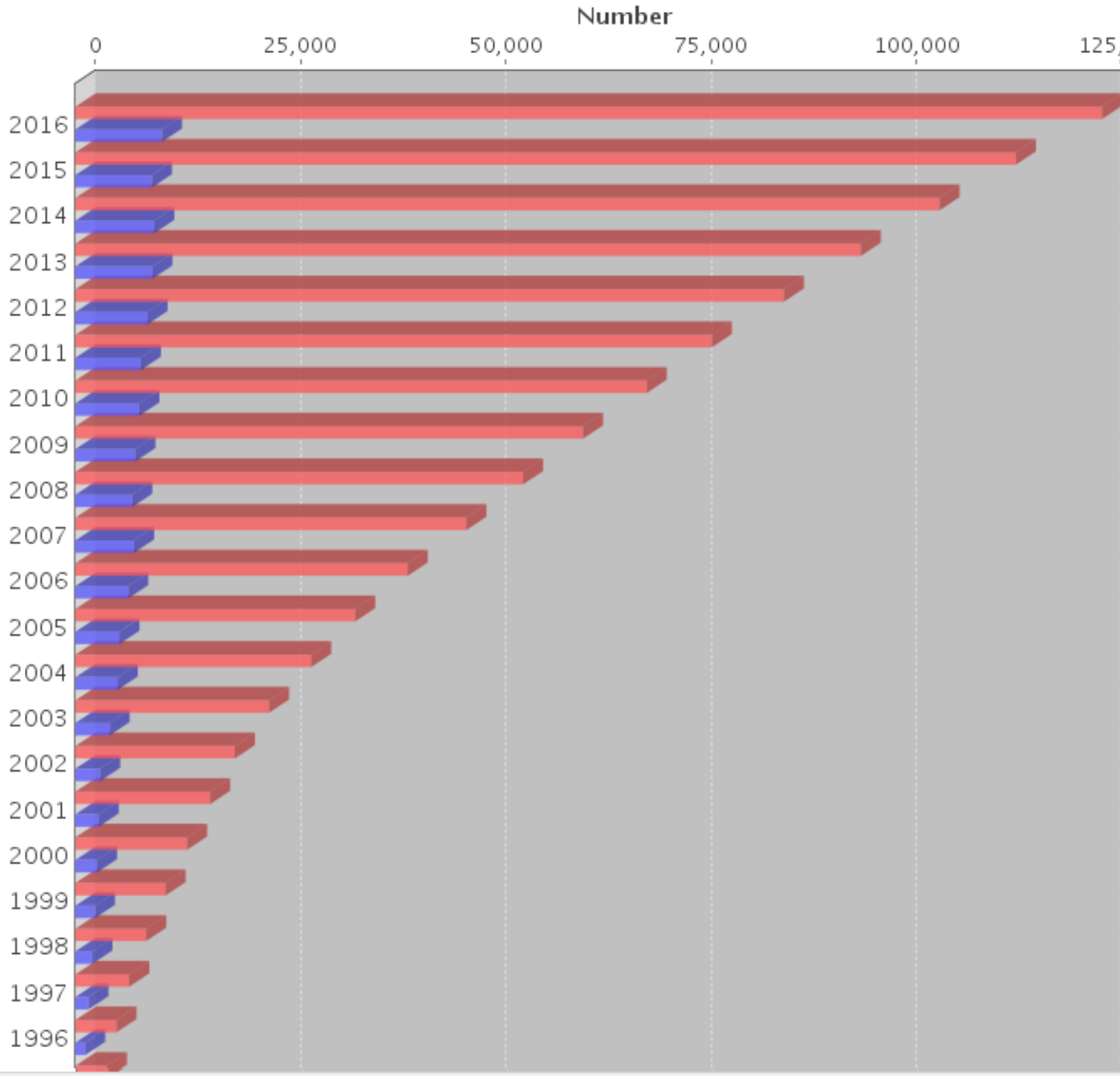
About About Us | Citing Us | Publications | Team | Careers | Usage & Privacy

Help Contact Us | Help Topics | Website FAQ | Browser Test | Glossary

RCSB PDB (citation) is managed by two members of the Research Collaboratory for Structural Bioinformatics: Rutgers and UCSD/SDSC

Yearly Growth of Total Structures

number of structures can be viewed by hovering mouse over the bar



Major Protein Structure Classification Repositories

- **SCOP**
  http://scop.mrc.lmb.cam.ac.uk/scop/

- **CATH**
  http://www.biochem.ucl.ac.uk/bsm/cath/

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).**

Protein Structure Prediction-Folding

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

Target sequence

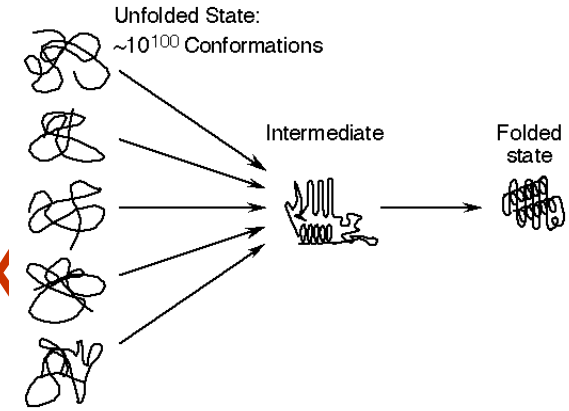
M	A	A	G	Y	A	V	L	S

structural model

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)

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

Arne Elofsson
(arne@bioinfo.se)

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.

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

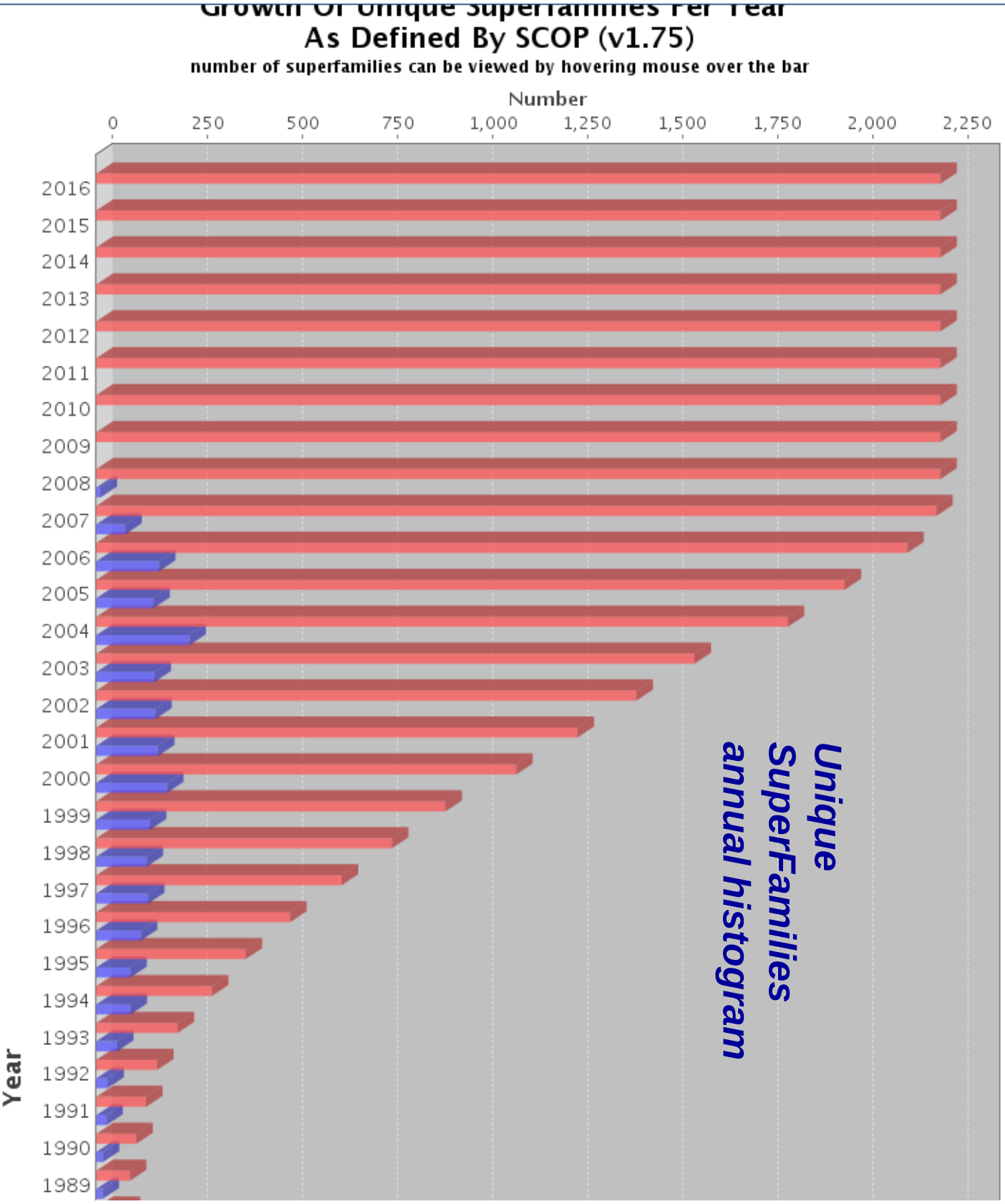
Fold Recognition – Threading

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.

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.


Growth Of Unique Superfamilies Per Year
As Defined By SCOP (v1.75)
number of superfamilies can be viewed by hovering mouse over the bar
Number
0
250
500
750
1,000
1,250
1,500
1,750
2,000
2,250
Year
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
1996
1995
1994
1993
1992
1991
1990
1989
Unique
SuperFamilies
annual histogram

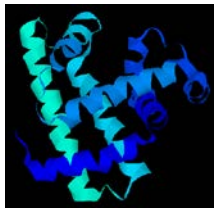
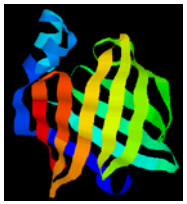
Protein Threading

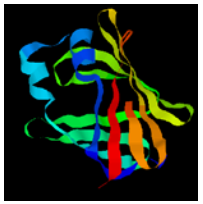
- Make a structure prediction through finding an optimal placement (threading) of a protein sequence onto each known structure (structural template)
  - “placement” quality is measured by some statistics-based energy function
  - best overall “placement” among all templates may give a structure prediction

target sequence

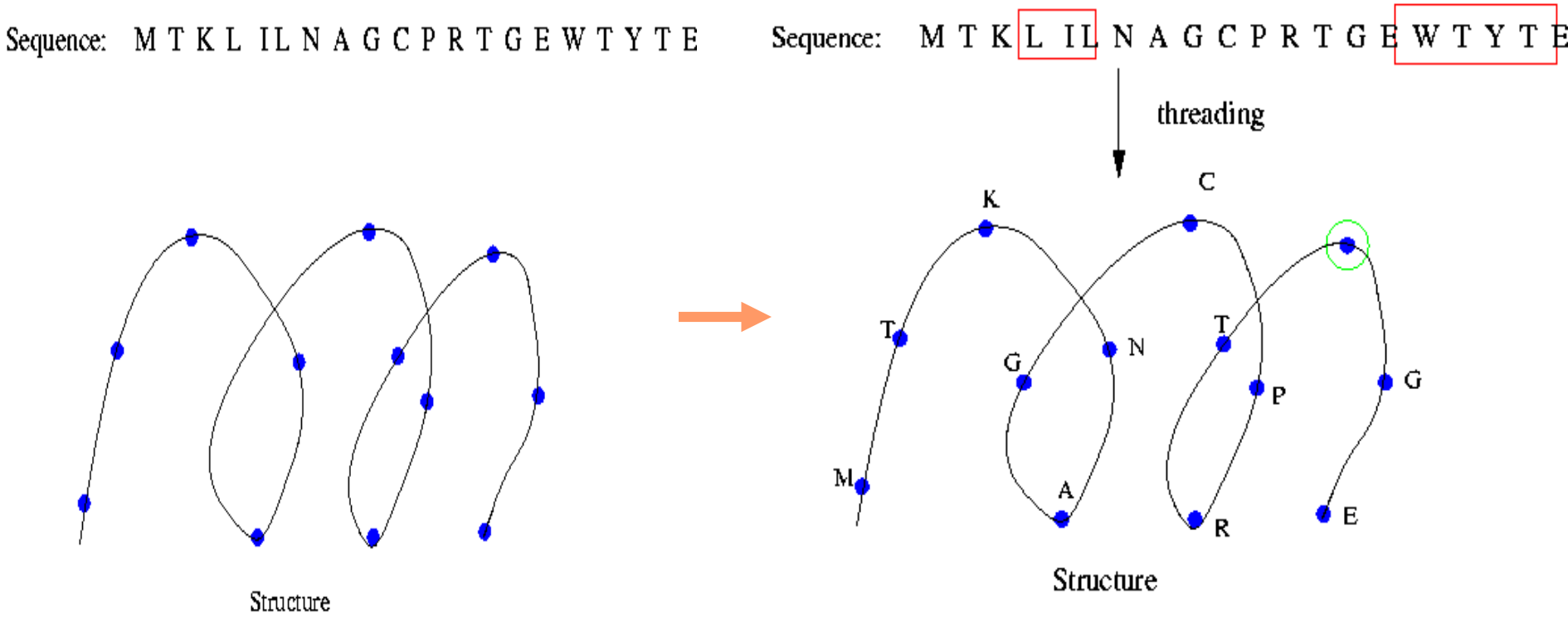
MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

template library



Threading Example



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

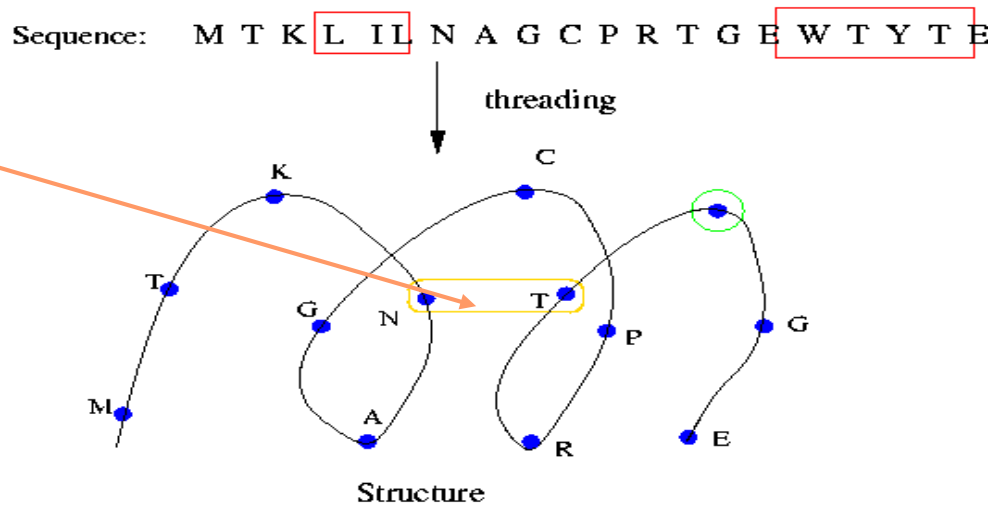
Assumptions :

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.

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)

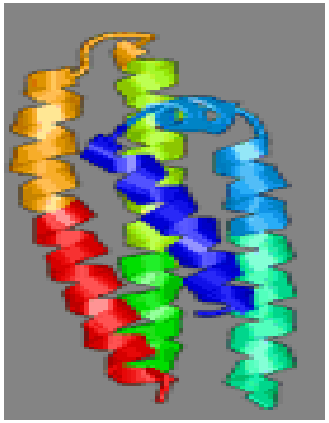
Consistency with the secondary structures: E_{ss}

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

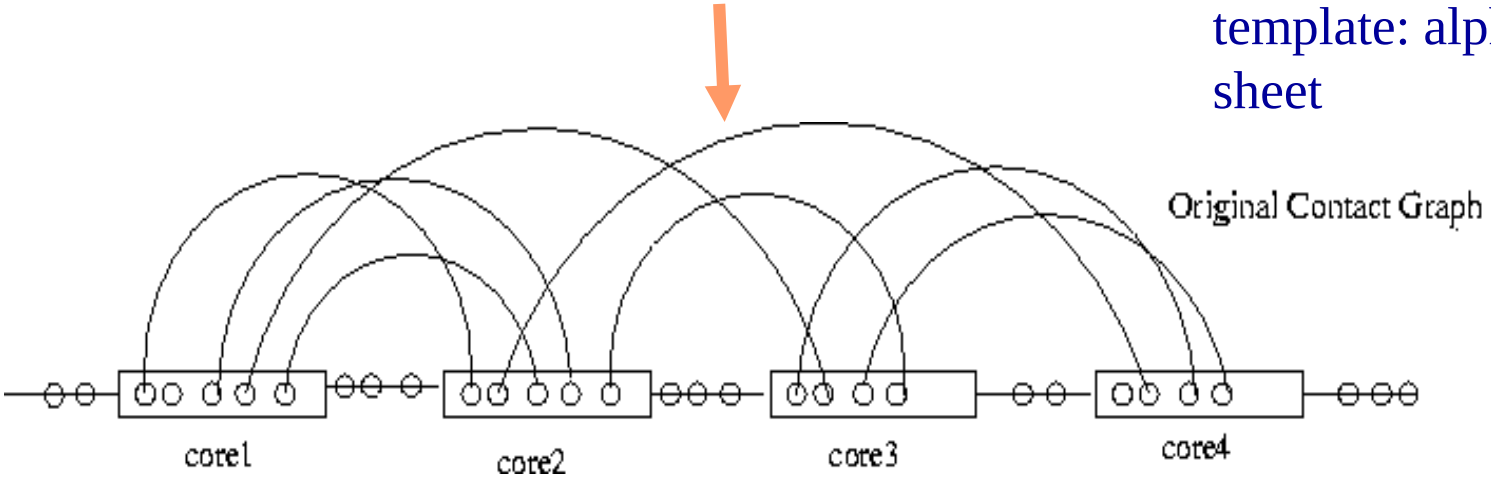
Minimize E to find a sequence-structure alignment

Contact Graph

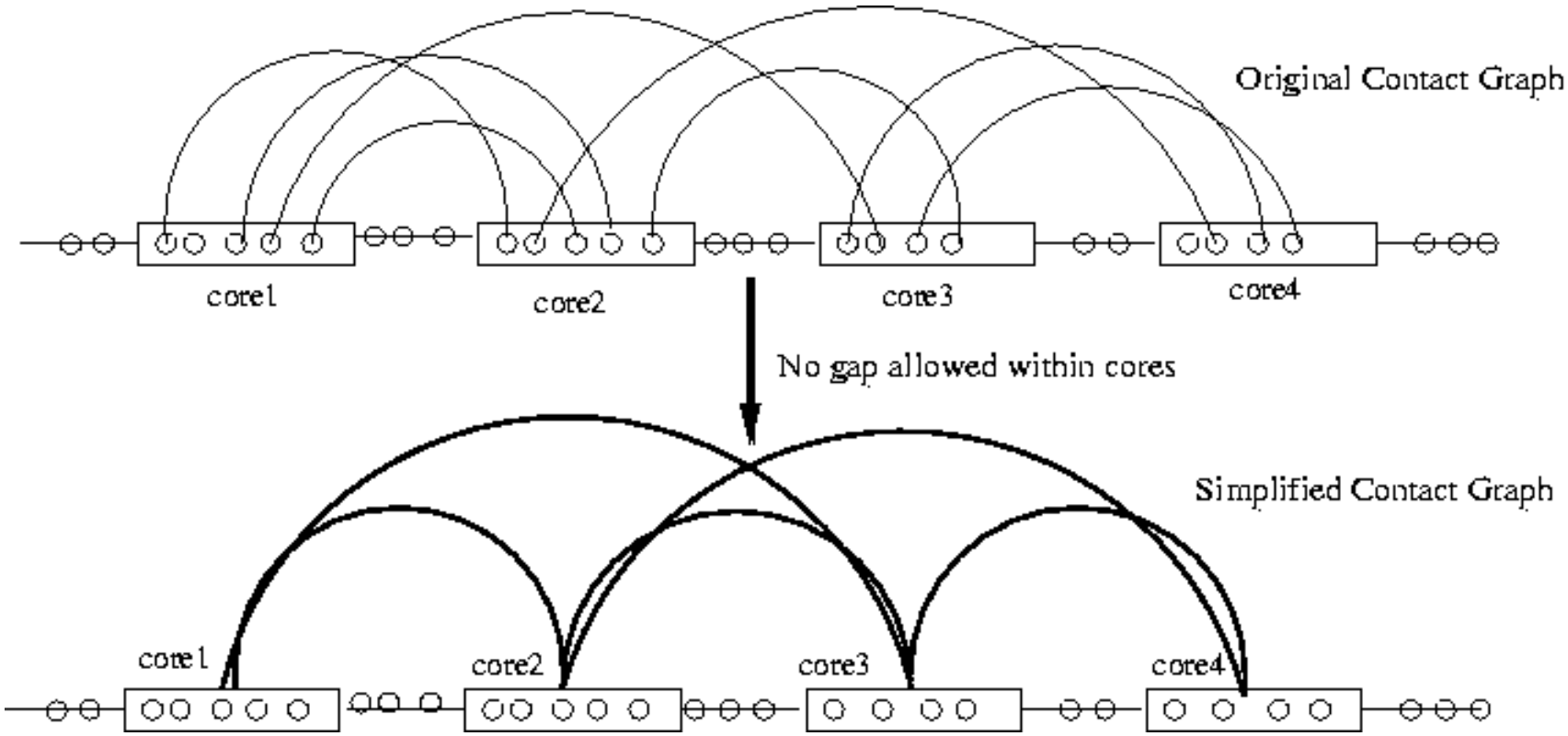
template



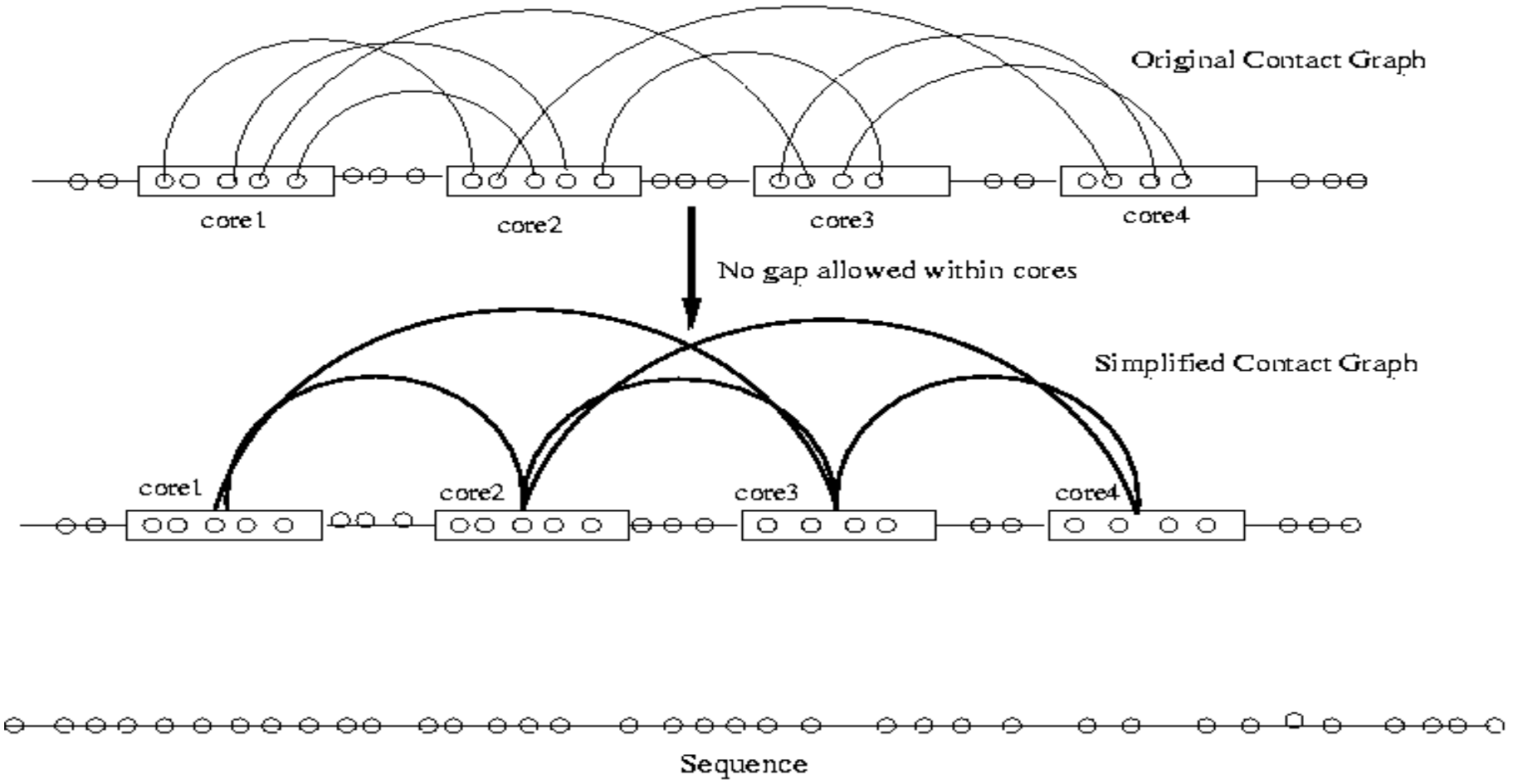
1. Each residue as a vertex
2. One edge between two residues if their spatial distance is within a given cutoff.
3. Cores are the most conserved segments in the template: alpha-helix, beta-sheet



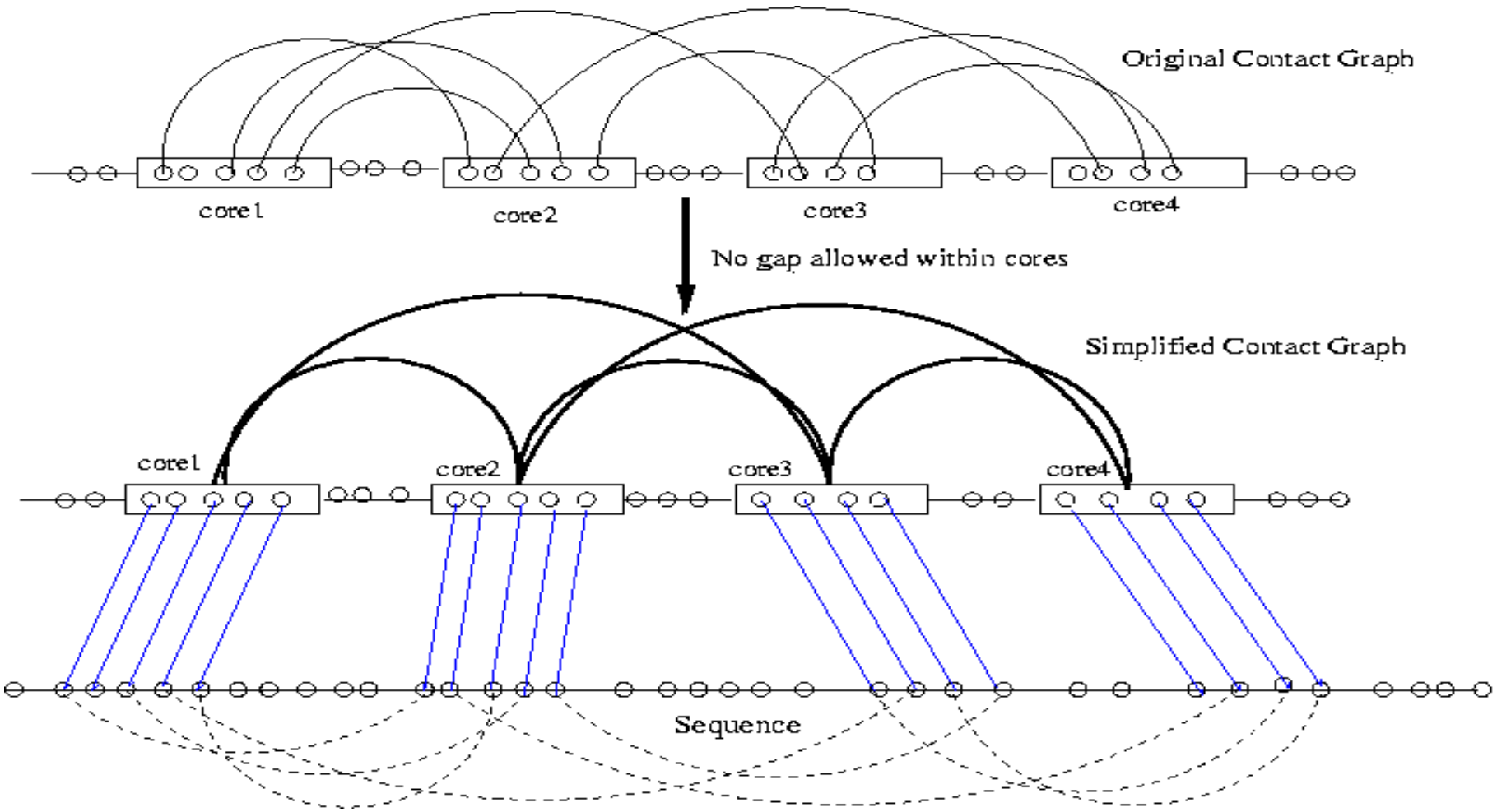
Simplified Contact Graph



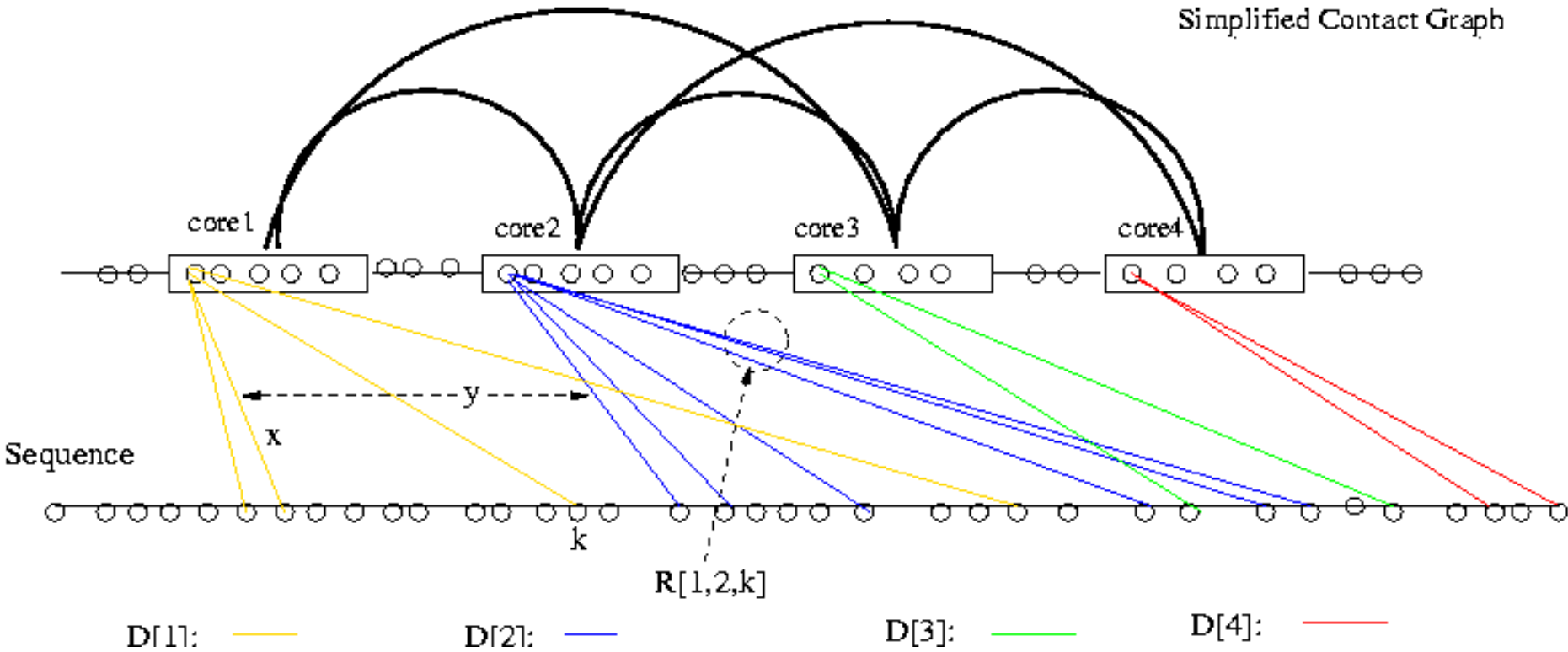
Contact Graph and Alignment Diagram



Contact Graph and Alignment Diagram



Variables



- 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).

Formulation 1

$$\textit{Minimize}$$

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

$$s.t.$$

$$x_{i,l} + x_{i+1,k} \leq 1$$

$$y_{(i,l)(j,k)} = x_{i,l} x_{j,k}$$

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

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

E_g , E_p

E_s , E_{ss} , E_m

Encodes scoring system

Encodes interaction structures: the first makes sure no crosses; the second is quadratic, but can be converted to linear: a=bc is equivalent to: a≤b, a≤c, a≥b+c-1

The constraint set is as follows:

$$\sum_{j \in D[i]} x_{i,j} = 1, \quad i = 1, 2, \ldots, M\,; \tag{8}$$

$$\sum_{l \geq l_0, l \in D[i]} x_{i,l} + \sum_{k \in D[i+1] - R[i,i+1,l_0]} x_{i+1,k} \leq 1, \quad l_0 \in D[i],\ i = 1, 2, \ldots, M-1\,; \tag{9}$$

$$\sum_{k \in R[i,j,l]} y_{(i,l),(j,k)} \leq x_{i,l}, \quad \forall l \in D[i],\ i, j = 1, 2, \ldots, M\,; \tag{10}$$

$$\sum_{l \in R[j,i,k]} y_{(i,l),(j,k)} \leq x_{j,k}, \quad \forall k \in D[j],\ i, j = 1, 2, \ldots, M\,; \tag{11}$$

$$\sum_{k \in R[i,j,l]} y_{(i,l),(j,k)} \geq x_{i,l} + \sum_{k \in R[i,j,l]} x_{j,k} - 1, \quad l \in D[i],\ i, j = 1, 2, \ldots, M\,; \tag{12}$$

$$\sum_{l \in R[j,i,k]} y_{(i,l),(j,k)} \geq x_{j,k} + \sum_{l \in R[j,i,k]} x_{i,l} - 1, \quad k \in D[j],\ i, j = 1, 2, \ldots, M\,; \tag{13}$$

$$x_{i,j} \in \{0, 1\}, \quad j \in D[i],\ i = 1, 2, \ldots, M\,; \tag{14}$$

$$y_{(i,l)(j,k)} \in \{0, 1\}, \quad \forall l \in D[i],\ k \in D[j],\ i, j = 1, 2, \ldots, M\,. \tag{15}$$

Constraint 8 says that one core can be aligned to a unique sequence position, i.e. given core i, only one of the $x_{i,j}$'s is 1, for $j \in D[i]$. Constraint 9 forbids the conflicts between the adjacent two cores. Based on the transitivity of *non-conflict* (see Lemma 2), this constraint guarantees that there are no conflicts between any two cores if variable x and y are integral. Therefore, it guarantees that the integral solution corresponds to a valid alignment. Constraints 10 and 11 say that at most one interaction variable can be 1 between any two cores that have interactions between them. Constraints 12 and 13 enforce that if two cores have their alignments to the sequence respectively and also have interactions between them, then at least one interaction variable should be 1. Constraints 14 and 15 guarantee x and y variables to be either 0 or 1. The integer program as formulated above does not

Formulation used in RAPTOR

$$\textit{Minimize}$$

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

$$s.t.$$

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

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

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

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

E_g, E_p

E_s, E_{ss}, E_n

Encodes scoring system

Encodes interaction structures

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

Docking

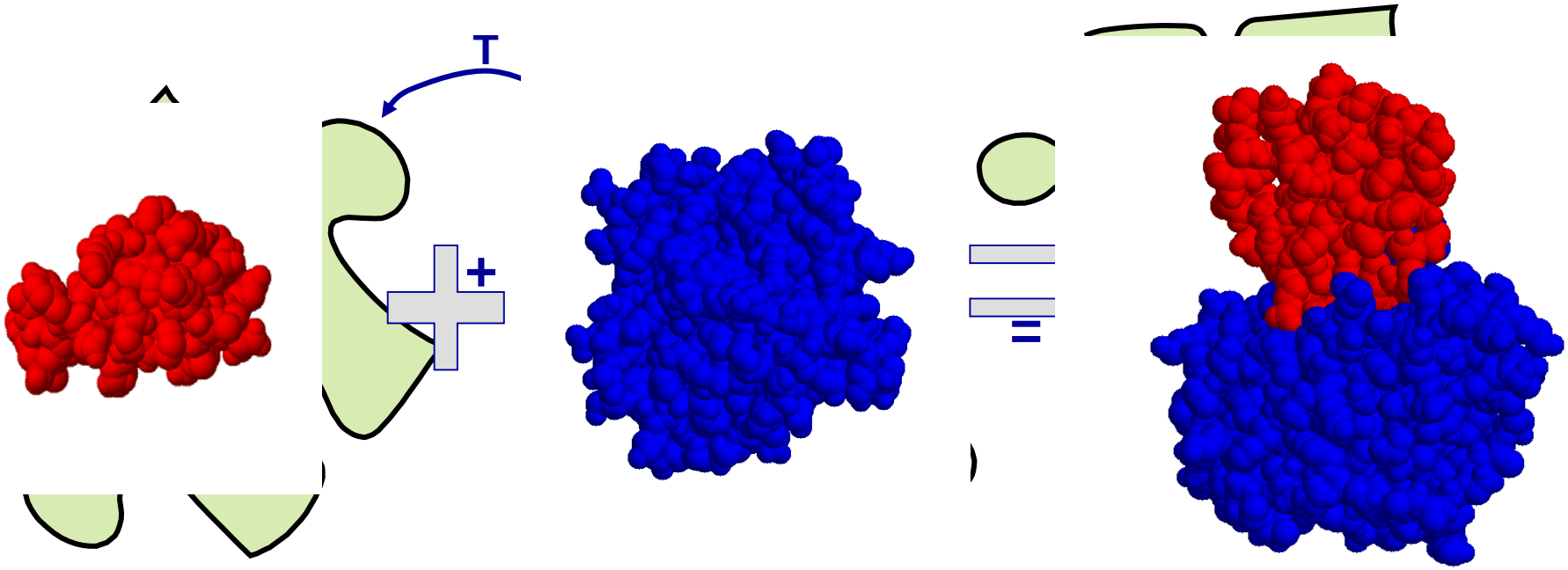
עגינה

<u>הגדרת הבעיה:</u>

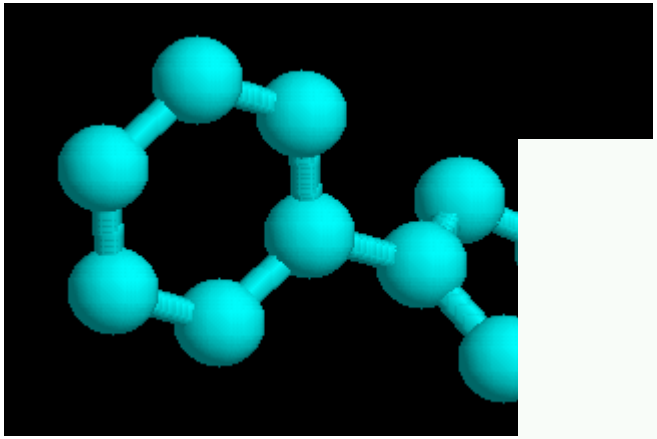
בהינתן שני מבנים מצא טרנספורמציה במרחב שתביא למקסימום את אינטראקציה ביניהם (תן חיזוי למבנה המשותף).

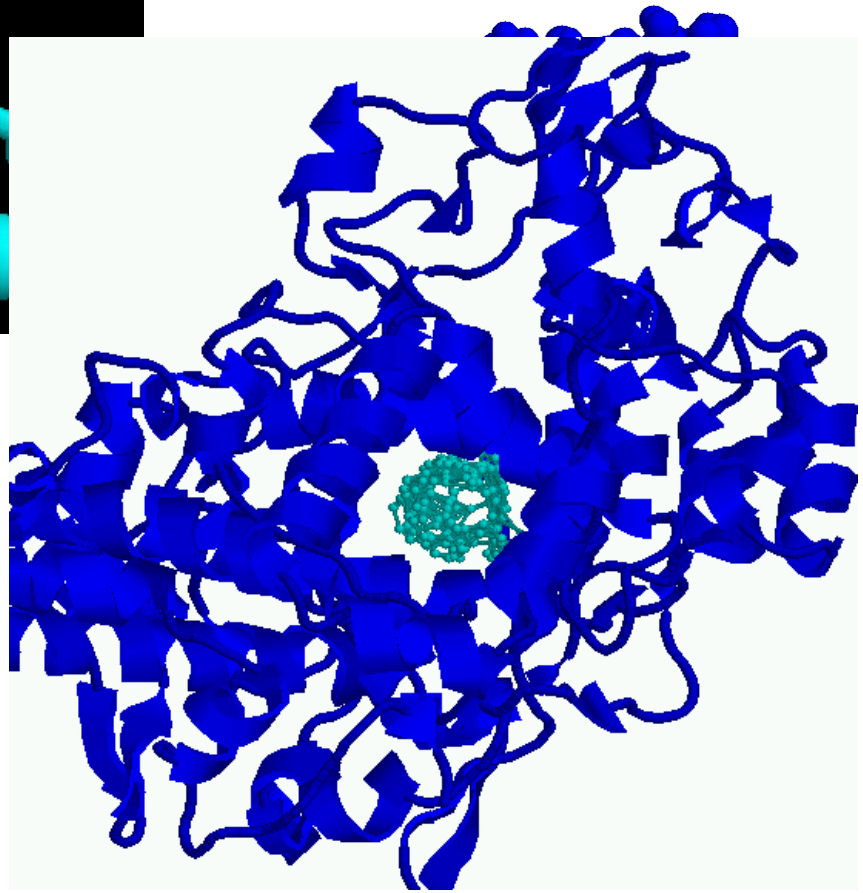
Docking Problem

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



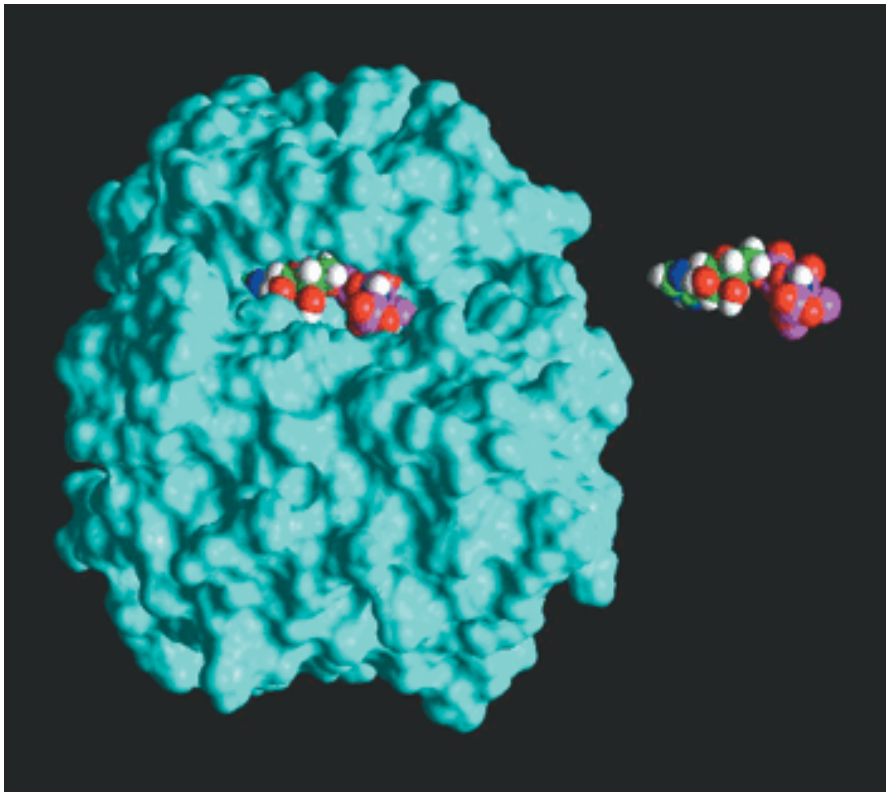
Docking Problem:





= ?

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



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.

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.

Forces Governing Biomolecular Recognition

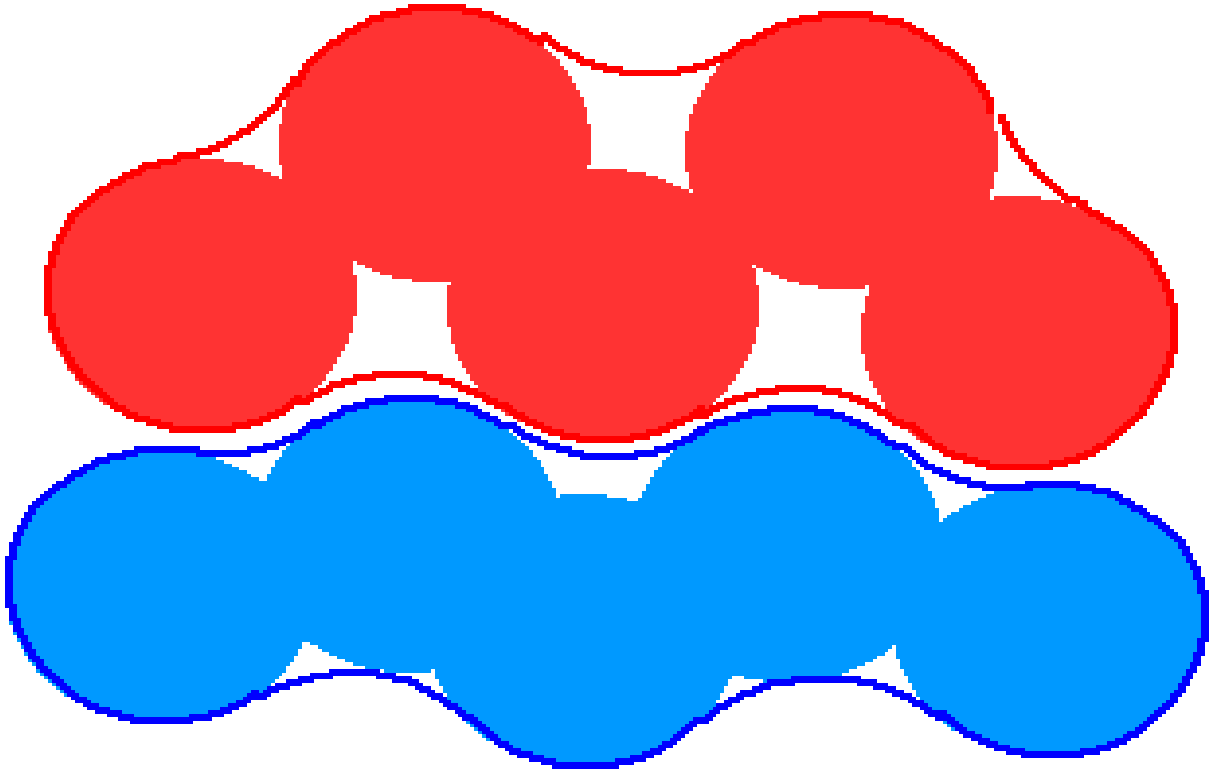
Depend on the molecules and the solvent.

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

All interactions act at short ranges.

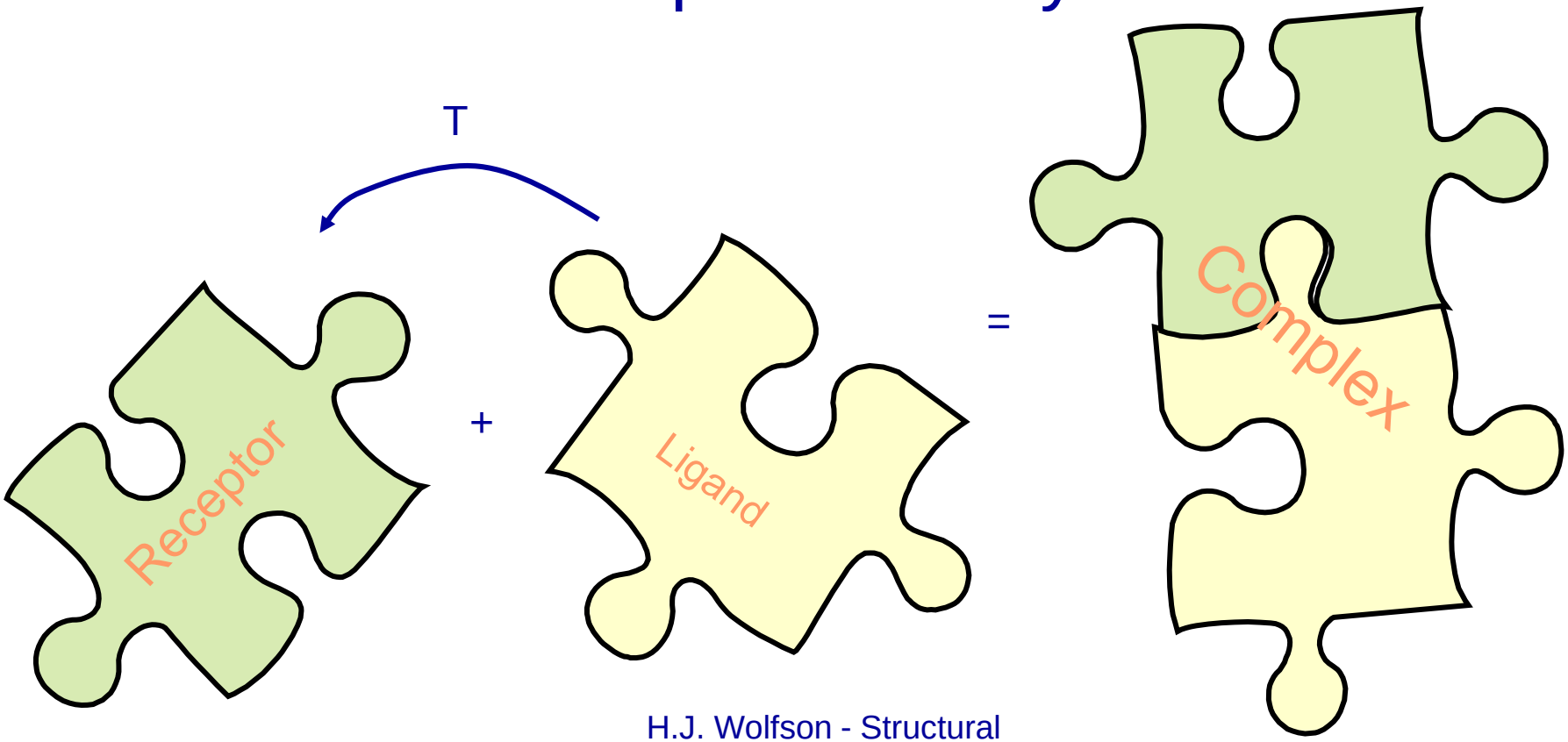
Implies that a necessary condition for tight binding is surface complementarity.

Shape Complementarity



Necessary Condition for Docking

- Given two molecules find significant surface complementarity.



Geometric Docking Algorithms

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

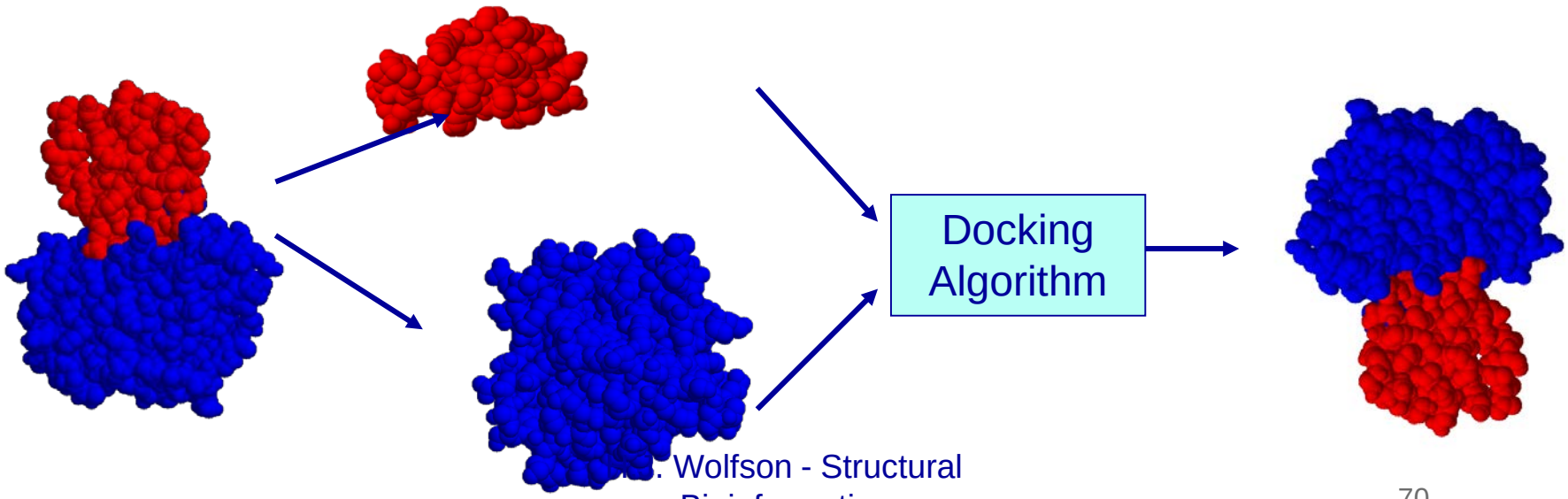
<u>Remark</u> : usually “protein” here can be replaced by “DNA” or “RNA” as well.

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.

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

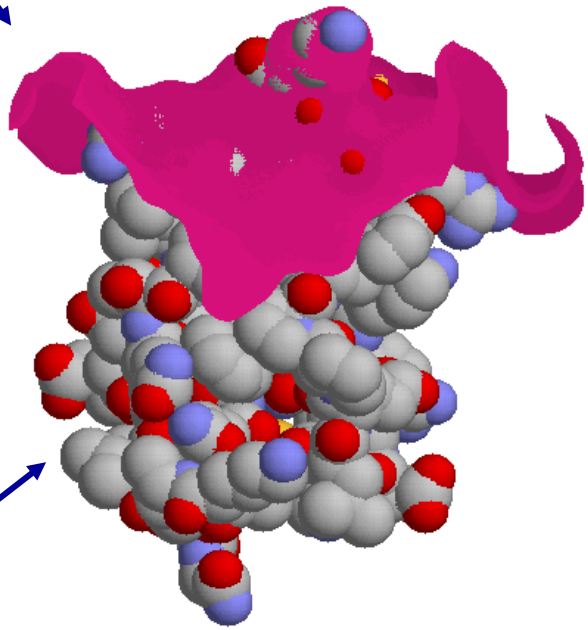
- In the unbound docking we are given 2 molecules in their native conformation.
- The goal is to find the correct association.

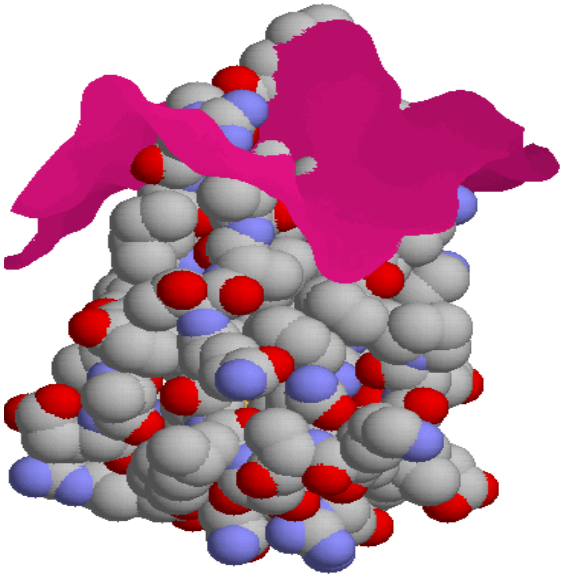
- **Problems:** conformational changes (side-chain and backbone movements), experimental errors in the structures.

Bound vs. Unbound

Receptor surface

10 highly penetrating residues





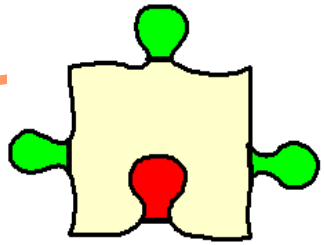
Ligand

Unbound ligand and receptor superimposed on the complex

Kallikrein A/trypsin inhibitor complex (PDB codes 2KAI,6PTI)

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**.

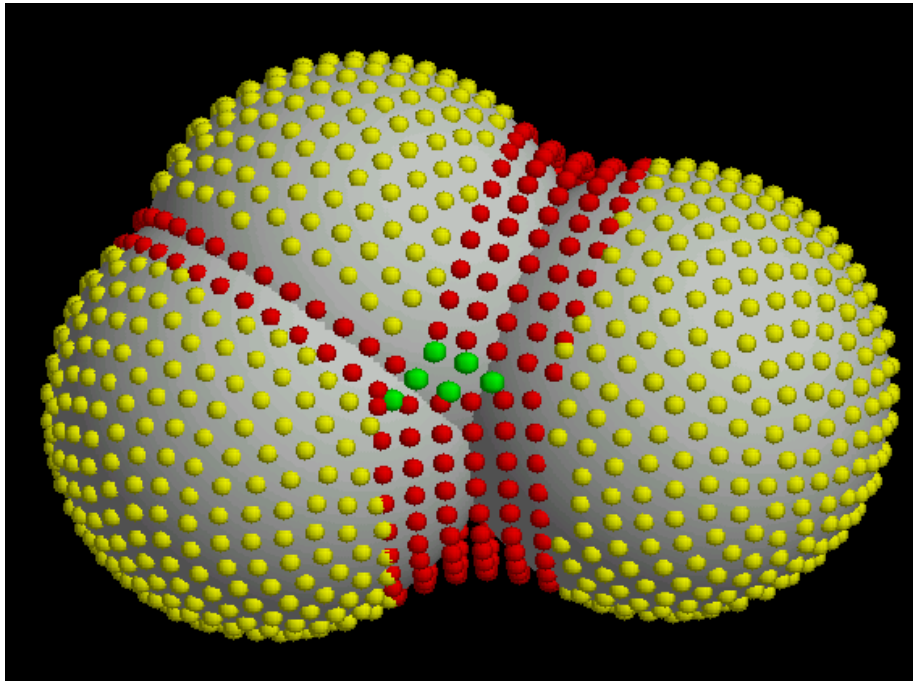


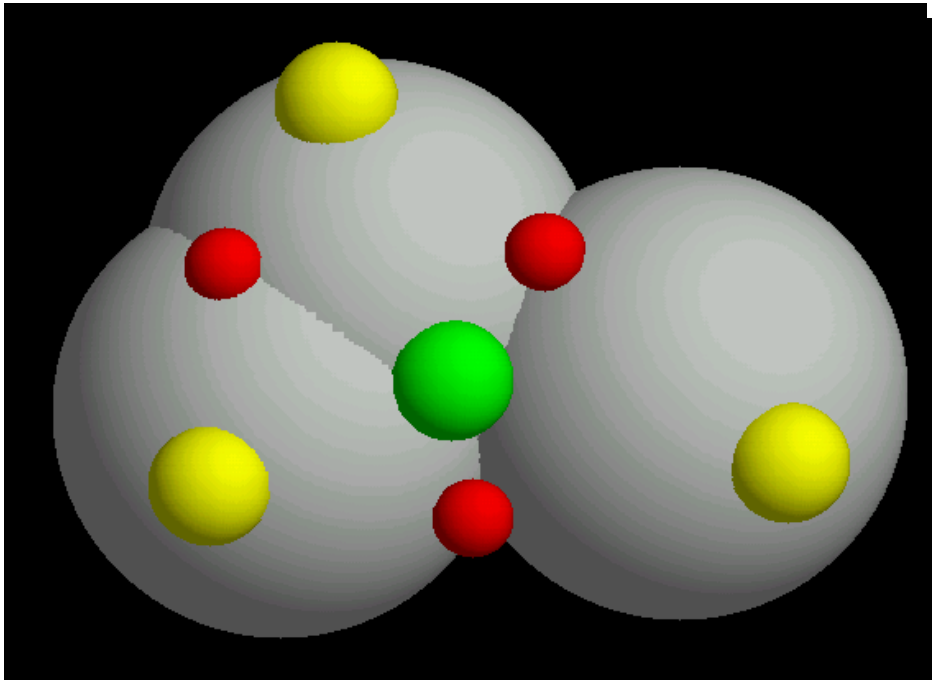
Docking Algorithm Scheme

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

1. Surface Representation

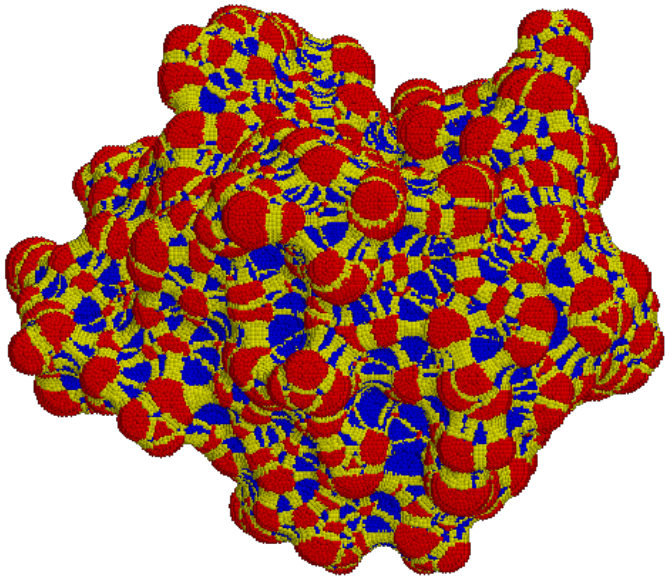
- Dense MS surface (Connolly)
- Sparse surface (Lin et al.)



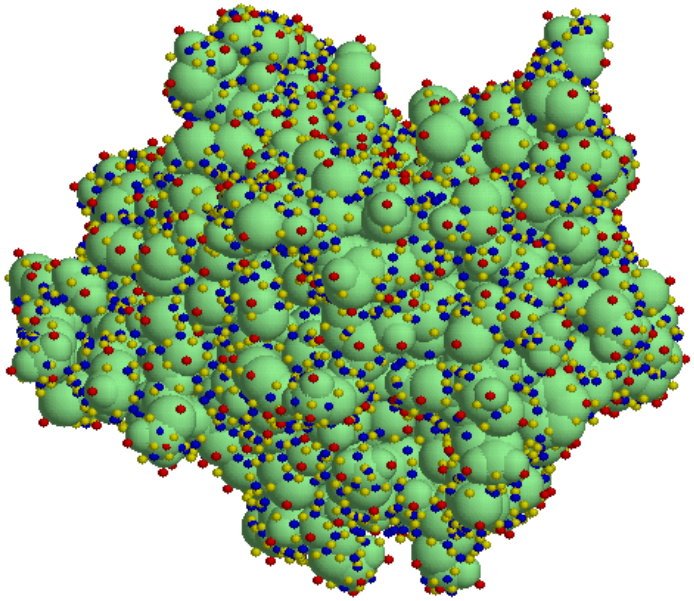


1. Surface Representation

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



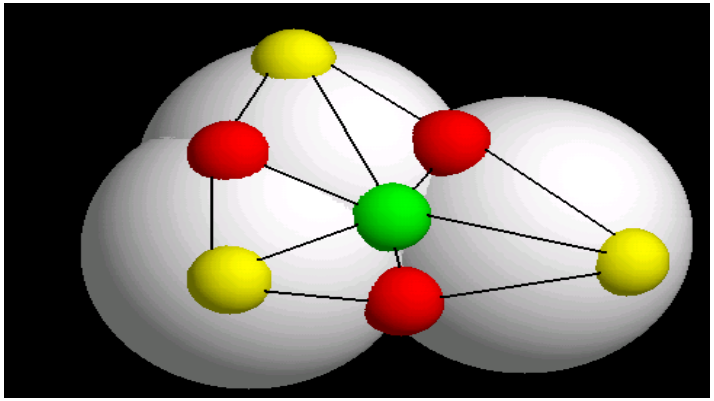
82,500 points



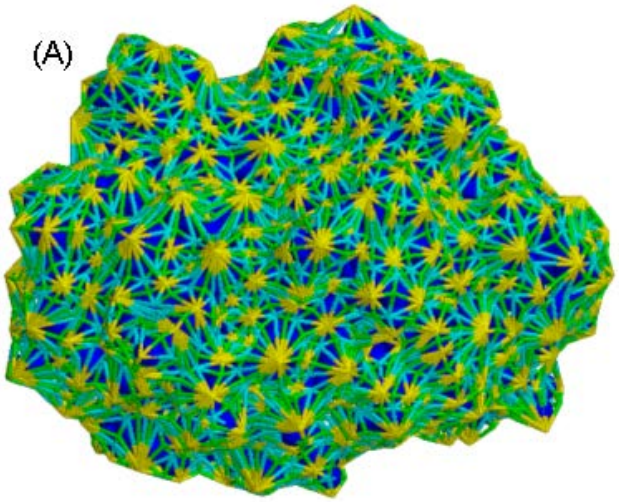
4,100 points

Sparse Surface Graph - G_{top}

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



- G_{top} – Surface topology graph:

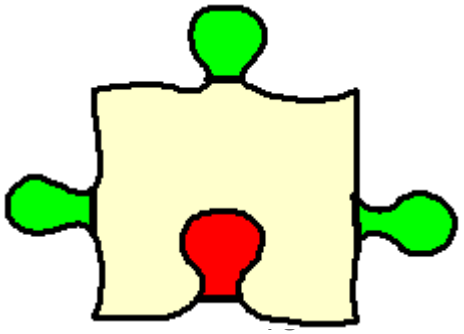


V = *surface points*

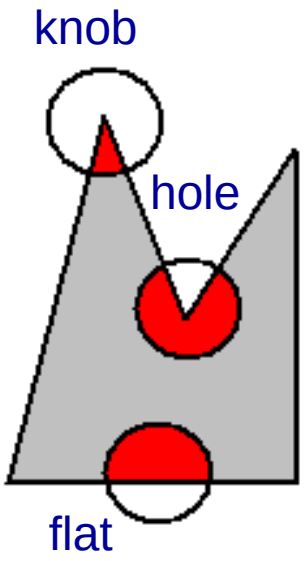
E = *{(u,v)| u,v belong to the same atom}*

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

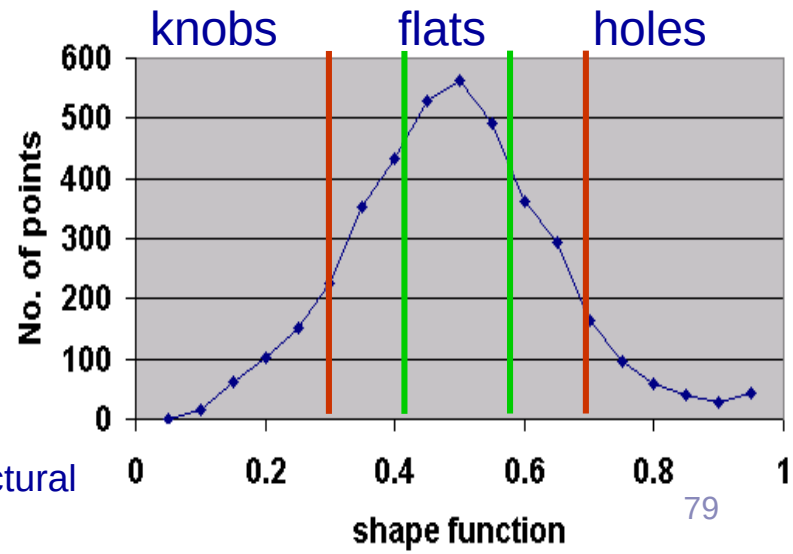


2.1 Curvature Calculation



- 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'.



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.

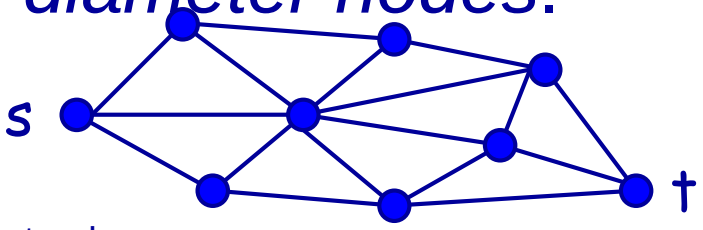
Patch Detection by Segmentation Technique

- 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*.



Split and Merge (cont.)

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

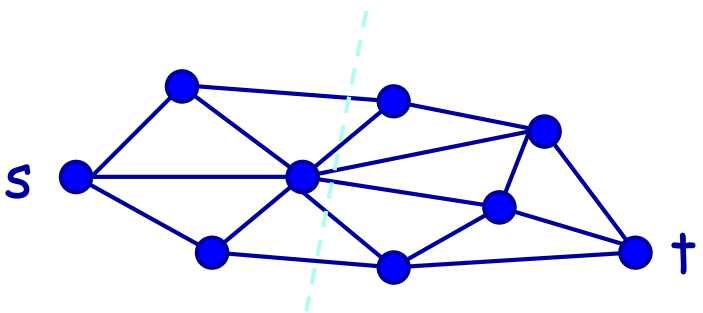
1. *low_patch_thr* ≤ ***diam*** ≤ *high_patch_thr* → **valid patch**
2. ***diam*** > *high_patch_thr* → **split**
3. ***diam*** < *low_patch_thr* → **merge**

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

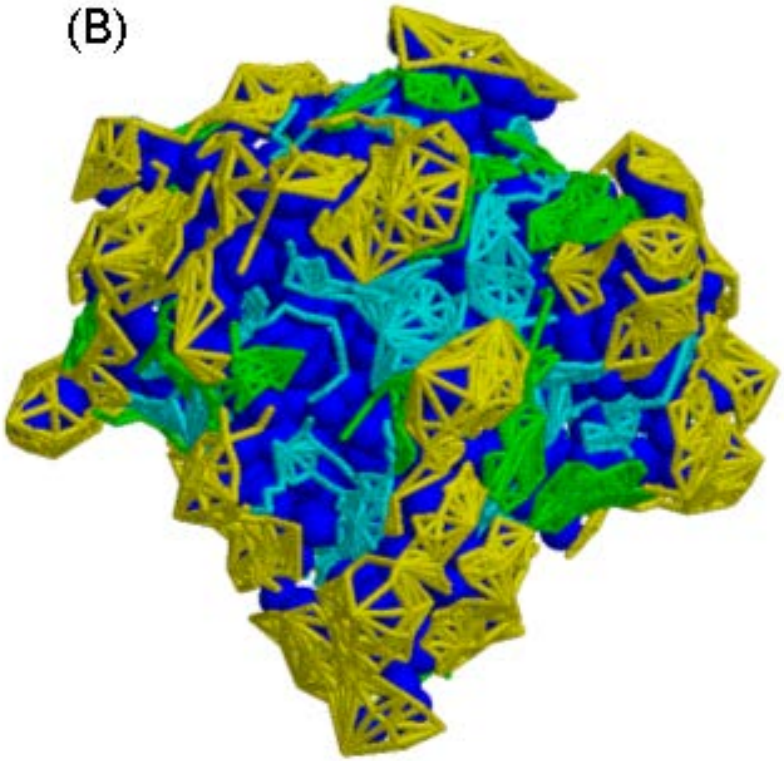
Split and Merge (cont.)

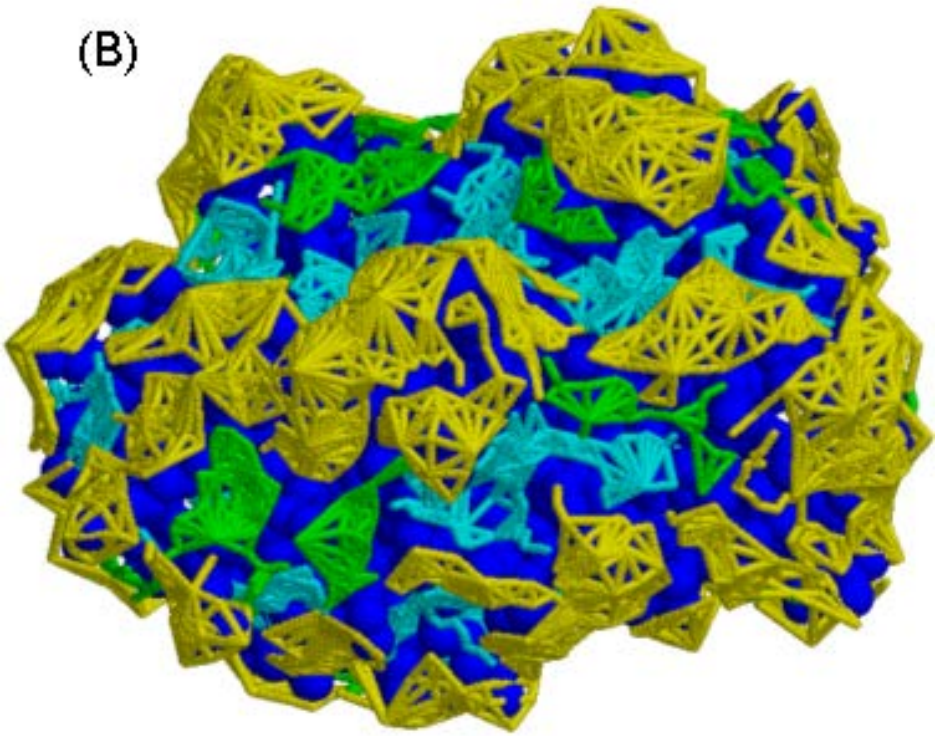
- **Split routine:** compute Voronoi cells of the diameter nodes *s,t*. Points **closer to *s*** belong to new component *S*, points **closer to *t*** belong to new component *T*. The split is applied until the new component has a valid diameter.
- **Merge routine:** compute the geodesic distance of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



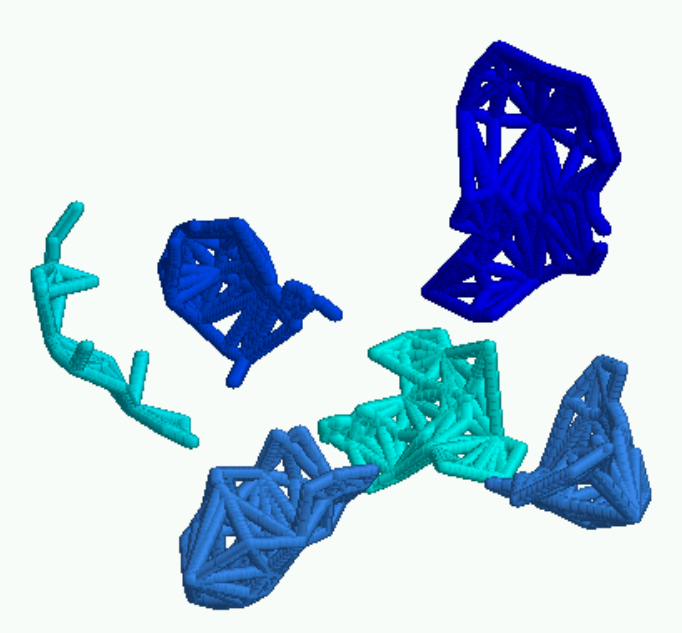
Examples of Patches for trypsin and trypsin inhibitor

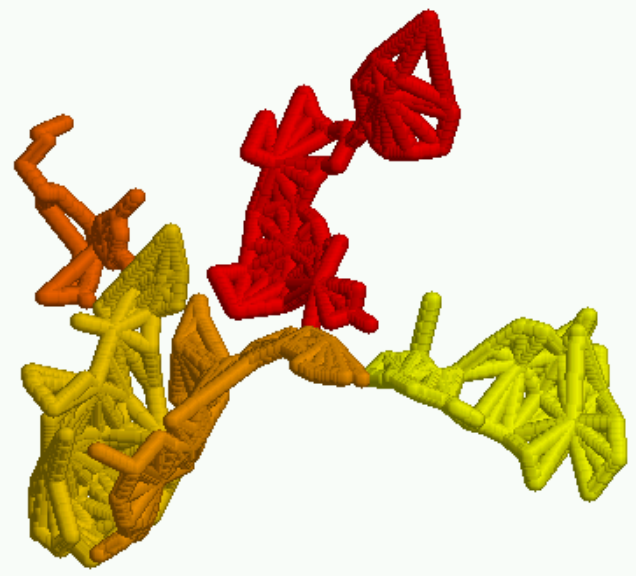




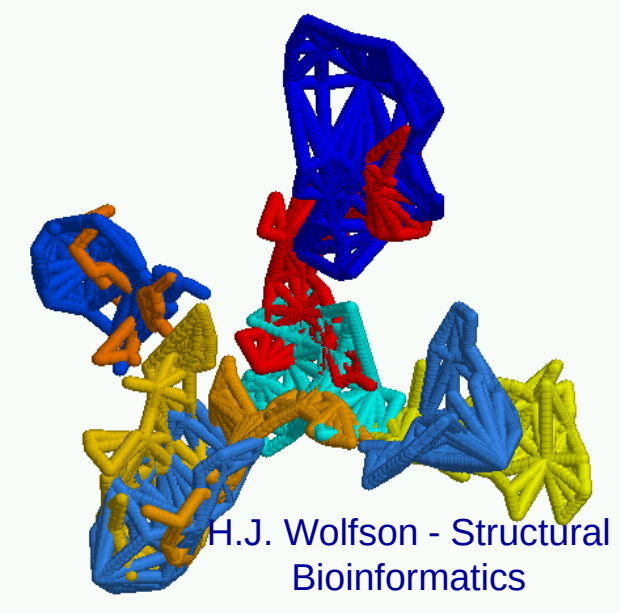
Yellow – knob patches, cyan – hole patches, green – flat patches, the proteins are in blue.

Complementarity of the Patches:



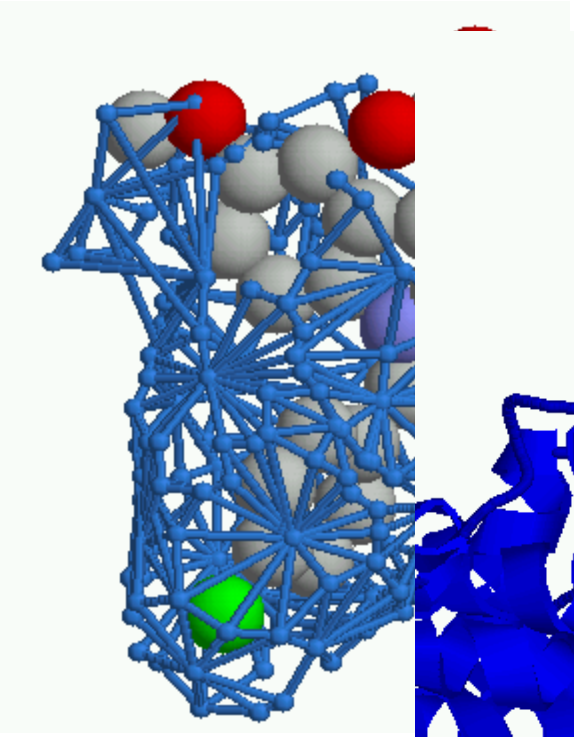


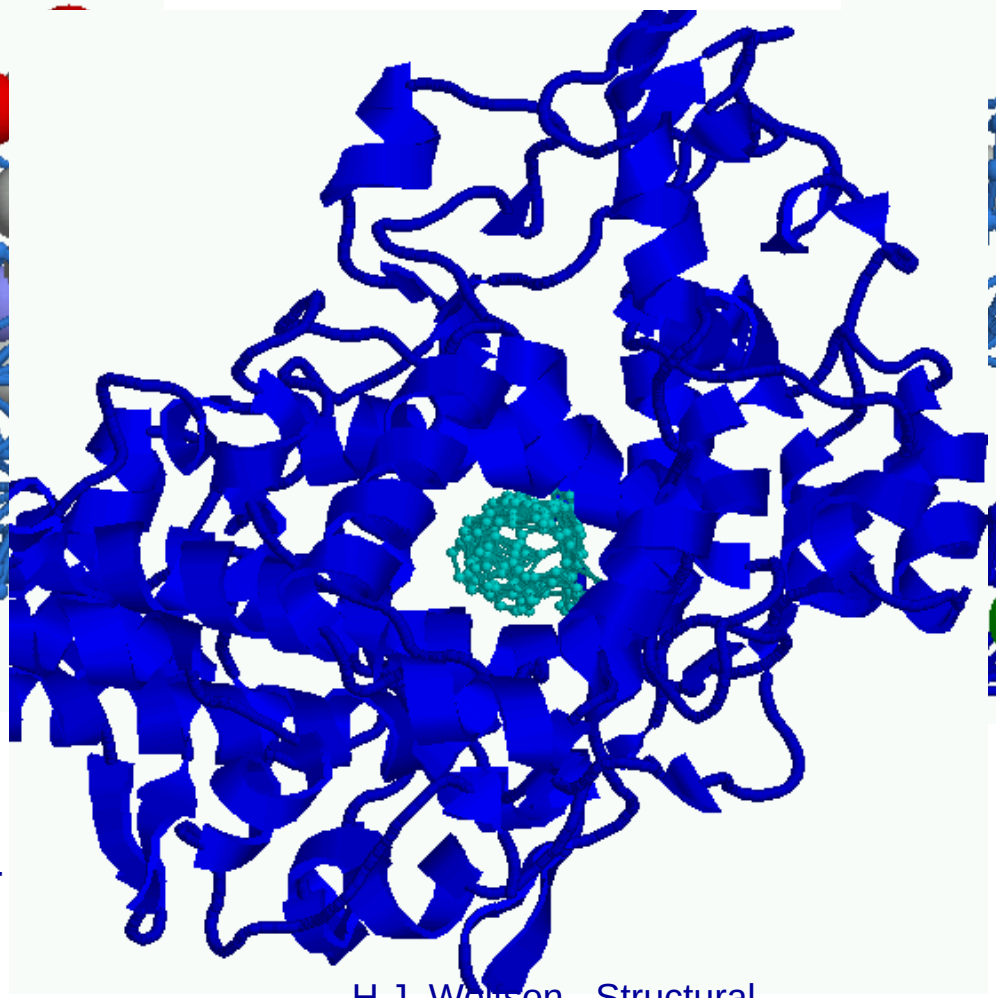
Interface knob patches of the ligand

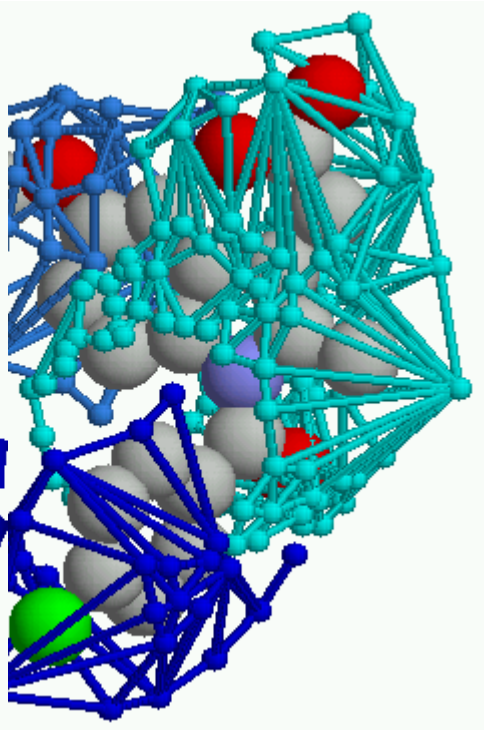


Interface hole patches of the receptor

Cox-2 cavity represented by a single hole patch:



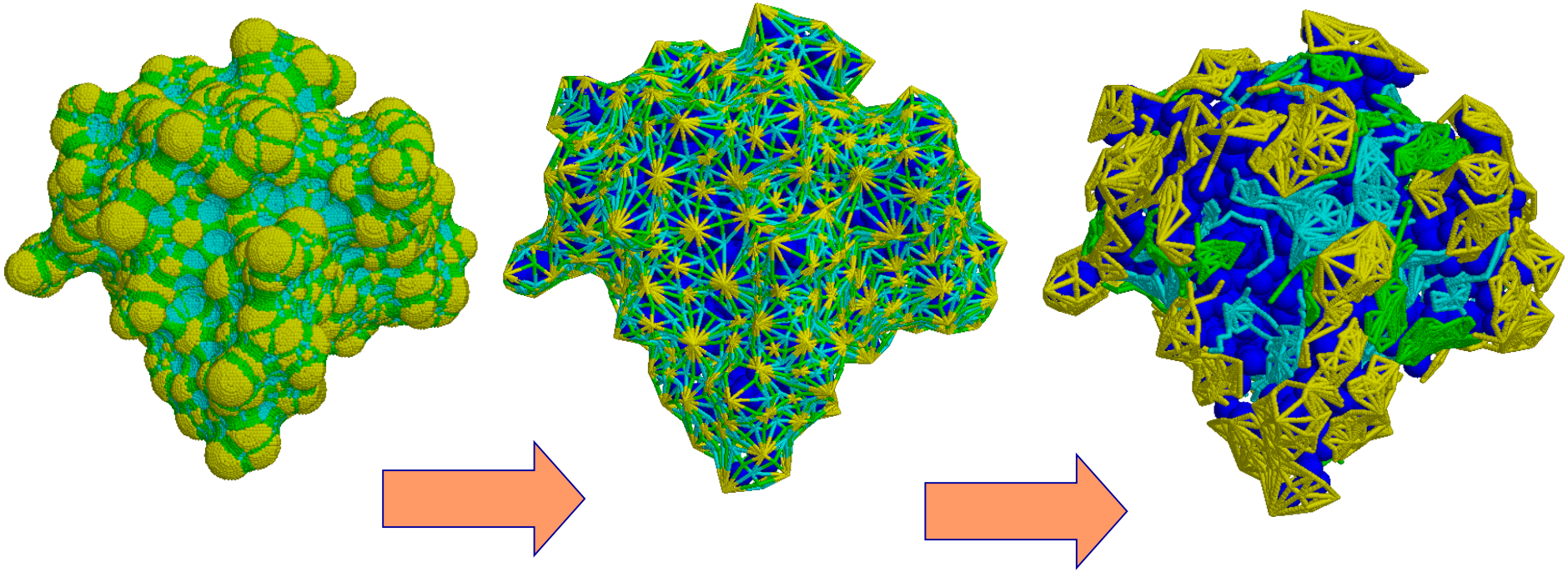




Indomethacin inside the COX-2 hole patch

Indomethacin inside its knob patches

Shape Representation Part



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.

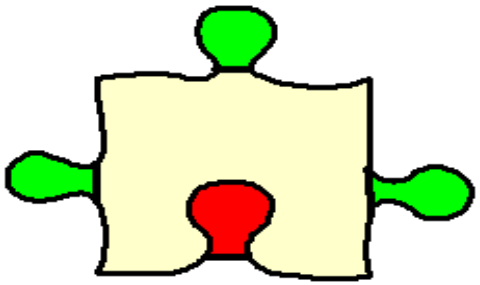
Enzyme/inhibitor –
Select patches with high enrichment of hot spot residues
(Ser, Gly, Asp and His for the enzyme;
Arg, Lys, Leu, Cys and Pro for the inhibitor).

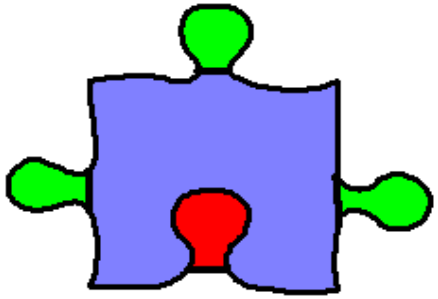
Antibody/antigen –
1. Detect CDRs of the antibody.
2. Select hot spot patches
(Tyr, Asp, Asn, Glu, Ser and Trp for antibody;
and Arg, Lys, Asn and Asp for antigen)

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

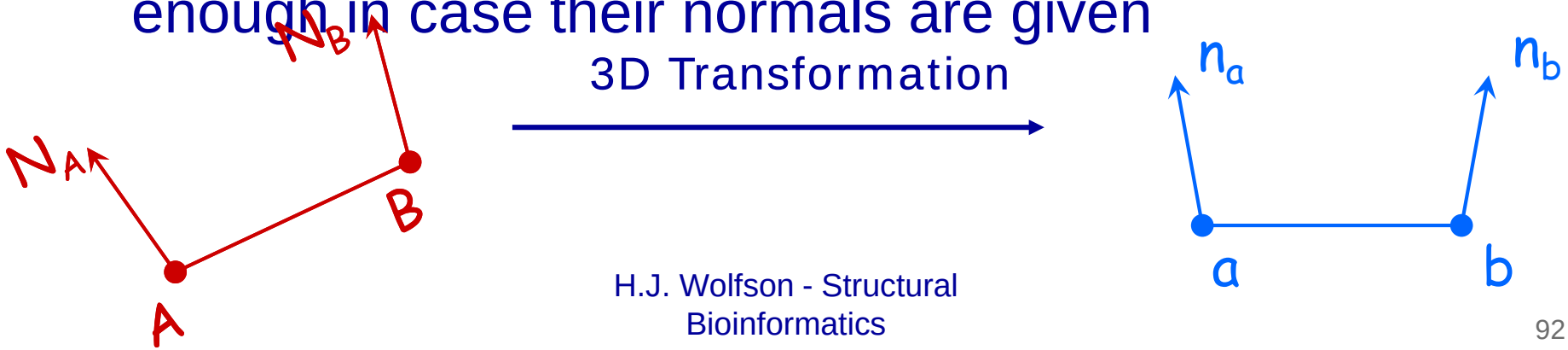
The aim is to align knob patches with hole patches, and flat patches with any patch. We use two types of matching:

- 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.

Creating Transformations in 3D Space

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

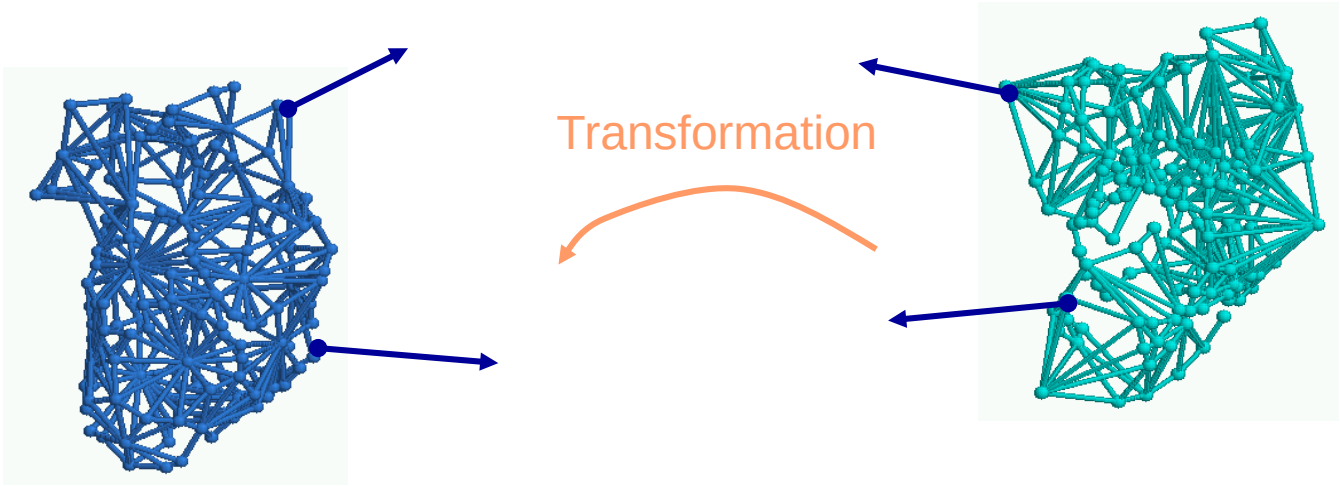
- A correspondence between a pair of 2 points is enough in case their normals are given



Single Patch Matching

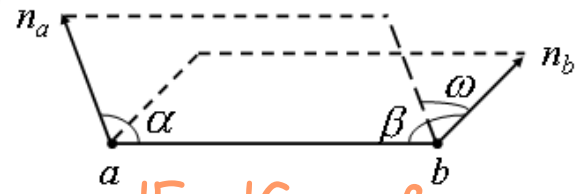
Receptor hole patch

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.

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

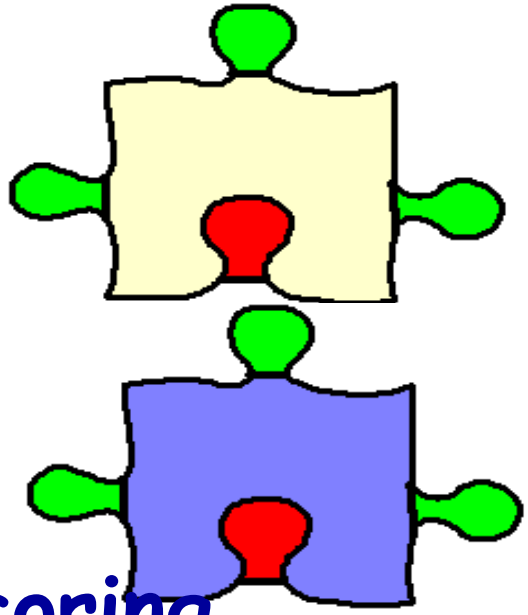
Two bases are compatible if their signatures match.

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.

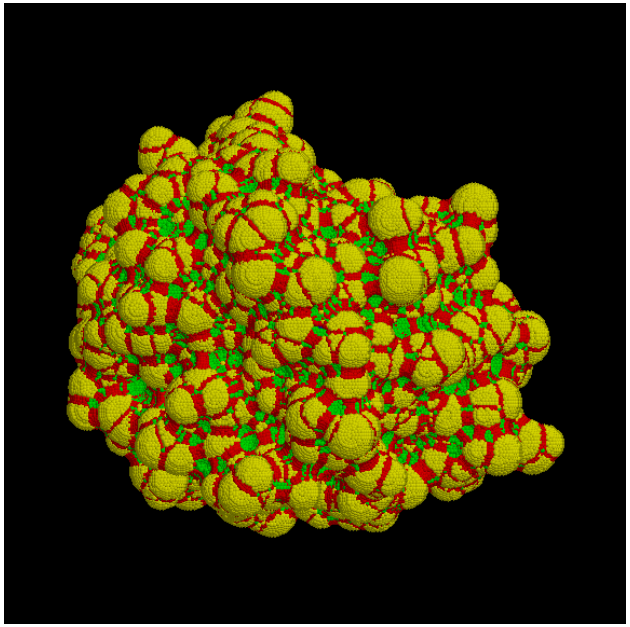
Docking Algorithm Scheme

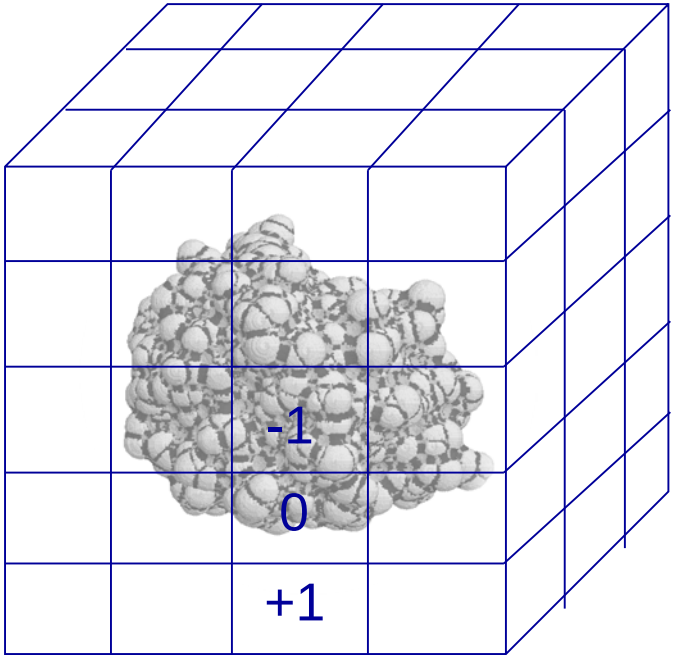
- Part 1: Molecular surface representation
- Part 2: Feature selection
- Part 3: Matching of critical features
- **Part 4: Filtering and scoring of candidate transformations**



Distance Transform Grid

Dense MS surface
(Connolly)





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
```

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

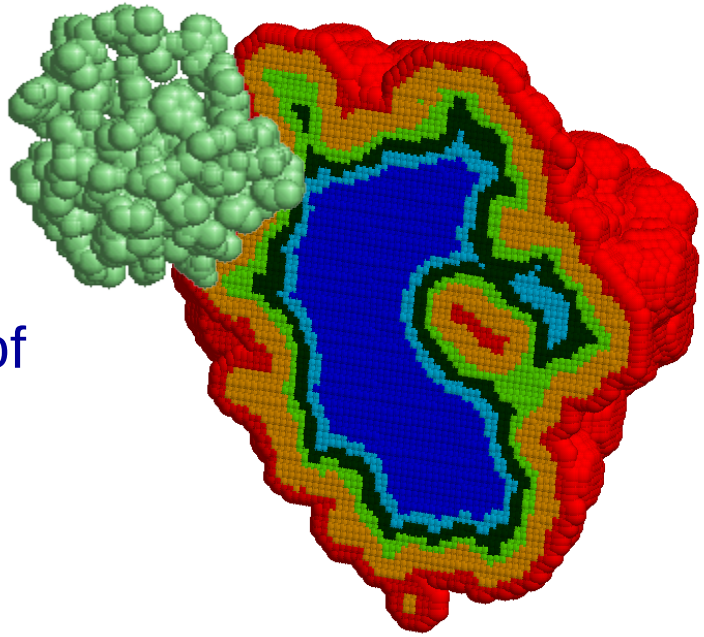
[-5.0,-3.6), [-3.6,-2.2), [-2.2, -1.0), [-1.0,1.0), [1.0→).

- 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:

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



Docking Algorithm Scheme

- Part 1: Molecular surface Representation
- Part 2: Features selection
- Part 3: Matching of critical features
- Part 4: Filtering and scoring of candidate transformations

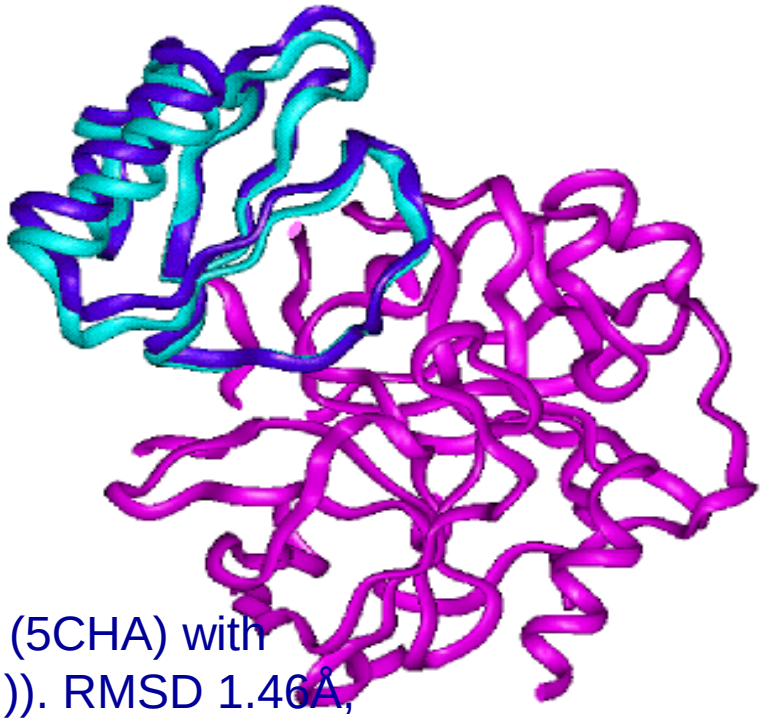
The correct solution is found in 90% of the cases with RMSD under 5A.

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

Refinement and Rescoring minimizing an Energy Function !

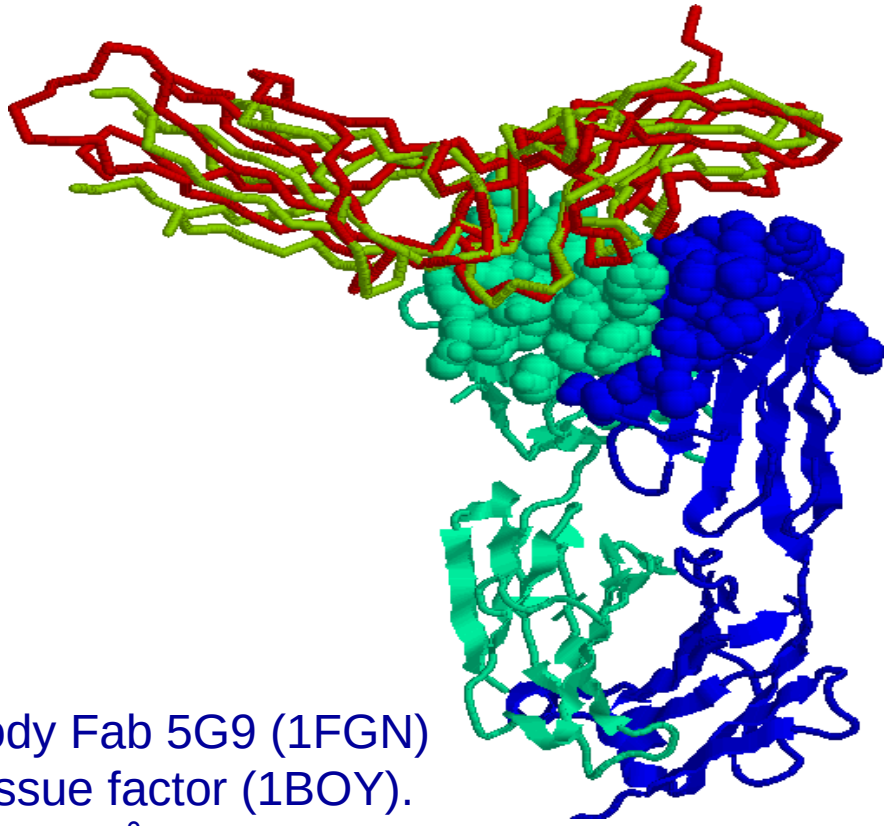
Example 1: Enzyme-inhibitor docking (unbound case)

- trypsin
- inhibitor from complex
- docking solution



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

Example 2: Antibody-antigen docking (unbound case)

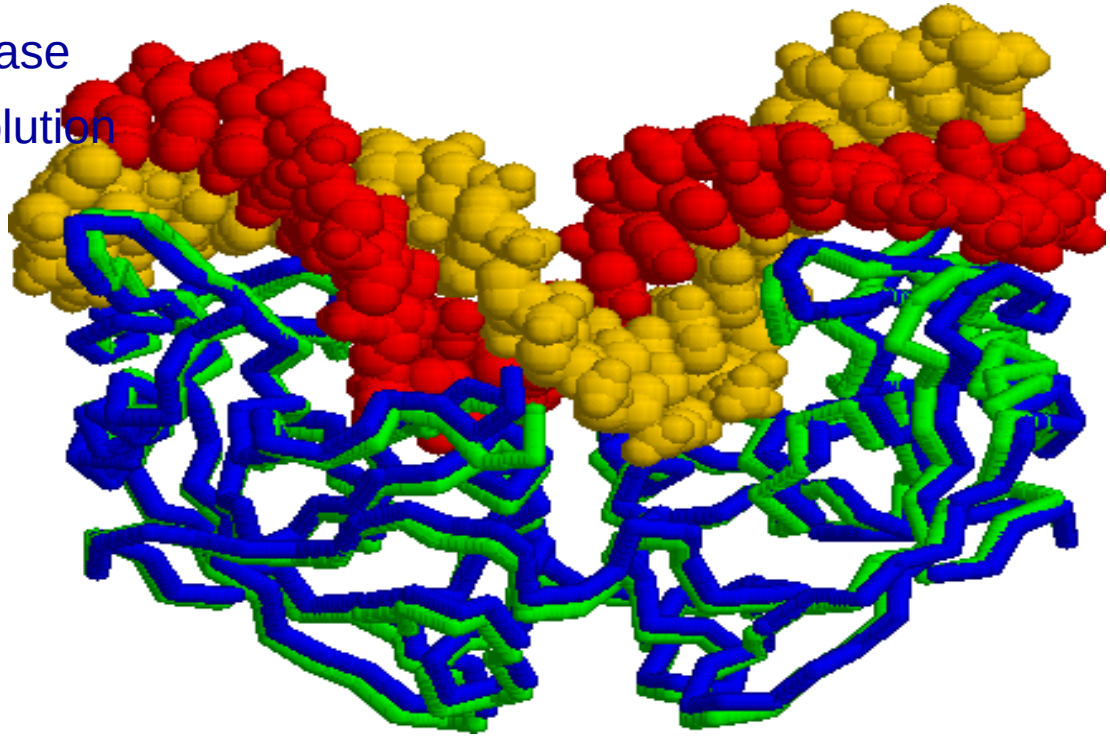


- antibody
- tissue factor from complex
- docking solution

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

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

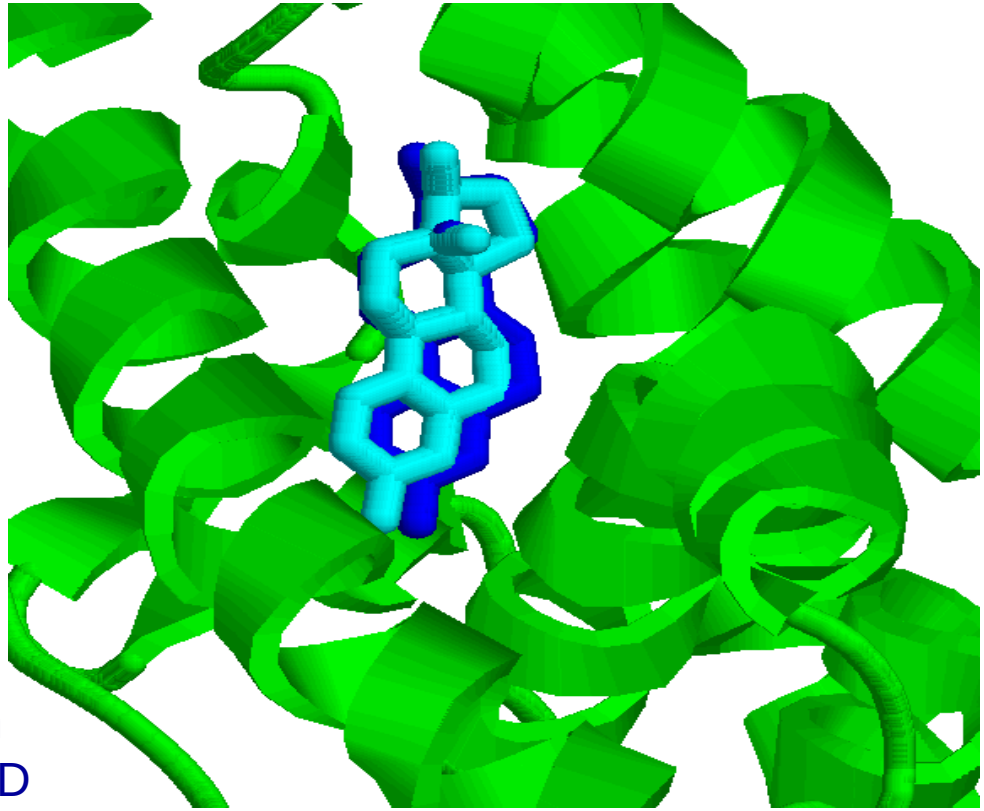
- DNA strand
- endonuclease
- docking solution



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

Example 4: Protein-drug docking (bound case)

- Estrogen receptor
- Estradiol from complex
- docking solution



Estrogen receptor with estradiol (1A52). RMSD 0.9Å, rank 1, running time: 11 seconds

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/