Protein Interactions Techniques and Challenges

Apostolos Axenopoulos
Electrical & Computer Engineer, MSc

Thessaloniki, October 2009

Outline

- **Protein Structure**
  - Proteins
  - amino acids
  - Protein Folding
    - Primary Structure
    - Secondary Structure
    - Tertiary Structure
    - Quaternary Structure
- **Protein Interactions**
  - Importance

Outline

- **Molecular Docking**
  - Definitions
  - Why docking is important?
  - Protein-protein docking / protein-ligand docking
  - Docking Aproaches
- **Binding Site Prediction**
- **Flexible Docking**

Proteins

- Proteins are usually large complex molecules, which have a fundamental role to cellular activity.
- They construct the cell skeleton.
- They demonstrate catalytic activity, accelerating biological reactions.
- Proteins consist of one or more polypeptides.
- A polypeptide is a single linear chain of amino acids (residues).
- amino acids are connected together with peptide bonds.

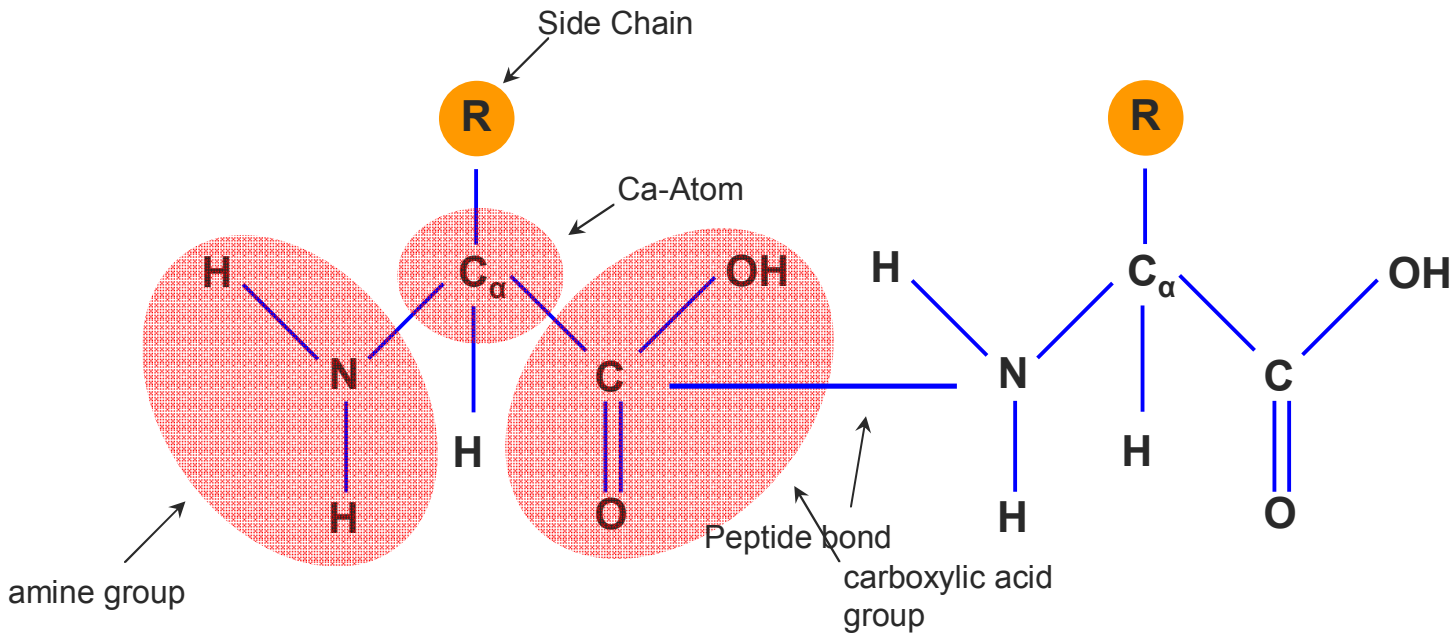
amino acids

- Proteins consist of 20 different amino acids.
- A protein sequence can be represented as a word, using an alphabet of 20 characters: Σ= {Ala, Arg, Asp, Asn, Cys, Glu, Gln, Gly, Hsi,Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val}.

Name	Code 1	Code 2
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Valine	Val	V
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Threonine	Thr	T
Tryptophan	Trp	W

Name	Code 1	Code 2
Isoleucine	Ile	I
Histidine	His	H
Cysteine	Cys	C
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Proline	Pro	P
Serine	Ser	S
Tyrosine	Tyr	Y
Phenylalanine	Phe	F

amino acids


Side Chain
R
Ca-Atom
H
Cα
OH
N
C
H
H
O
amine group
Peptide bond
carboxylic acid group
R
H
Cα
OH
N
C
H
H
O
amino acid 1
amino acid 2

Proteins Size

- Small: 3-10 amino acids
- Large: > 50000 amino acids
- Usual: 50 – 1000 amino acids

Protein Folding

- Protein folding is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure
- The biological activity of proteins is strongly related to their three-dimensional structure (i.e. protein folding).
- All information needed for a polypeptide to fold into its three-dimensional structure is encoded in the protein's amino acid sequence.

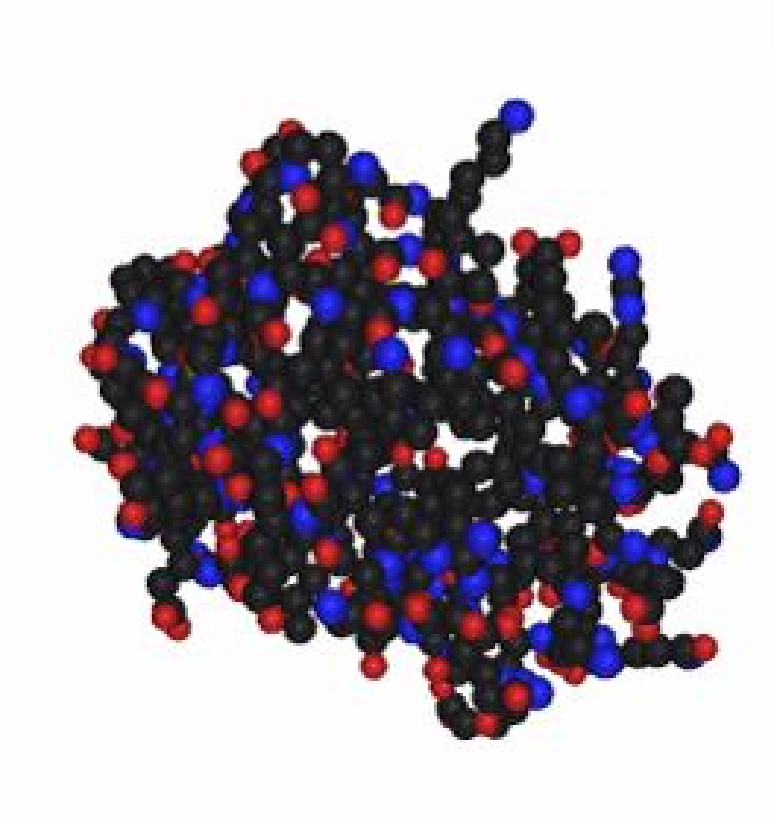
Protein Folding

- **Primary Structure:**
  amino acid sequence

```
F N V S G T V C L S A L P P E A T N T L N L I A S N G P F P Y S Q N G
V V F Q N R E S V L P T Q S Y G Y Y H E Y T V I T P G A R T R G T R R
I I T G E A T Q E S P Y Y T G N H Y A T F S L I N Q T C
```

- **Secondary Structure:**
  Folding of amino acid sequence into a-helices and/or b-sheets

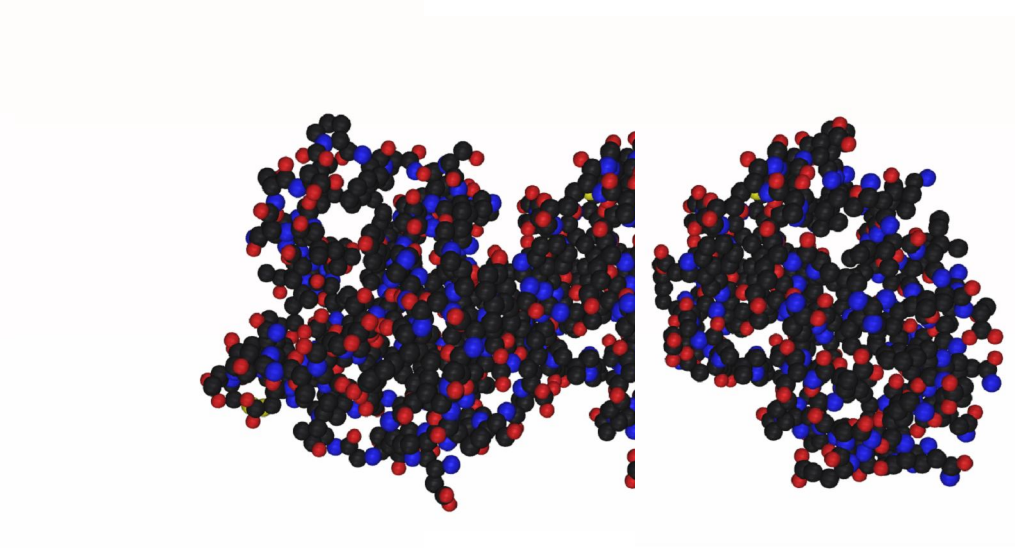
Protein Folding



- **Tertiary Structure:** Three-dimensional representation of a polypeptide chain.

Protein Folding

- **Quaternary Structure:** 3D representation of a complex protein (two or more polypeptide chains).



PDB File

```
HEADER     IMMUNOGLOBULIN                                 03-MAR-97   2PSK
TITLE      THEORETICAL MODEL OF AN FAB FRAGMENT COMPLEXED WITH THE
TITLE     2 MELANOMA-ASSOCIATED GD2 GANGLIOSIDE
COMPND     MOL_ID: 1;
COMPND    2 MOLECULE: ANTIBODY;
COMPND    3 CHAIN: L, H;
COMPND    4 FRAGMENT: FAB
...
AUTHOR     S.L.PICHLA,R.MURALI,R.M.BURNETT
REVDAT    1   04-SEP-97 2PSK     0
...
JRNL          AUTH    S.L.PICHLA,R.MURALI,R.M.BURNETT
SEQRES    1 L   213   GLN ILE VAL LEU THR GLN SER PRO ALA ILE MET SER ALA
SEQRES    2 L   213   SER PRO GLY GLU LYS VAL THR ILE THR CYS SER ALA SER
SEQRES    3 L   213   SER SER VAL SER ASN ILE HIS TRP PHE GLN GLN LYS PRO
...
HELIX     1    1 SER L   121   SER L   126   1                                   6
HELIX     2    2 LYS L   182   TYR L   185   1                                   4
...
SHEET     1    A 4 LEU L     4   SER L     7   0
SHEET     2    A 4 VAL L    19   ALA L    25  -1   N   SER L   24    O   THR L    5
SHEET     3    A 4 SER L    69   ILE L    74  -1   N   ILE L   74    O   VAL L   19
...
ATOM   1   N    GLN L    1   40.444    0.114   53.530   1.00 44.06           L     N
ATOM   2   CA   GLN L    1   39.136   -0.460   53.239   1.00 38.84           L     C
ATOM   3   C    GLN L    1   39.210   -1.920   52.815   1.00 33.95           L     C
ATOM   4   O    GLN L    1   39.943   -2.274   51.886   1.00 34.91           L     O

HETATM 3854  O  HOH   1      -1.229   -1.762    5.590   1.00 15.50           W     O
HETATM 3855  O  HOH   3      23.399  -21.858   56.848   1.00 10.79           W     O
HETATM 3856  O  HOH   4       6.748   17.422   37.138   1.00 28.29           W     O
...
CONECT 1815 1196 1814
CONECT 2209 2208 2944
CONECT 2944 2209 2943
...
END
```

PDB File

- HEADER: classification, date entered the database
- TITLE, COMPOUND, SOURCE, KEYWDS, EXPDTA, AUTHOR, JRNL, REMARK
- SEQRES: sequence of aminoacids
- HELIX: all α-helices included in the chain
- SHEET: all β-sheets included in the chain
- ATOM: coordinates of all atoms (X,Y,Z)

- Protein Data Bank (PDB) (http://www.rcsb.org/)
  - More than 54000 structures

Protein Interactions

Thessaloniki, October 2009

Protein Interactions

- Proteins interact in order to perform their functions
- The patterns of such interactions can be very informative about the functional organisation of the cell
- Knowledge about where (and how) a protein binds to another protein or other molecules gives us a better understanding of its biological function.

Protein Interactions

- Many important applications:
  - drug design
  - protein mimetics engineering
  - elucidation of molecular pathways
  - understanding of disease mechanisms

Computational Biology

- Experimental determination of protein structures, protein-protein complexes, is a highly time-consuming task.
- Computational Biology applies the techniques of computer science, applied mathematics and statistics to address biological problems.
- Computational Biology provides a faster solution to identify protein-protein interaction sites, in particular, identify surface residues that are associated with protein-protein interaction.
  - Experimental validation is still necessary!

Protein Interactions

Protein interaction problems that involve computational biology

- **Binding Site Prediction**: to identify which parts of the protein structure are likely to participate in protein interactions (binding sites)
- **Molecular Docking**: given two protein structures, determine:
  - Whether the two molecules “interact”
  - If so, what is the 3D structure of the resulting complex

Protein Interactions

- **Docking** usually refers to computation of the geometric complementarity between the surfaces of the interacting proteins
- **Binding Site Prediction** computes the probability of a surface patch to interact, based on:
  - Electrostatic potential
  - Hydrophobicity
  - residue interface propensity
  - salvation potential
- Efficient binding site prediction can improve protein docking

Why is docking important?

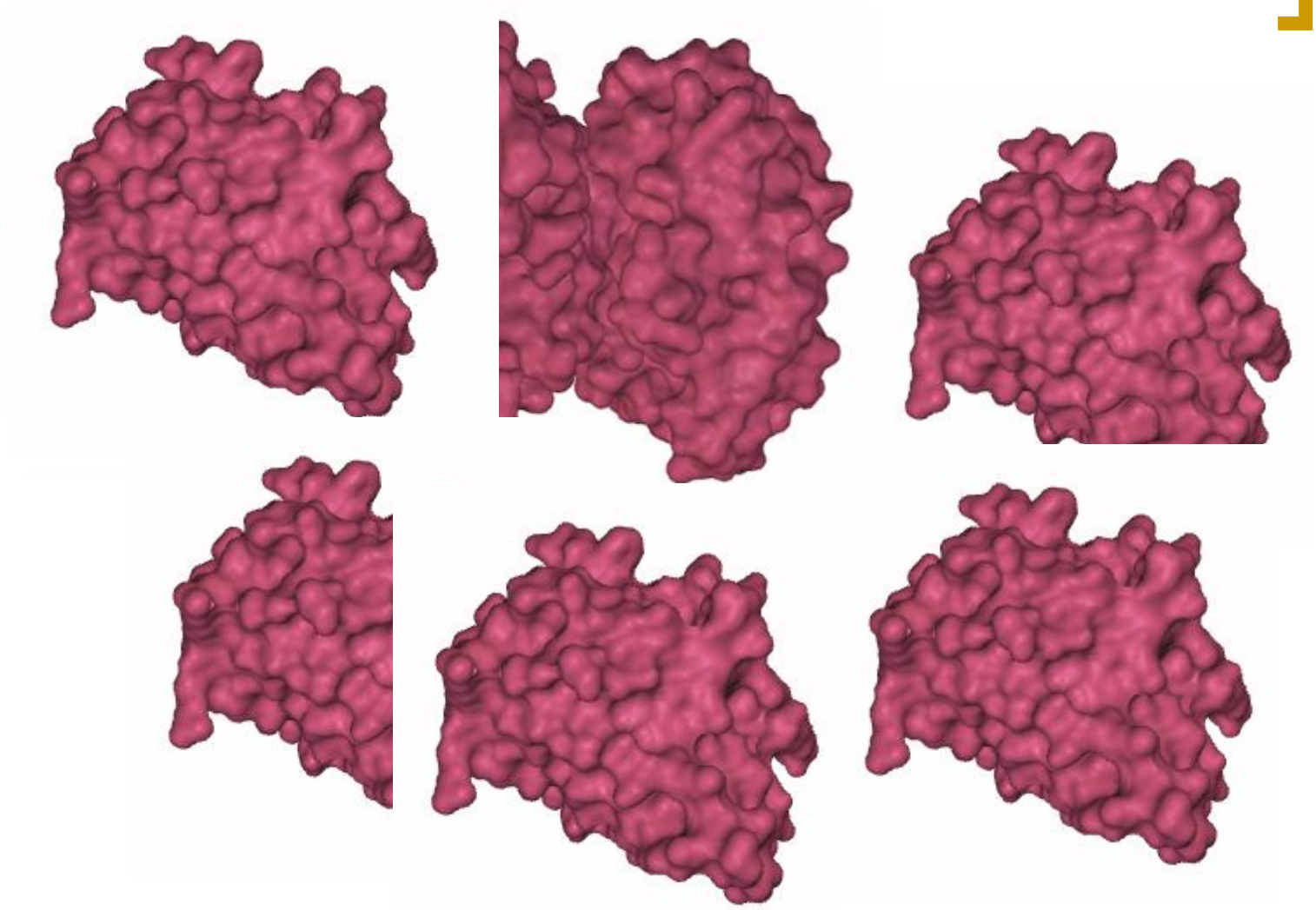
- It is of extreme relevance in **cellular biology**, where function is accomplished by proteins interacting with themselves and with other molecular components

- It is the key to rational **drug design**: The results of docking can be used to find inhibitors for specific target proteins and thus to design new drugs. It is gaining importance as the number of proteins whose structure is known increases

Types of Docking studies

- Protein-Protein Docking
  - Molecules have approximately the same size
  - Both molecules usually considered rigid
  - 6 degrees of freedom
  - First apply steric constraints to limit search space and the examine energetics of possible binding conformations
- Receptor-Ligand Docking
  - The ligand is usually a small molecule comparing with the receptor
  - Flexible ligand, rigid-receptor
  - Search space much larger
  - Either reduce flexible ligand to rigid fragments connected by one or several hinges, or search the conformational space using monte-carlo methods or molecular dynamics

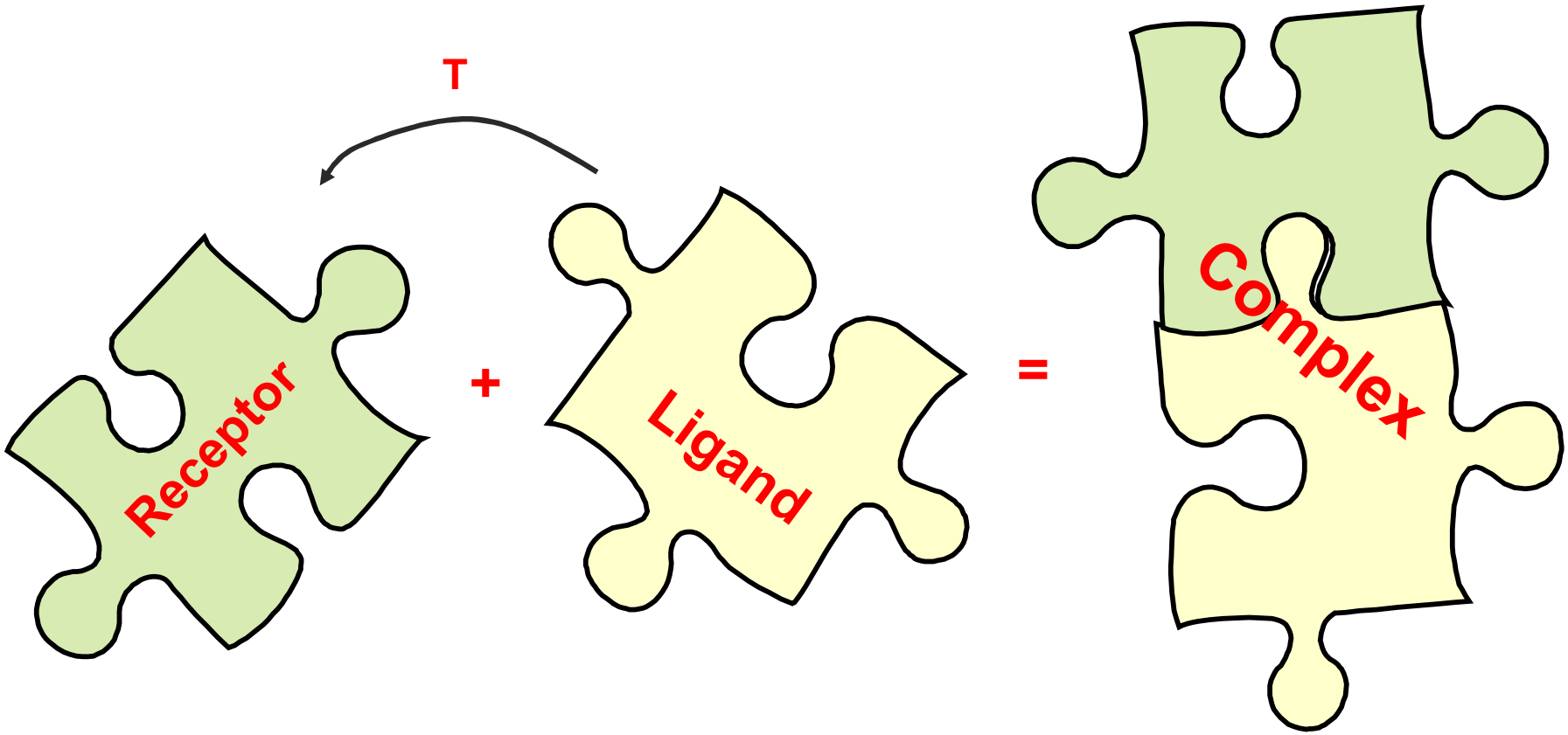
Protein Docking Problem



Protein Docking Problem

- Protei-Protein docking is based on the following principles:
  - "**Geometric Complementarity**": the binding sites of the two interacting molecules have complementary shapes
  - "**Biochemical Complementarity**": it has been proven that several non-geometric factors (hydrogen bonds, electrostatic potential, hydrophobocity, residue interface propensity) can affect the interaction of two molecules.

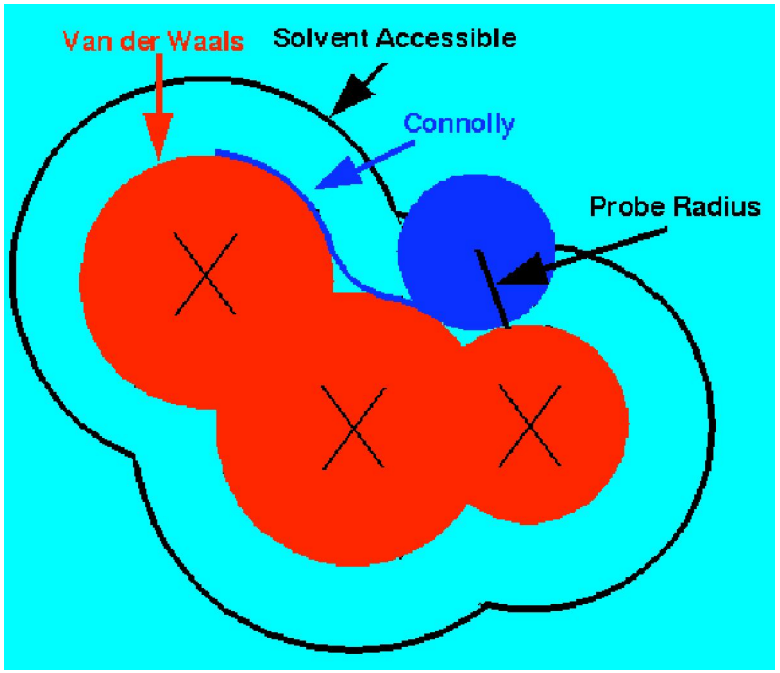
Geometric Complementarity



Docking Scheme

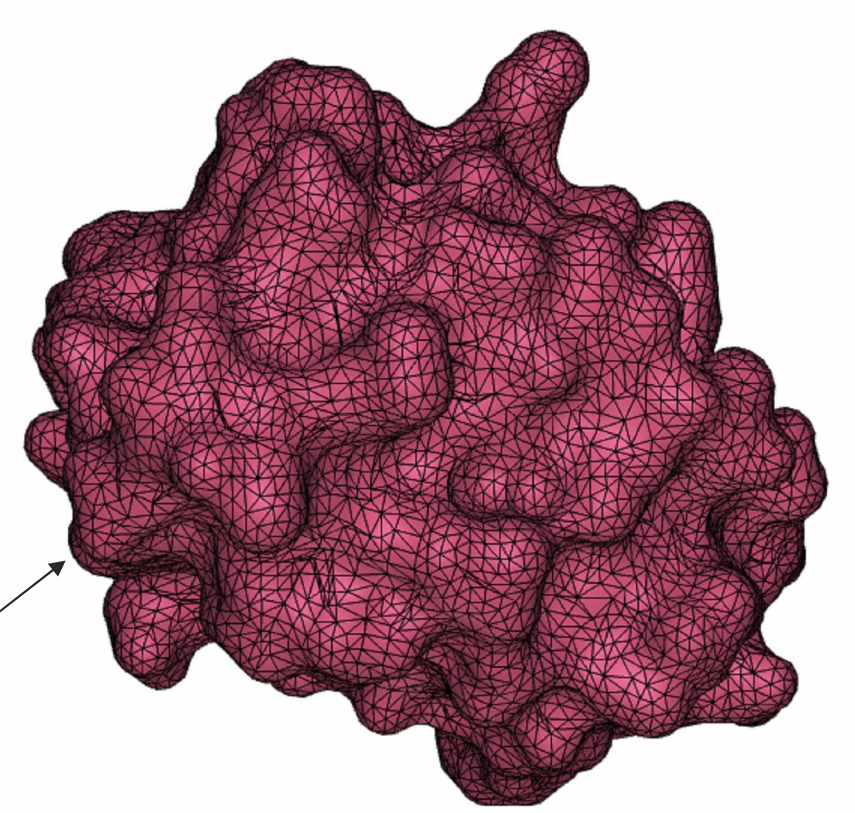
- **Step 1: Surface representation**
  - **Dense MS surface Connolly surface**
  - Sparse surface (Shuo Lin et al.)
  - Lenhoff technique
  - Kuntz et al. Clustered-Spheres
  - Alpha shapes
- **Step 2: Identify the regions of interest**
  - cavities and protrusions
  - Surface patches of specific size
- **Step 3: Matching of critical features and compute transformation**
  - Geometric Hashing
  - Context Shapes
- **Step 4: Scoring of candidate transformations**
  - Distance transform grid

Surface Representation



- Rolling a Probe Sphere over the molecule
- Everywhere the center of the sphere goes is the **Solvent Accessible Surface** (SAS)
- Everywhere the sphere touches (including empty space) is the **Solvent Excluded (or "Connolly") Surface** (SES)

Dense MS Surface (Connolly)



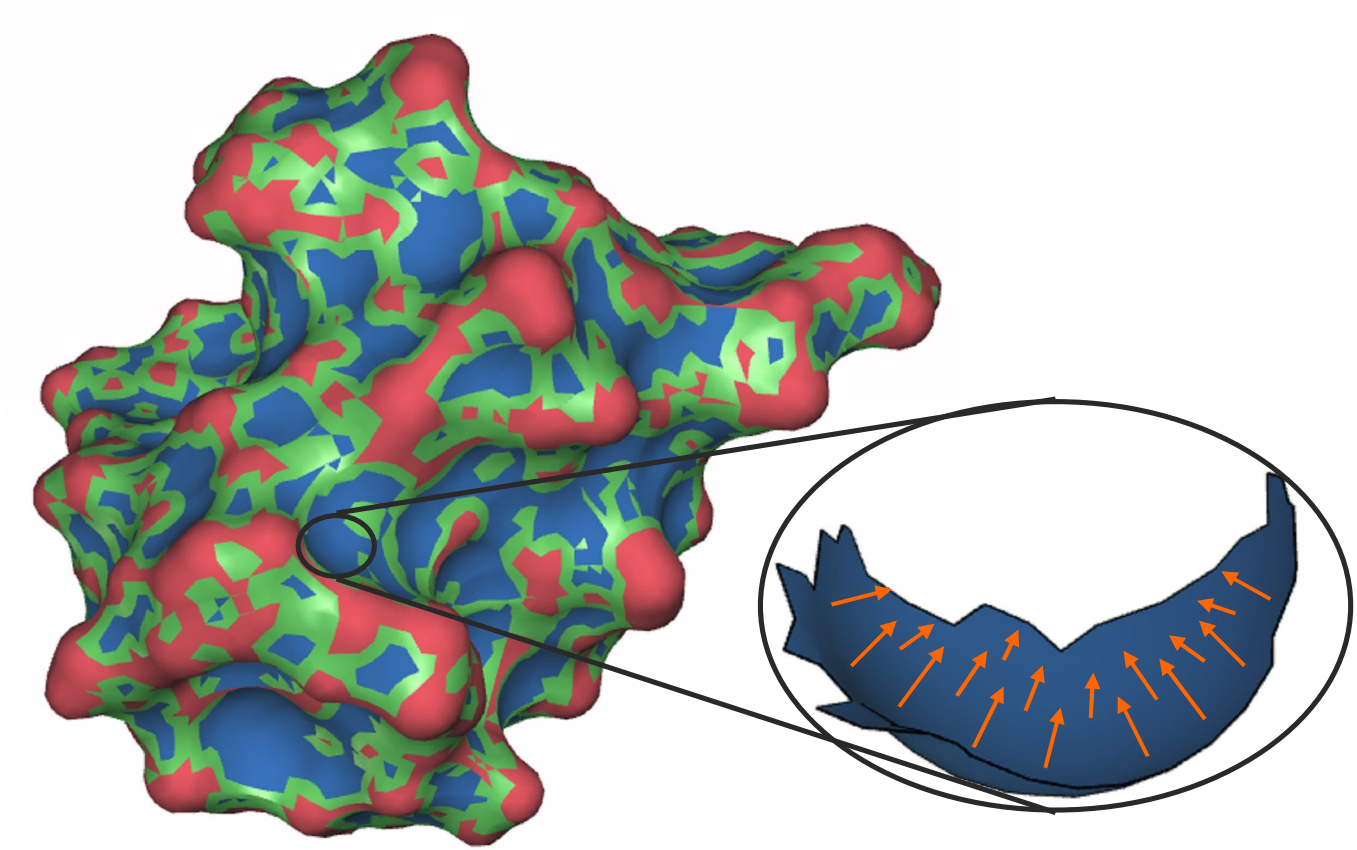
11244 points

22488 triangles

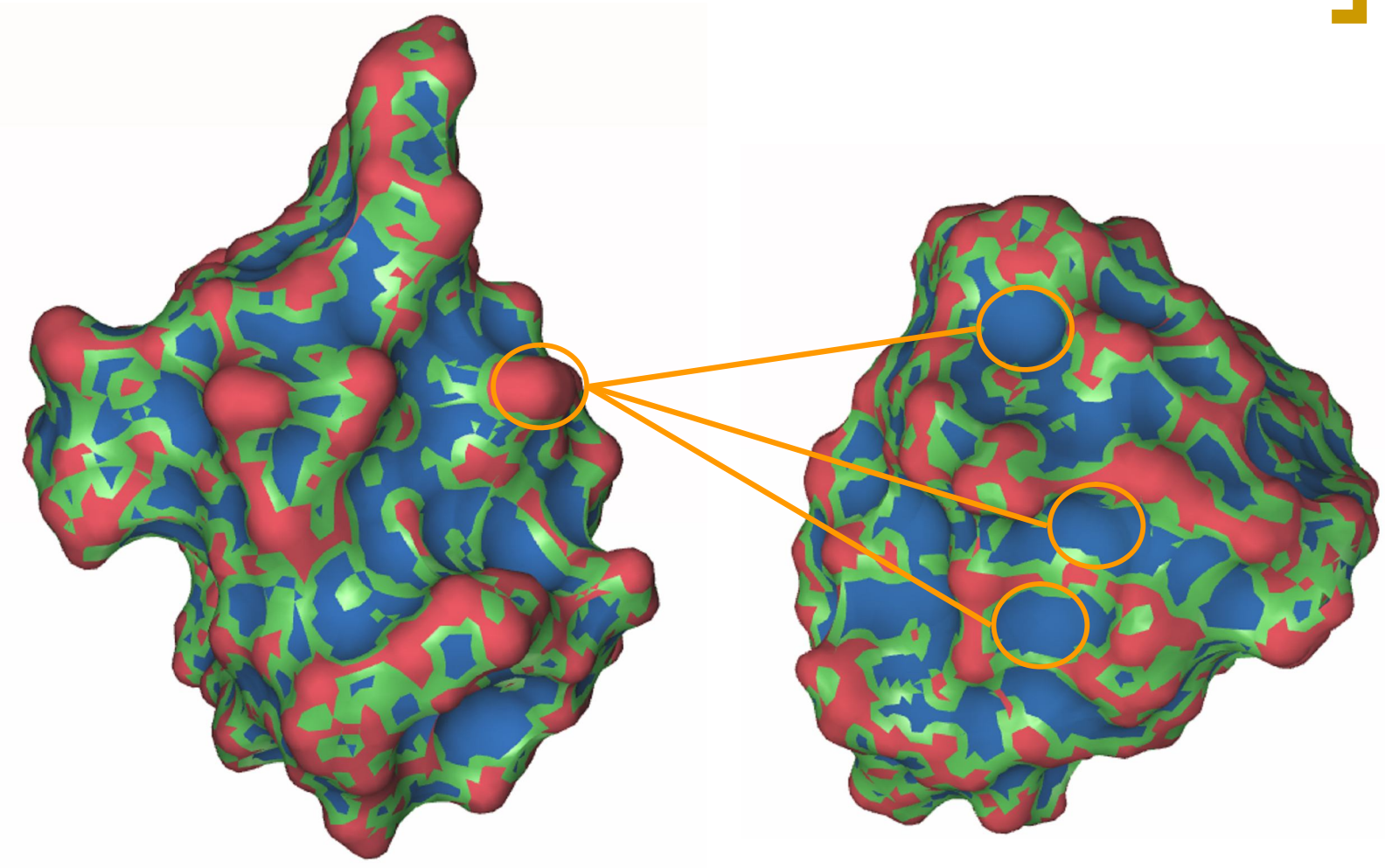
Docking Scheme

- Step 1: Surface representation
  - Dense MS surface Connolly surface
  - Sparse surface (Shuo Lin et al.)
  - Lenhoff technique
  - Kuntz et al. Clustered-Spheres
  - Alpha shapes
- **Step 2: Identify the regions of interest**
  - cavities and protrusions
  - Surface patches of specific size
- Step 3: Matching of critical features and compute transformation
  - Geometric Hashing
  - Context Shapes
- Step 4: Scoring of candidate transformations
  - Distance transform grid

Docking Scheme



Docking Scheme

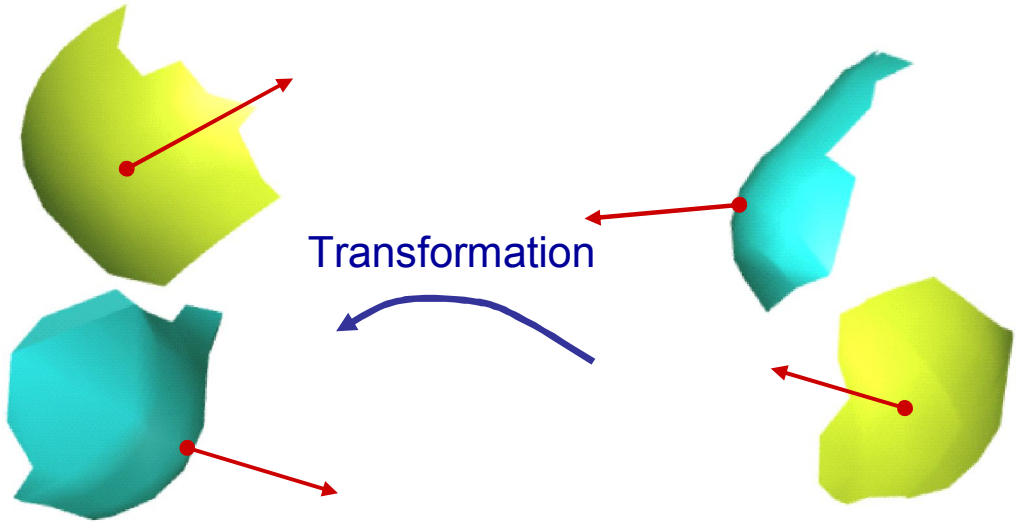


Docking Scheme

- Step 1: Surface representation
  - Dense MS surface Connolly surface
  - Sparse surface (Shuo Lin et al.)
  - Lenhoff technique
  - Kuntz et al. Clustered-Spheres
  - Alpha shapes
- Step 2: Identify the regions of interest
  - cavities and protrusions
  - Surface patches of specific size
- **Step 3: Matching of critical features and compute transformation**
  - Geometric Hashing
  - Context Shapes
- Step 4: Scoring of candidate transformations
  - Distance transform grid

PatchDock* Method

- **Base**: 1 critical point with its normal from one patch and 1 critical point with its normal from a neighboring patch.

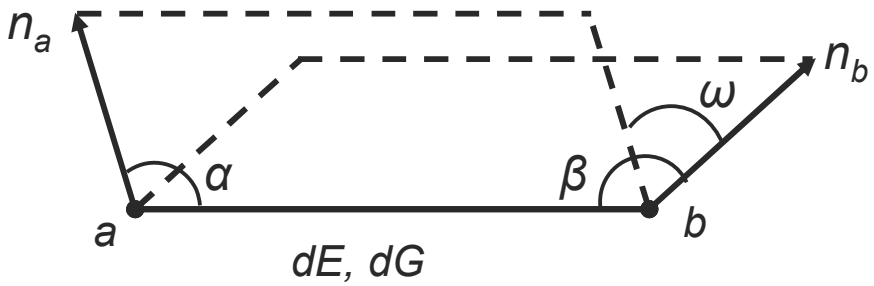


- Match every base from the receptor patches with all the bases from complementary ligand patches.
- Compute the transformation for each pair of matched bases.

* D. Duhovny, R. Nussinov, and H. J. Wolfson. Efficient unbound docking of rigid molecules. In 2'nd Workshop on Algorithms in Bioinformatics, pages 185–200, 2002.

PatchDock - Base Signature

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



Two bases match when their signatures match

PatchDock – Geometric Hashing

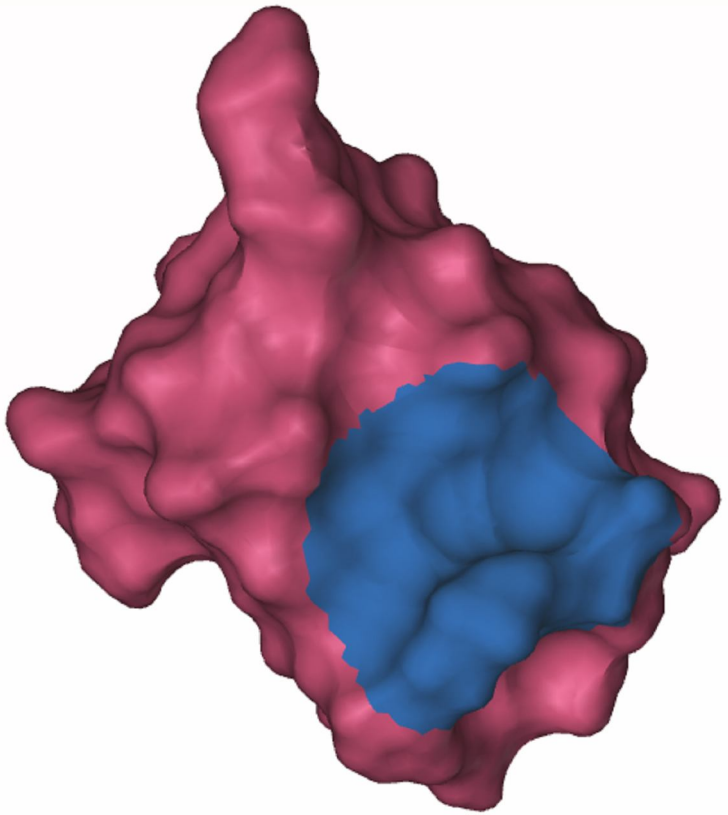
- **Preprocessing**: for all patch pairs of the ligand, compute the bases and store them to the hash table according to base signature.
- **Recognition**: for each patch pair of the receptor, compute base signature and access hash table. The transformations set is computed for all compatible bases.

Results:

- Small patches (convex or concave) do not carry quite significant shape information
- A large number of compatible bases (and transformations) may occur

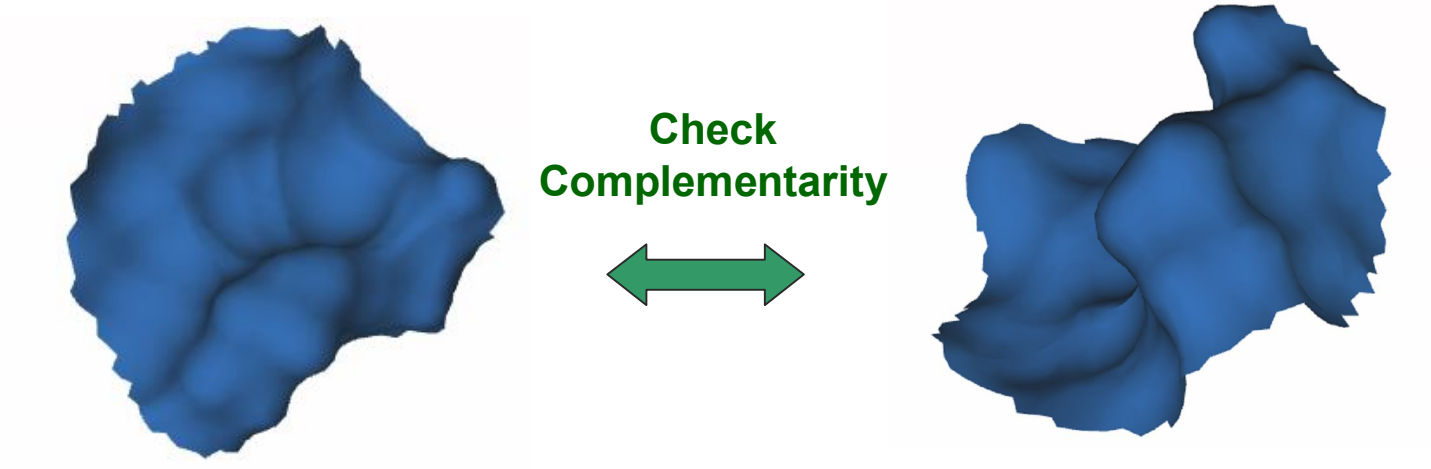
Context Shapes* Approach

- Generate an initial set of *Critical Surface Points* (centers of convex/ concave patches)
- For each CSP extract a wider surface patch (radius of sphere ~ 10 Å)



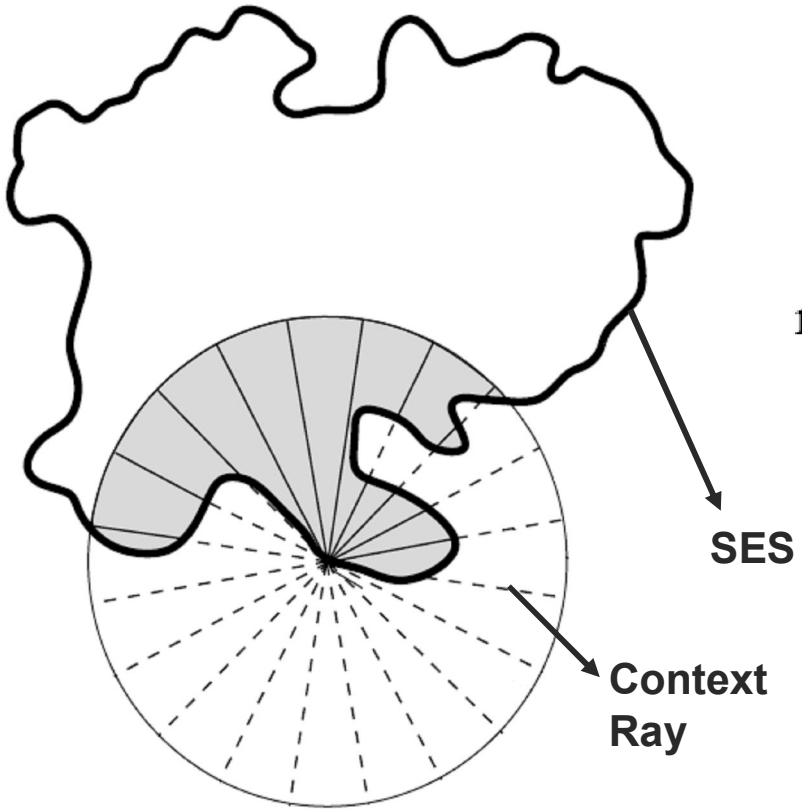
* SHENTU Z., AL HASAN M., BYSTROFF C., ZAKI M.: Context shapes: Efficient complementary shape matching for protein-protein docking. Proteins: Structure, Function and Bioinformatics (2007).

Context Shapes Approach



- Match the wide patches of the receptor with the (equally-sized) patches of ligand in terms of complementarity
- Now the patches enclose more significant shape information (local)

Context Shapes Approach



Context Ray

11111111111111111000000000000000 Binary String

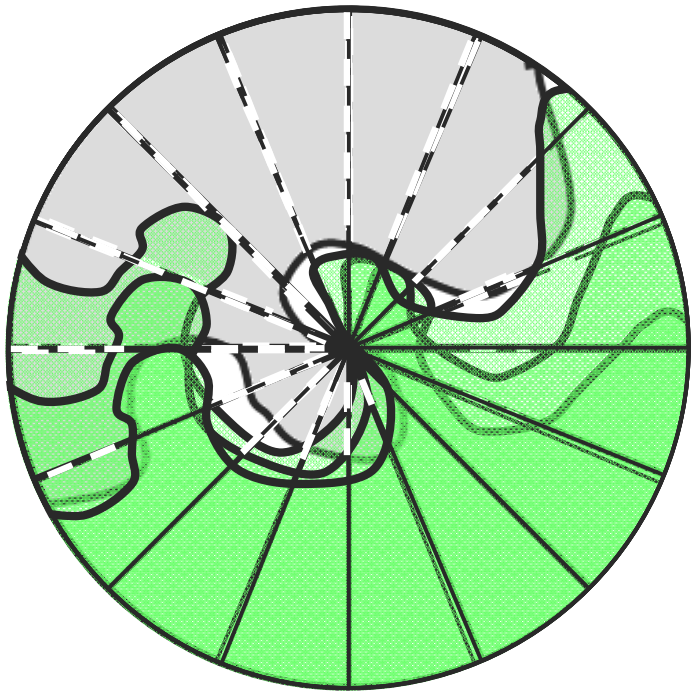
Complementarity Matching of 2 CRs:

CR^R **XOR** CR^L

Context Shapes Approach

Score of a pose π

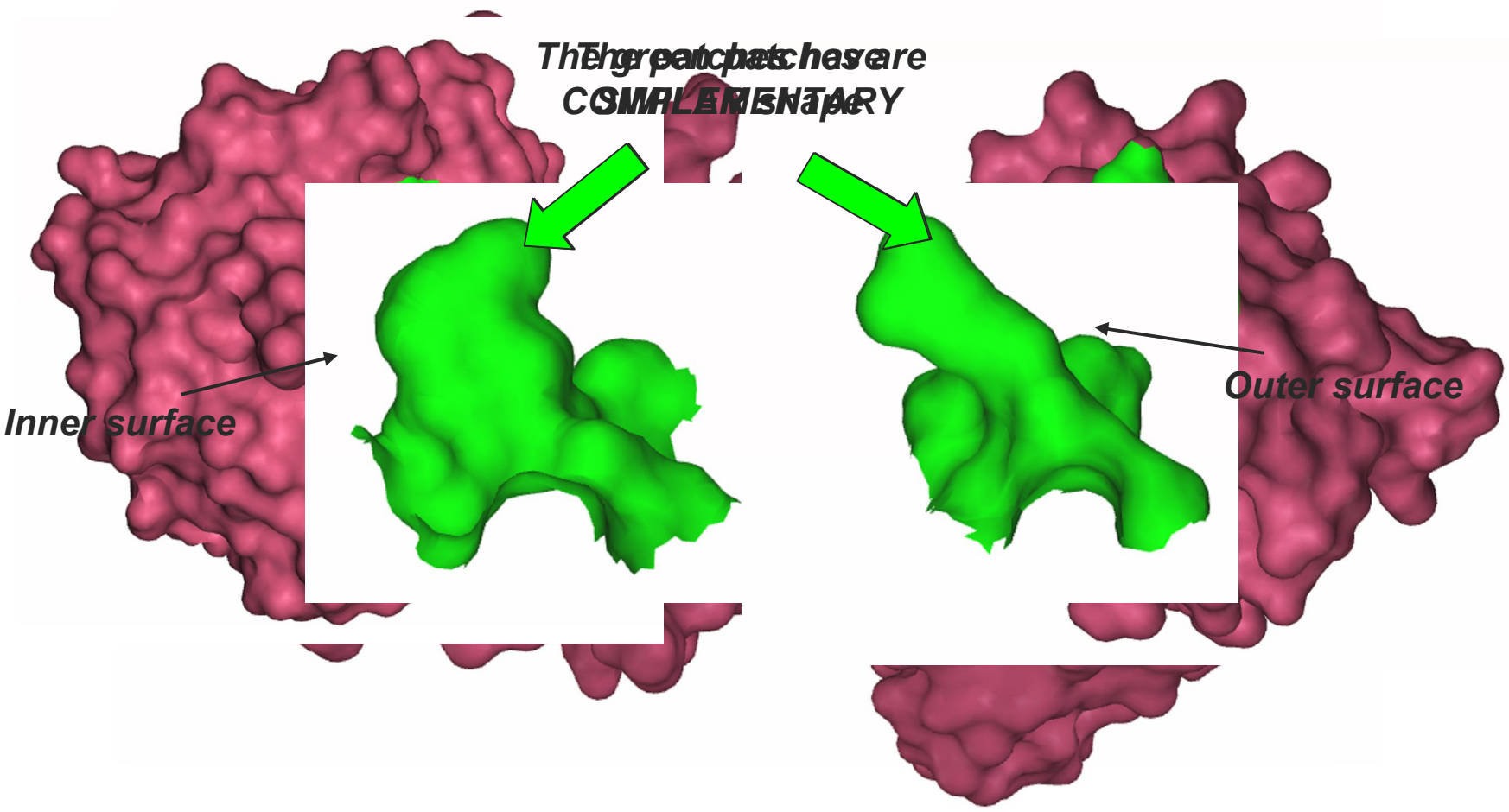
- One-to-one maching of the context rays between the two context shapes



Results:

- Achieves better results than PatchDock
- The ligand patch is rotated several times until the best match is found

Similarity vs Complementarity

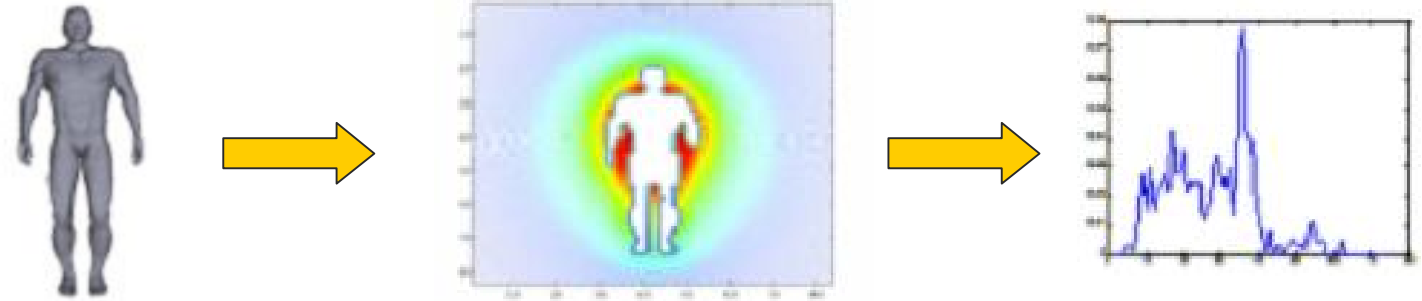


3D Shape Similarity

- Use 3D shape similarity matching approaches to identify the pairs of complementary patches
- The methods should be invariant to rotation
- The methods should be applied to the surface of 3D objects

Shape Impact Descriptor (SID)

- The key idea of Impact Descriptor* is the indirect description of the 3D object's geometry, by computing features that describe the impact of the 3D object in the surrounding space.
- Every object is treated as a distributed mass and the gravitational impact is described



3D Object **Field Computation** **Shape Impact Descriptor**

* A.Mademlis, P.Daras, D.Tzovaras, and M.G.Strintzis, "3D Object Retrieval using the 3D Shape Impact Descriptor" ELSEVIER, Pattern Recognition, Volume 42 , Issue 11, pp. 2447-2459, Nov 2009

SID Features

- Every 3D shape (patch) is described by a descriptor vector
- Similarity matching is performed by histogram matching
- Rotation Invariant

* A.Mademlis, P.Daras, D.Tzovaras, and M.G.Strintzis, "3D Object Retrieval using the 3D Shape Impact Descriptor" ELSEVIER, Pattern Recognition, Volume 42 , Issue 11, pp. 2447-2459, Nov 2009

SID Performance

- The method was tested in Docking Benchmark v2.0 (contains 84 test complexes)
- In all complexes, at least one correct match was found in the first places of the ranked patch pairs.

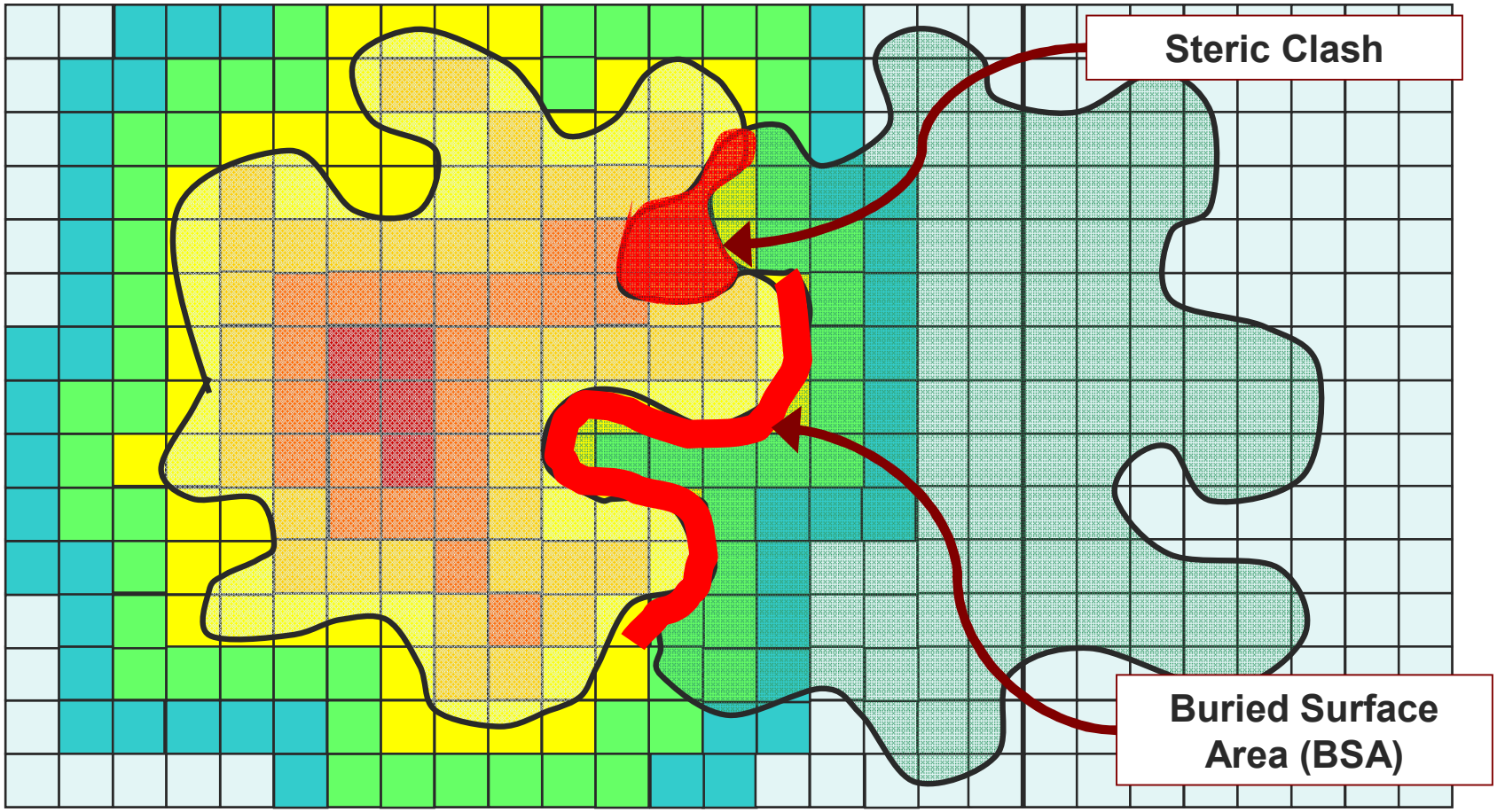
Results:

- More than 98% of the total pairs of patches (in Step 3) can be discarded using a fast method (without losing the correct solution).
- A solution to align the correct matches (compute translation/rotation) is needed.

Docking Scheme

- Step 1: Surface representation
  - Dense MS surface Connolly surface
  - Sparse surface (Shuo Lin et al.)
  - Lenhoff technique
  - Kuntz et al. Clustered-Spheres
  - Alpha shapes
- Step 2: Identify the regions of interest
  - cavities and protrusions
  - Surface patches of specific size
- Step 3: Matching of critical features and compute transformation
  - Geometric Hashing
  - Context Shapes
- **Step 4: Scoring of candidate transformations**
  - **Distance transform grid**

Scoring



Scoring

$$\textbf{\textit{Score} = } w_1 N_{BSA} - w_2 N_{Clash}$$

- N_{BSA}: *number of surface points of the ligand within Buried Surface Area*
- N_{clash}: *number of surface points that penetrate the receptor's surface*
- w_1, w_2: *appropriately selected weights*
- Finally, we keep the pose (or poses) with the highest score

Geometric Complementarity

Is Geometric Complementarity enough?

- Geometric algorithms usually propose more than one potential complexes
- Only one is the correct solution
- There are a lot of false positive solutions that produce similar scores
- In some cases the correct solution is not among the first ranked results

Biochemical Complementarity

- There are some types of residues (amonoacids) that have a preference to bind, while other types prefer not to bind.
- **Binding Site Prediction** is the task to identify specific regions on a protein surface that have a binding preference (**hot spots**)
- Binding Site Prediction can improve docking since it constrains the search space in geometric complementarity matching tasks.
- The following non-geometric factors are taken into account for binding site prediction:
  1. Electrostatic interaction energy
  2. Buried hydrophobic solvent accessible surface
  3. Hydrogen bonding energy
  4. Atom-contact surface area
  5. Overlap volume
  6. Residue conservation score
  7. Residue interface propensity

Binding Site Prediction

The binding score for a surface residue *i* is given by:

$$Score(i) = w_1 E_{electro} + w_2 E_{hydrophobic} + w_3 E_{hbe} + w_4 E_{acsa} + w_5 E_{ov} + w_6 E_{conservation} + w_7 E_{propensity}$$

- The first 5 energy scores are extracted directly from the atom types and the protein structure
- $E_{conservation}$ measured by the self-substitution score from the sequence profile.

$$E_{propensity}(i) = \left(\ln \frac{p_r^{interface}}{p_r^{surface}} \right) \cdot \frac{S_r}{S_r^{ave}}$$

Contribution of residue *r* to Binding Site

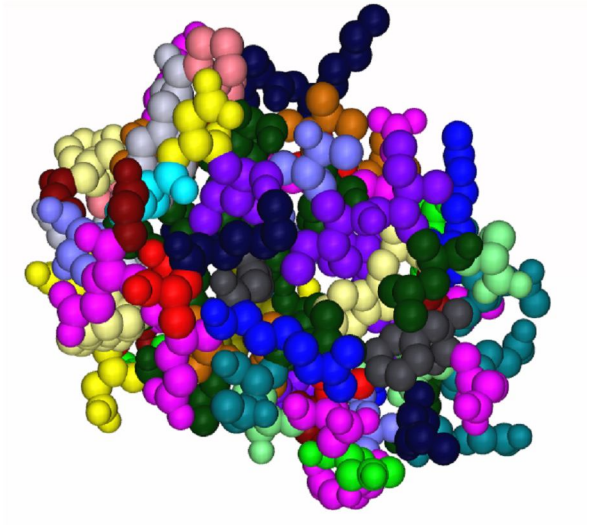
Contribution of residue *r* to surface

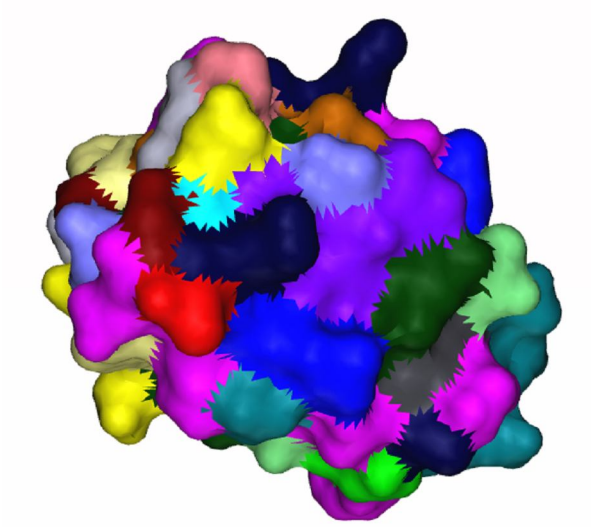
Residue Interface Propensity

- Select a large number of known complexes
  - >3000 structures of known dimers were selected from Protein Data Bank (PDB)
- Extract statistical data regarding the preference of residue types to take part in protein interactions

Residue Interface Propensity

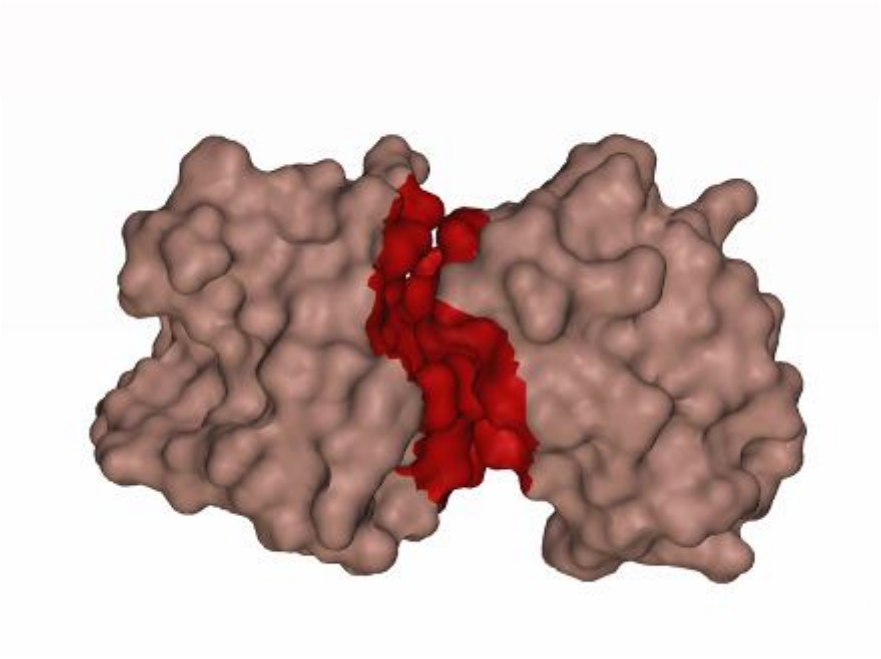
- 1st Step: extract 3D structure and “Connolly” surface for the two interacting molecules, incorporating the residue information.





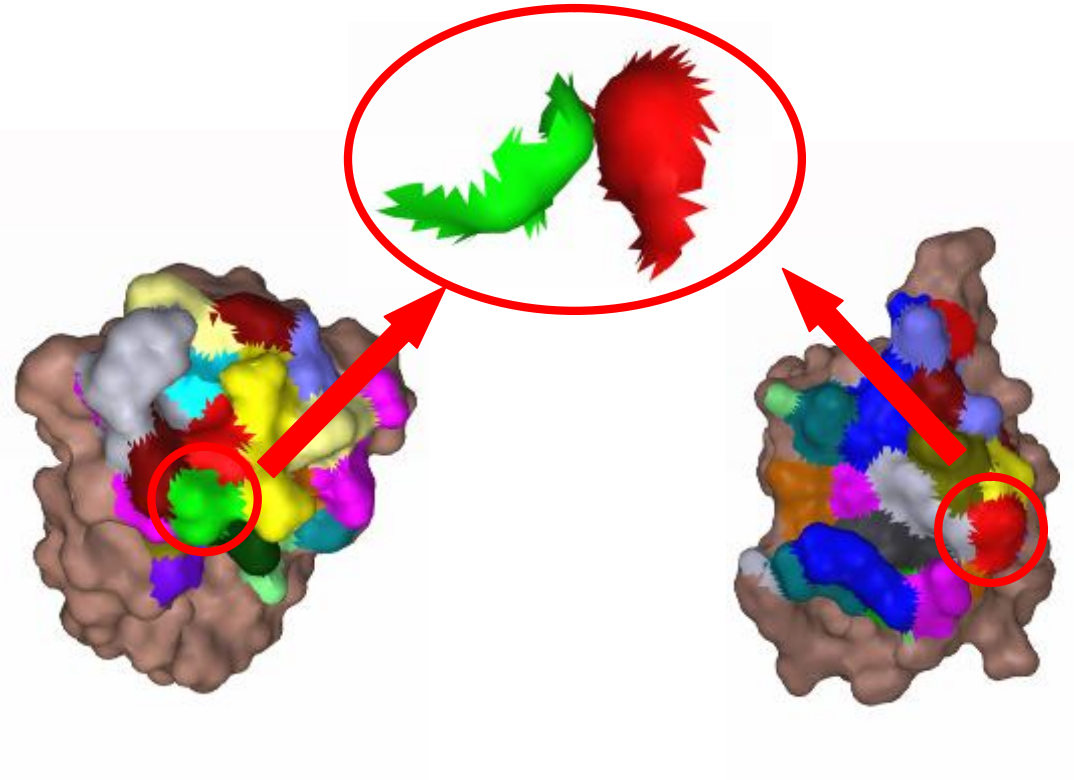
Residue Interface Propensity

- 2nd Step: Simulate the interaction of the 2 molecules in order to create the complex (the structure of the complex is also known).
  - Calculate the surface points that belong to the binding sites (distance < 4.7 Å)



Residue Interface Propensity

- 3rd Step: For each pair of interacting residues (are close enough) add a vote



Residue Interface Propensity

X Nr	Nr	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nr	Name	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
0	ALA	0,0022	0,0059	0,004	0,0031	0,002	0,0032	0,004246	0,003	0,0042	0,0036	0,0043	0,0037	0,0044	0,0051	0,0036	0,0032	0,0033	0,0039	0,0053	0,003426
1	ARG	0	0,0063	0,0088	0,0121	0,0044	0,011	0,008699	0,0069	0,0072	0,0063	0,0067	0,0076	0,0066	0,008	0,0072	0,0077	0,0072	0,0066	0,0102	0,00619
2	ASN	0	0	0,0039	0,0052	0,0022	0,0053	0,006706	0,0047	0,0057	0,0039	0,0044	0,0055	0,0041	0,0052	0,0052	0,005	0,0049	0,0042	0,0063	0,003645
3	ASP	0	0	0	0,0024	0,0016	0,0036	0,005247	0,0031	0,0058	0,0029	0,0035	0,008	0,0035	0,004	0,004	0,0044	0,0036	0,0035	0,0059	0,002735
4	CYS	0	0	0	0	0,0035	0,0016	0,002555	0,0018	0,0041	0,0027	0,0024	0,002	0,0026	0,0028	0,0021	0,0022	0,0018	0,0031	0,0027	0,002012
5	GLU	0	0	0	0	0	0,0026	0,005801	0,0032	0,006	0,0033	0,004	0,0085	0,0036	0,0047	0,0044	0,0049	0,0047	0,004	0,0062	0,003203
6	GLN	0	0	0	0	0	0	0,004894	0,0046	0,0059	0,0045	0,0052	0,0061	0,0051	0,0056	0,0054	0,0056	0,0056	0,0046	0,0066	0,004553
7	GLY	0	0	0	0	0	0	0	0,0017	0,0049	0,0029	0,0035	0,0046	0,004	0,0048	0,004	0,0034	0,0038	0,0042	0,0054	0,002722
8	HIS	0	0	0	0	0	0	0	0	0,0054	0,0042	0,0053	0,0043	0,0046	0,0057	0,0056	0,0057	0,0055	0,0046	0,0082	0,004208
9	ILE	0	0	0	0	0	0	0	0	0	0,0035	0,0059	0,0033	0,0057	0,0073	0,0043	0,0038	0,0043	0,0051	0,0064	0,0046
10	LEU	0	0	0	0	0	0	0	0	0	0	0,0043	0,0044	0,0064	0,008	0,0047	0,0041	0,0045	0,0051	0,0068	0,005334
11	LYS	0	0	0	0	0	0	0	0	0	0	0	0,003	0,004	0,0045	0,005	0,005	0,0053	0,0038	0,0066	0,003862
12	MET	0	0	0	0	0	0	0	0	0	0	0	0	0,0051	0,0071	0,0049	0,0042	0,0045	0,0054	0,0063	0,004881
13	PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0062	0,0063	0,0049	0,0053	0,0071	0,0081	0,00629
14	PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0029	0,0047	0,0046	0,0058	0,0075	0,003962
15	SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0029	0,0038	0,0044	0,0056	0,003231
16	THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0028	0,0043	0,0061	0,00392
17	TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0039	0,0047	0,004222
18	TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0052	0,005726
19	VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,002836

Some preferences:

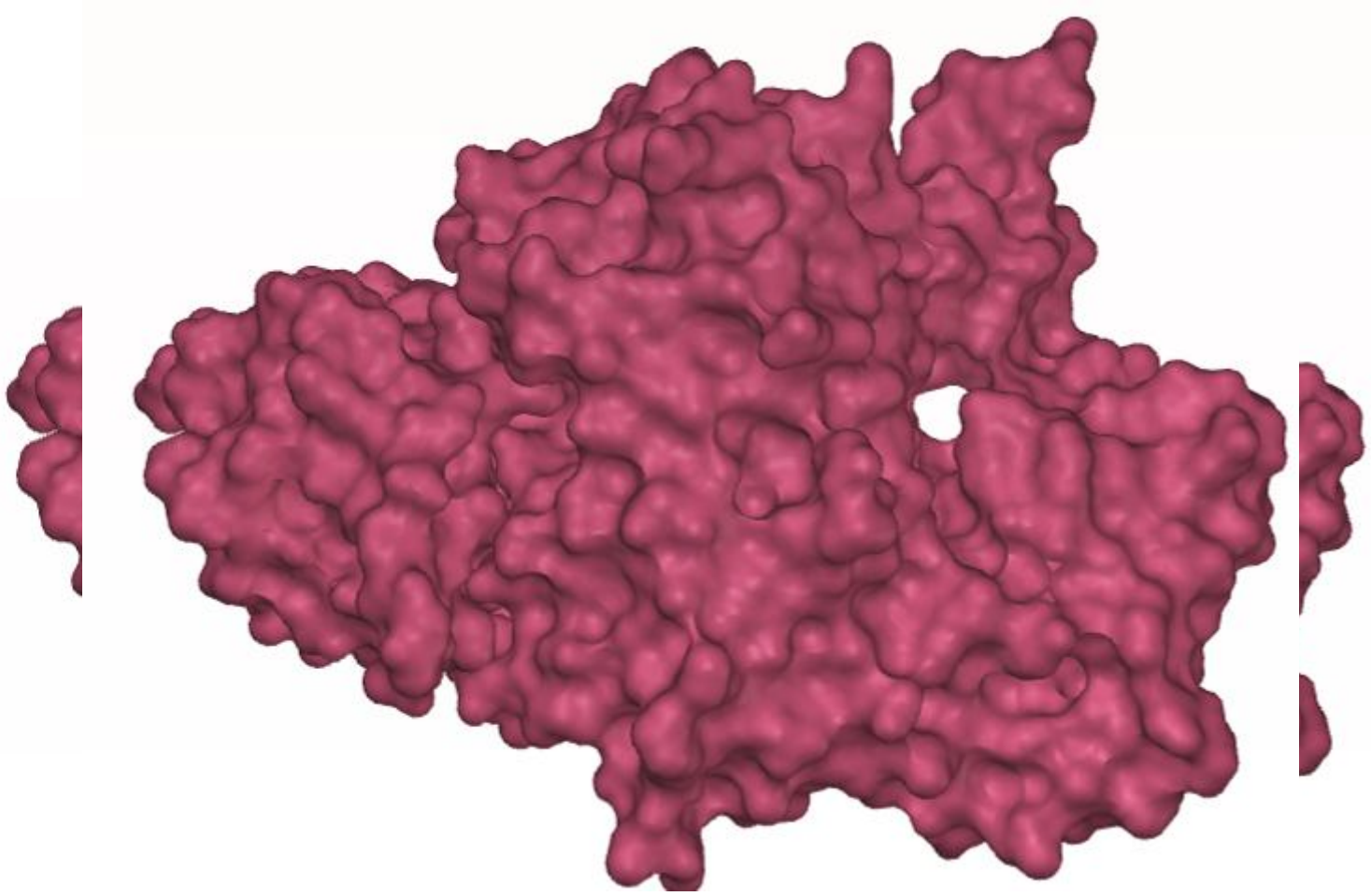
- **ARG – ASP**
- **ARG – GLU**
- **ARG – SER**
- **ARG – TYR**
- **ASP – LYS**
- **GLU – LYS**

Thessaloniki, October 2009

Rigid vs Flexible Docking

- In all methods of geometric docking described above, the interacting molecules are considered as rigid bodies.
- In fact, when proteins interact tend to change their shape on order to achieve better shape complementarity
- Protein flexibility adds thousands degrees of freedom apart from translation and rotation

Flexible Docking



Flexible Docking Approaches

- Soft Docking: modify scoring functions so as to be tolerant to small side-chain deformations (soft scoring functions)
- Modeling side-chain flexibility: use a predefined library of side chain conformations (rotamers), apply the conformations to the interacting molecules and select the one with the best score
- Modeling backbone flexibility: study the potential deformations of the protein backbone (similar to protein folding problem)

Conclusions

- There are several factors that we need to take into account for correct prediction of protein interactions:
  - Geometric complementarity
  - Protein flexibility
  - Binding preferences (binding site prediction)
- The integration of docking algorithms with protein interface prediction software, structural databases and sequence analysis techniques will help produce better and more accurate predictions.