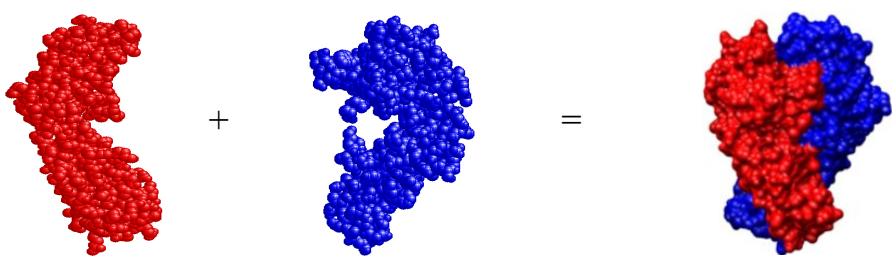
Lecture 16

Protein-protein
Docking

Protein-Protein Docking

Ab initio **Protein-protein docking** is the determination of the molecular structure of *complexes* formed by two or more proteins without the need for *experimental* measurement.



Background

It will take few decades to experimentally determine all the protein complex structures at atomic level resolution.

An alternative: computational modeling of protein-protein interactions; commonly known as protein-protein docking.

Proteins are the building blocks of the cells; perform bulk of the functions of the cell.

The functionality of a protein is determined by its interaction with other proteinous or non-proteinous molecules.

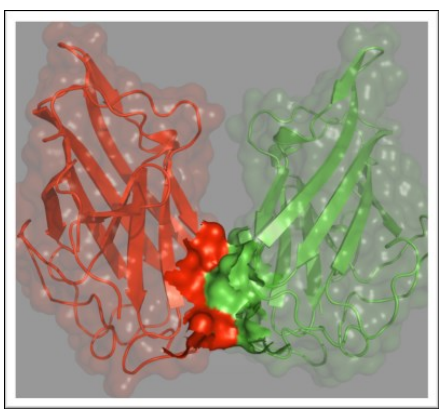
Epstein–Barr virus(

Aloy et al. (2004). *Nat. Biotechnology*

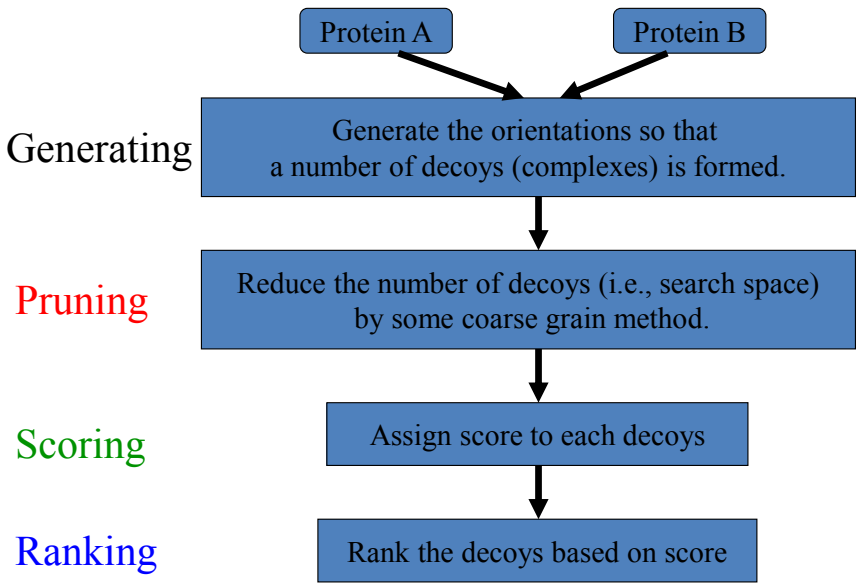
Calderwood M A et al. (2007) *PNAS* **104**, 7606-7611.

Protein Interaction

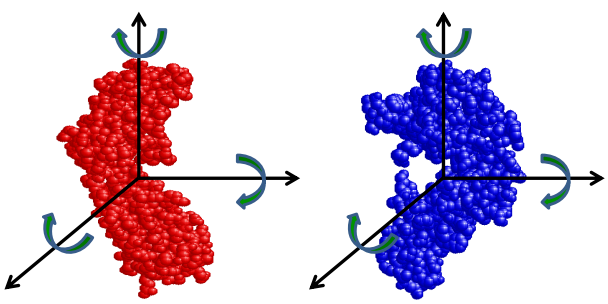
- Identify the interacting surface for two given protein molecules.
  - To model the protein complex structure formed out of two protein molecules – protein-protein docking



Docking Strategy



Generation methods



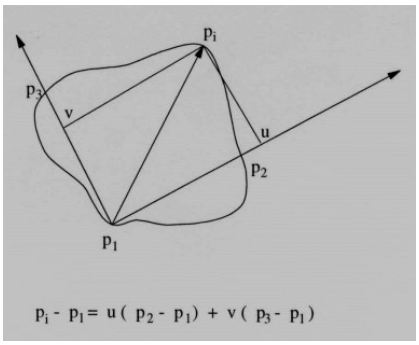
- Tagline – “Higher the decoys; better the possibility of having a hit”
- How many is good?
- Move to discrete space

Geometric Hashing

- Models are represented in a redundant affine invariant way and stored in a table (off-line).
- Hashing is used for organizing and searching the table.

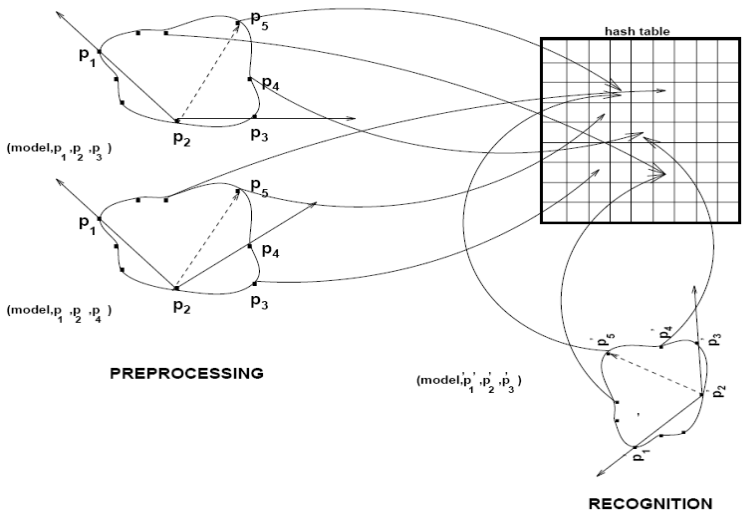
Affine Invariants

- Each triplet of non-collinear model points forms a basis of a coordinate system that is invariant under affine transformations.
- Represent model points in an affine invariant way by rewriting them in terms of this coordinate system.

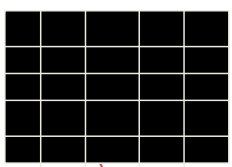


(u,v) are affine invariant!

Preprocessing and Recognition



Preprocessing Step



For each model do:

(1) Extract model's point features.

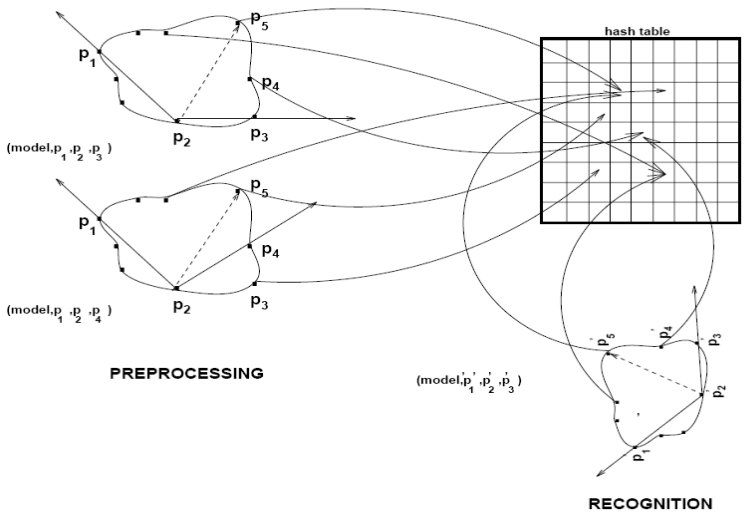
(2) For each ordered set of three, non-collinear, points (p_1, p_2, p_3)

(a) Compute the coordinates **(u,v)** of the remaining features in the coordinate frame defined by the model basis **(p_1, p_2, p_3)**

(b) After a proper quantization, use the computed coordinates **(u,v)** as an index to a two dimensional hash table, and record in the corresponding hash table bin the information **(model, (p_1, p_2, p_3))**

Hash Function: **h(Q(u), Q(v))** →

Preprocessing and Recognition



Recognition Step

(1) Extract the image point features

(2) Choose an arbitrary ordered pair **(p'_1, p'_2, p'_3)**

(3) Compute the coordinates (u',v'), of the remaining feature points in the coordinate frame defined by the image basis **(p'_1, p'_2, p'_3)**

(4) After quantization, use the computed coordinates as an index to the hash table. For every entry **(model, (p_1, p_2, p_3))** found in the corresponding bin, cast a vote.

Recognition Step (cont'd)

(5) Histogram all the hash table entries that received one or more votes. Determine those entries that received more than a certain number of votes -- each such entry corresponds to a potential match (hypothesis generation).

(6) For each potential match, consider all the model-image feature pairs which voted for a particular entry, and recover the affine transformation A that results in the best least-squares match between all the corresponding feature points.

Recognition Step (cont'd)

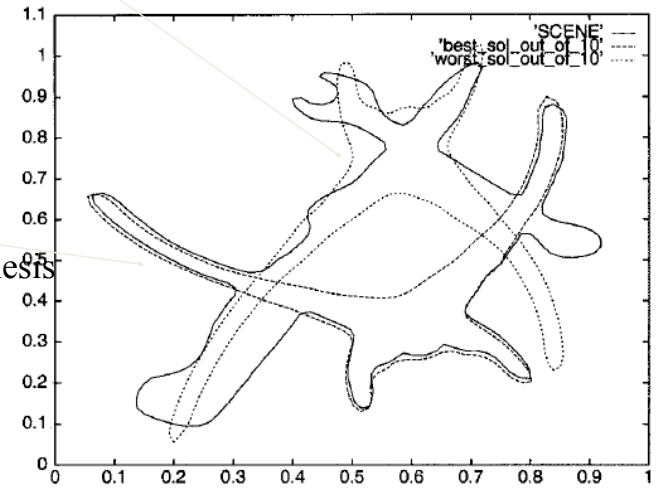
(7) Map the model onto the image using the computed transform and compare the model edges with the image edges (verification step).

(8) If the verification fails for all the models computed in step (5), go back to step (2) and repeat the procedure using a different image basis.

Recognition Example

Bad hypothesis

Good hypothesis



Complexity

- Preprocessing Step:

$O(Mm^4)$

- Recognition Step:

worst case: $O(i^4Mm^4)$

(M: #models, m: #model points, i: #scene points)

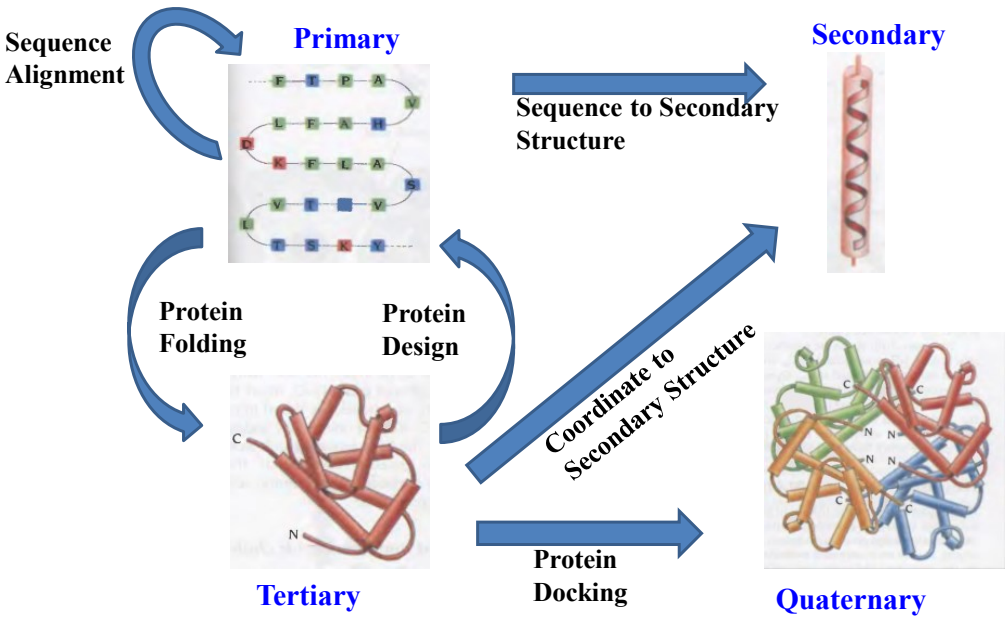
Geometric Hashing

- Pro:
  - Faster
- Con:
  - Storage requirement is very high and increases with the increase in object points.
  - Proper identification of object points are crucial for the success.

Lecture 17

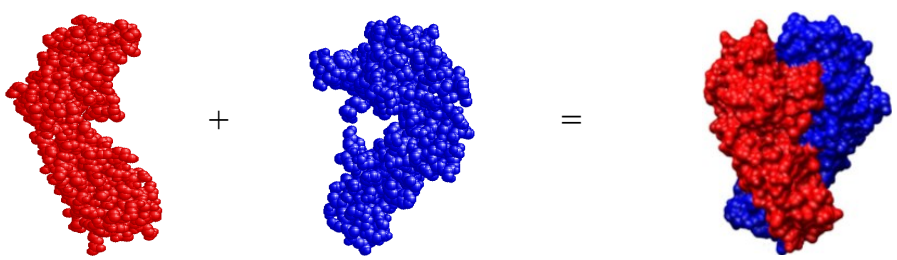
Protein-protein
Docking

Computational Methods in Proteins



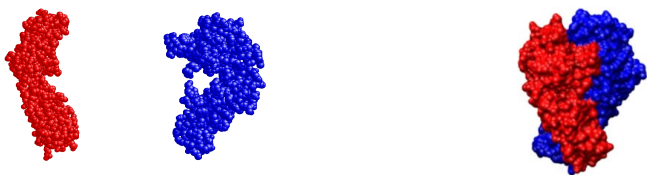
Protein-Protein Docking

Ab initio **Protein-protein docking** is the determination of the molecular structure of *complexes* formed by two or more proteins without the need for *experimental* measurement.



Docking Types

- Based on crystallization information
  - Bound docking
  - Unbound docking/predictive docking



(A) (B) (C)

- Based on protein flexibility
  - Rigid Body
  - Flexible Body

Rigid to Flexible

- Protein as a rigid body – a conceptual model
  - No bond stretching/bending
  - Less number of degrees of freedom for the atoms
  - Manageable possibilities
  - Not a practical scenario.

- Protein as a flexible body - a reality
  - Real life situation
  - Possibilities are enormous
  - Mathematical modeling of flexible body is difficult.

Protein Flexibility

Bad News	Good News
Dimensionality of the problem	Whole protein is not flexible ☺
If whole protein is flexible	☹
It varies from case to case	Predict the flexible regions a prior ☺
If flexibility occurs at some distance place not at the contact area	☹
If conformational changes occur after the binding/docking	☹

Protein flexibility is largely infeasible.
Ligand flexibility is possible.

Algorithms for Protein Docking

- Shape Complementarity
  (Connolly 1986, Cazals *et al* 2003)
- Fourier Technique
  (Katchalski-Katzir *et al*, 1992 (FTDock))
- Shape Polynomial and Shape Descriptor
  (Taubin *et al* 1989, Goldman *et al* 2000, Kazhdan *et al* 2003)
- Docking using Biochemical Information
  (Gabb *et al* 1997, Ritchie and Kemp 2000, Chen and Weng 2002, Fernandez-Recio *et al* 2002, Dominguez *et al* 2003 (HADDOCK), Murakami and Jones 2006 (SHARP))
- Meta-heuristic Approach
  (SA - Goodsell and Olson 1990, GA - Poirrette *et al* 1997, Gardiner *et al* 2001)
- Machine Learning Approach
  (Neuvirth *et al* 2004, Bradford and Westhead 2005, Qi *et al* 2006)

Docking Search Strategies

- **Pseudo Random**
  - Simulated Annealing / Monte Carlo
  - Genetic Algorithms
- **Directed Search**
  - Geometric Hashing
  - Spherical Harmonic Surface Triangles
- **Brute-Force Search**
  - Explicit Grid Correlations
  - Fast Fourier Transform (FFT) Correlations
  - Spherical Polar Fourier Correlations

Scoring in docking methods

Integrated Scoring

FTDock (Katchalaski-Katzir (1992), PNAS)

Edge Scoring

PRUNE (Mitra and Pal (2011), Nucleic Acids Res.)