Computational Modeling of Protein-Protein Interaction

Yinghao Wu
Department of Systems and Computational Biology
Albert Einstein College of Medicine
Fall 2014

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

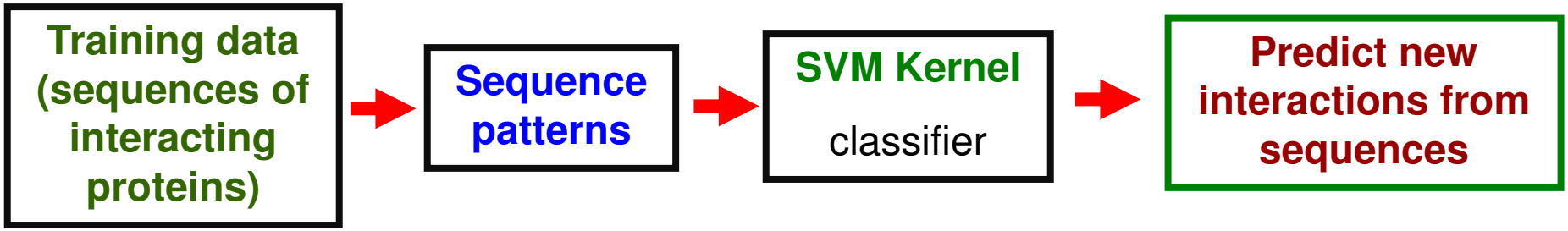
Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI

Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI

Binary prediction of PPI: General procedure

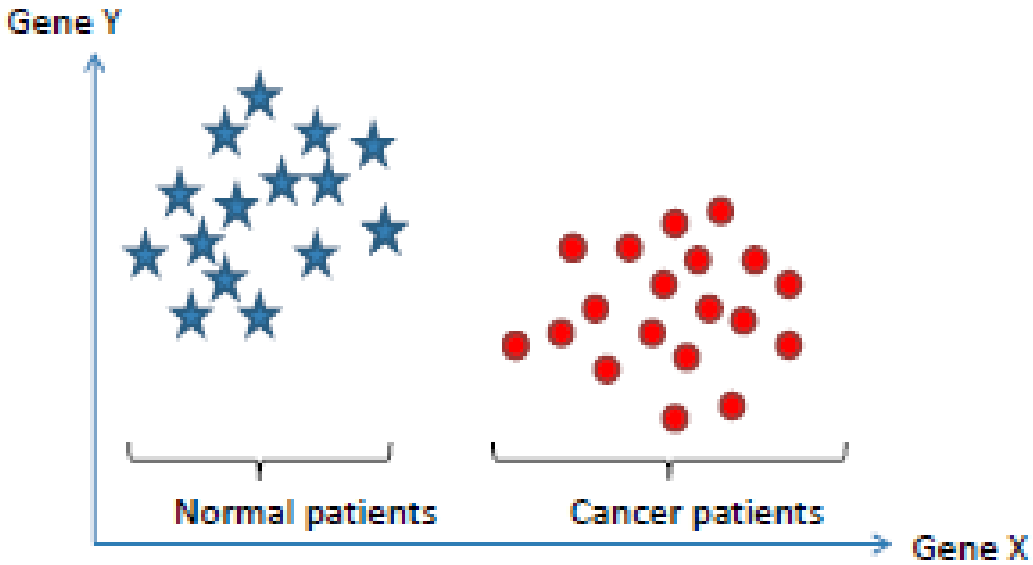


Training set for SVM kernel classifier

= Positive training set (experimental interactions, some for training, some for validation)

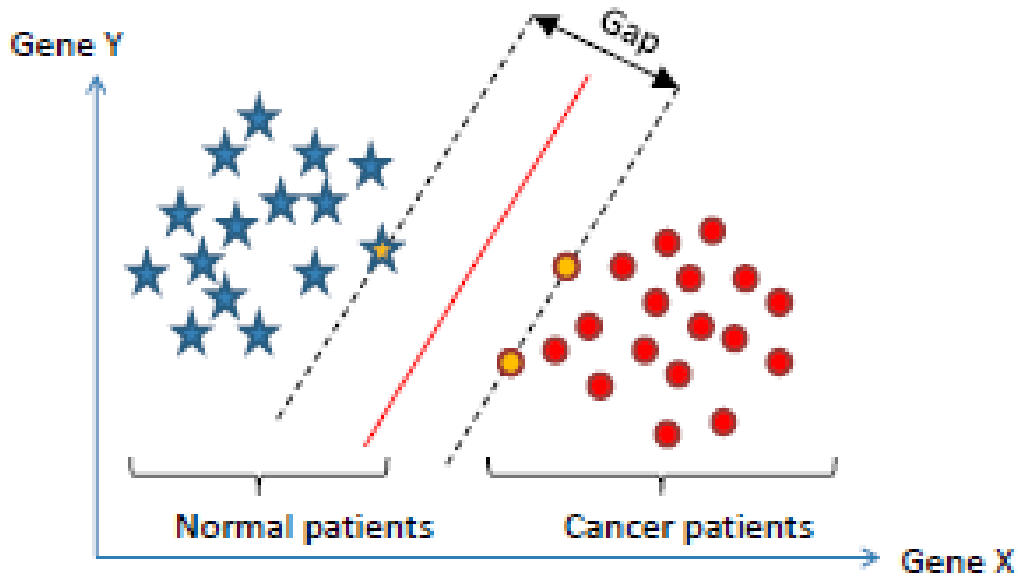
+ Negative training set (mostly random generated pairs)

Main ideas of SVMs



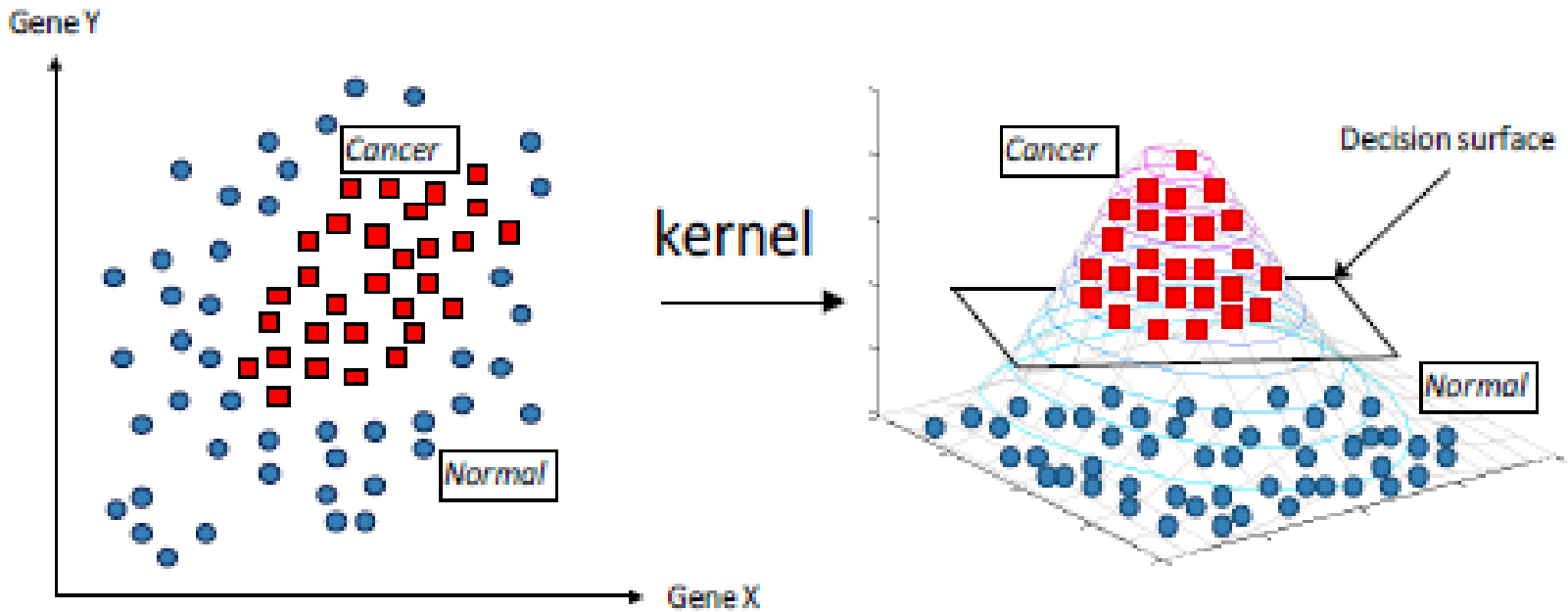
- Consider example dataset described by 2 genes, gene X and gene Y
- Represent patients geometrically (by "vectors")

Main ideas of SVMs



- Find a linear decision surface ("hyperplane") that can separate patient classes <u>and</u> has the largest distance (i.e., largest "gap" or "margin") between border-line patients (i.e., "support vectors");

Main ideas of SVMs



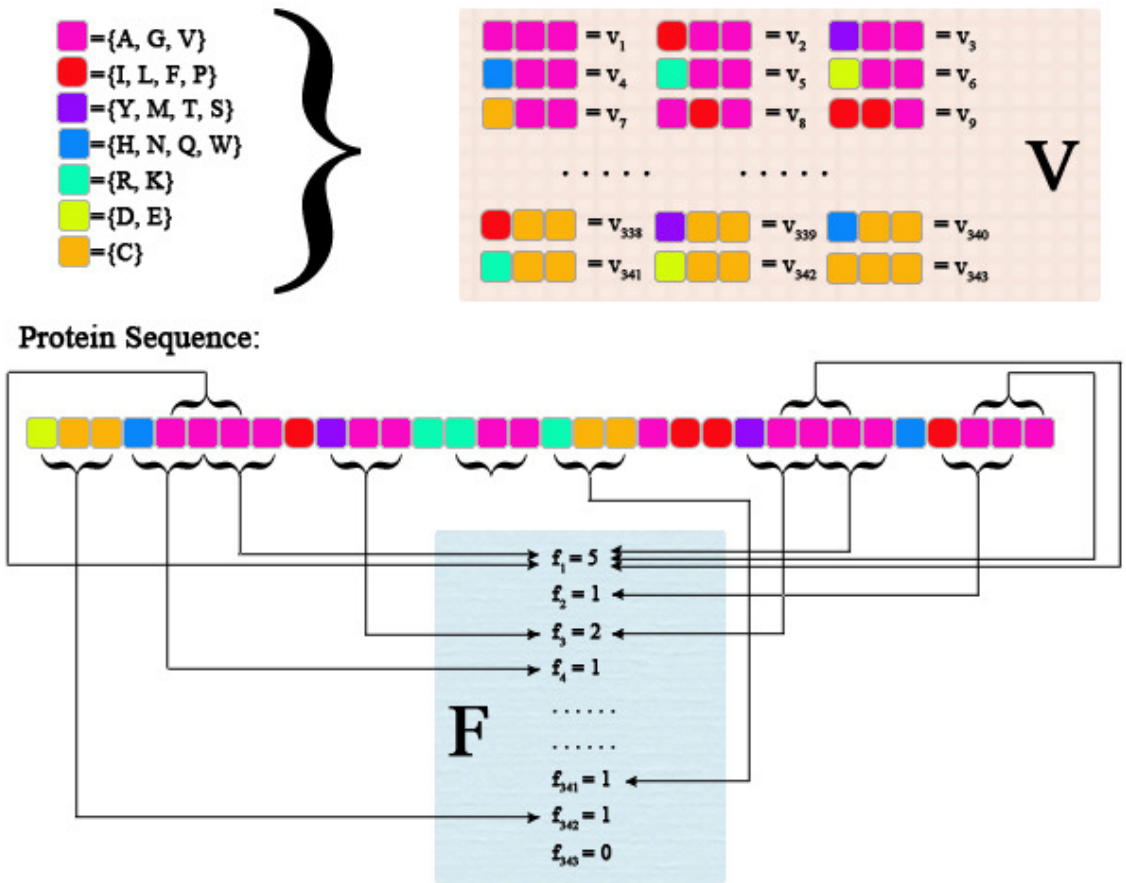
- If such linear decision surface does not exist, the data is mapped into a much higher dimensional space ("feature space") where the separating decision surface is found;
- The feature space is constructed via very clever mathematical projection ("kernel trick").

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Classification of Amino Acid(AA)

No.	Dipole scale[a]	Volume scale[b]	Class
1	-	-	Ala, Gly, Val
2	-	+	Ile, Leu, Phe, Pro
3	+	+	Tyr, Met, Thr, Ser
4	++	+	His, Asn, Gln, Tpr
5	+++	+	Arg, Lys
6	+'+'+'	+	Asp, Glu
7	+ [c]	+	Cys

Using Conjoint Triads for sequence pattern construction



Reduced-alphabet sequence pattern training:

1. Classify 20 AA types into 7 classes based on their properties (hydrogen bonding, hydrophobic, volumes of sidechains, etc).
2. Build AA triplets using 7 classes, called “conjoint triad” (343 unique types). Save in **V**
3. Calculate frequency of each triad for each protein sequence.

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Kernel Function

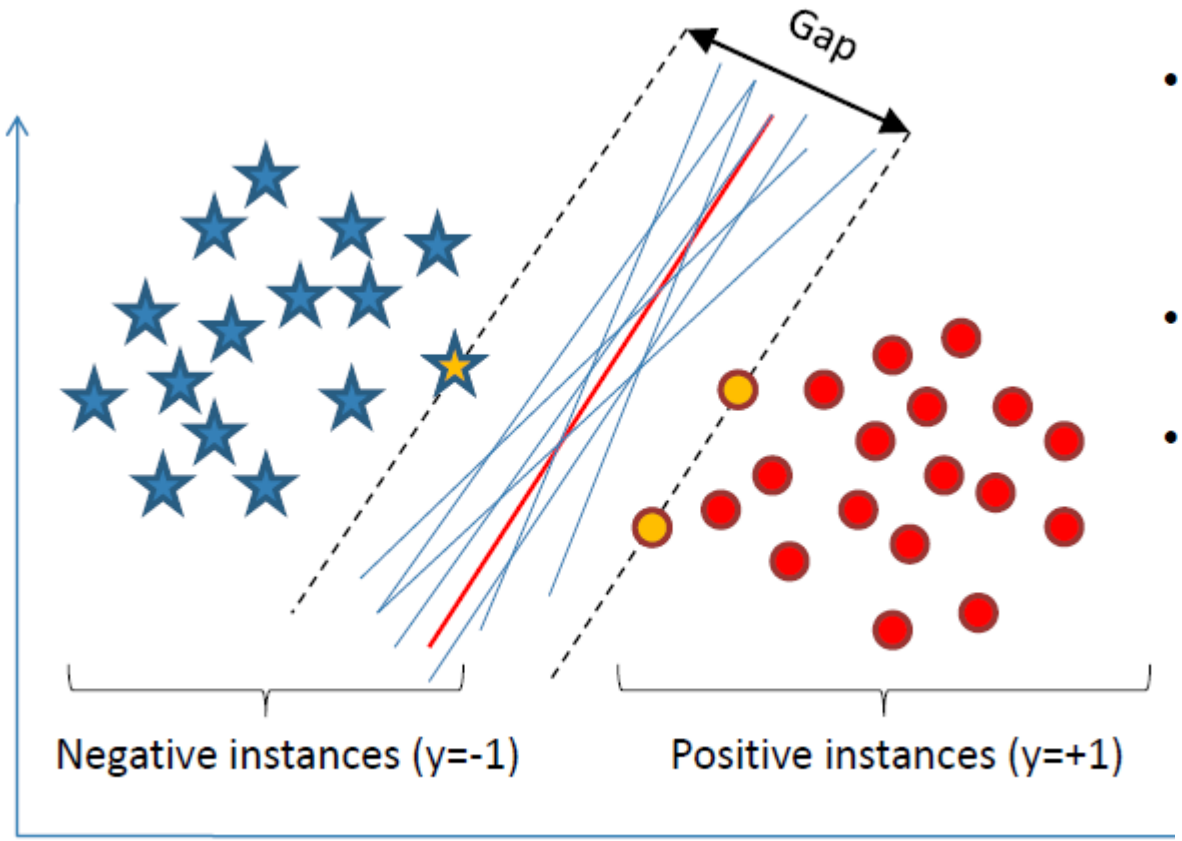
- **di** = (**fi** - min {**f**1, **f**2, ..., **f**343})/max{**f**1, **f**2, ..., **f**343}
- $\mathbf{D}_A$={$\mathbf{d}_A$1,$\mathbf{d}_A$2,……,$\mathbf{d}_A$343}
- {$\mathbf{D}_{AB}$} = {$\mathbf{D}_A$}⊕{$\mathbf{D}_B$}: a 686 dimensional vector
- **Kernel Function**:

$$K(\mathbf{D_{AB}}, \mathbf{D_{EF}}) = \exp(-\gamma\|\mathbf{s}\|^2)s = \min\{(\|\mathbf{D_A} - \mathbf{D_E}\|^2 + \|\mathbf{D_B} - \mathbf{D_F}\|^2), (\|\mathbf{D_A} - \mathbf{D_F}\|^2 + \|\mathbf{D_B} - \mathbf{D_E}\|^2)\}.$$

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Given training data: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \in R^n$

$y_1, y_2, \ldots, y_N \in \{-1, +1\}$



- Want to find a classifier (hyperplane) to separate negative instances from the positive ones.
- An infinite number of such hyperplanes exist.
- SVMs finds the hyperplane that maximizes the gap between data points on the boundaries (so-called "support vectors").

Kernel Function and parameter adjustment

C=128
Γ=0.25

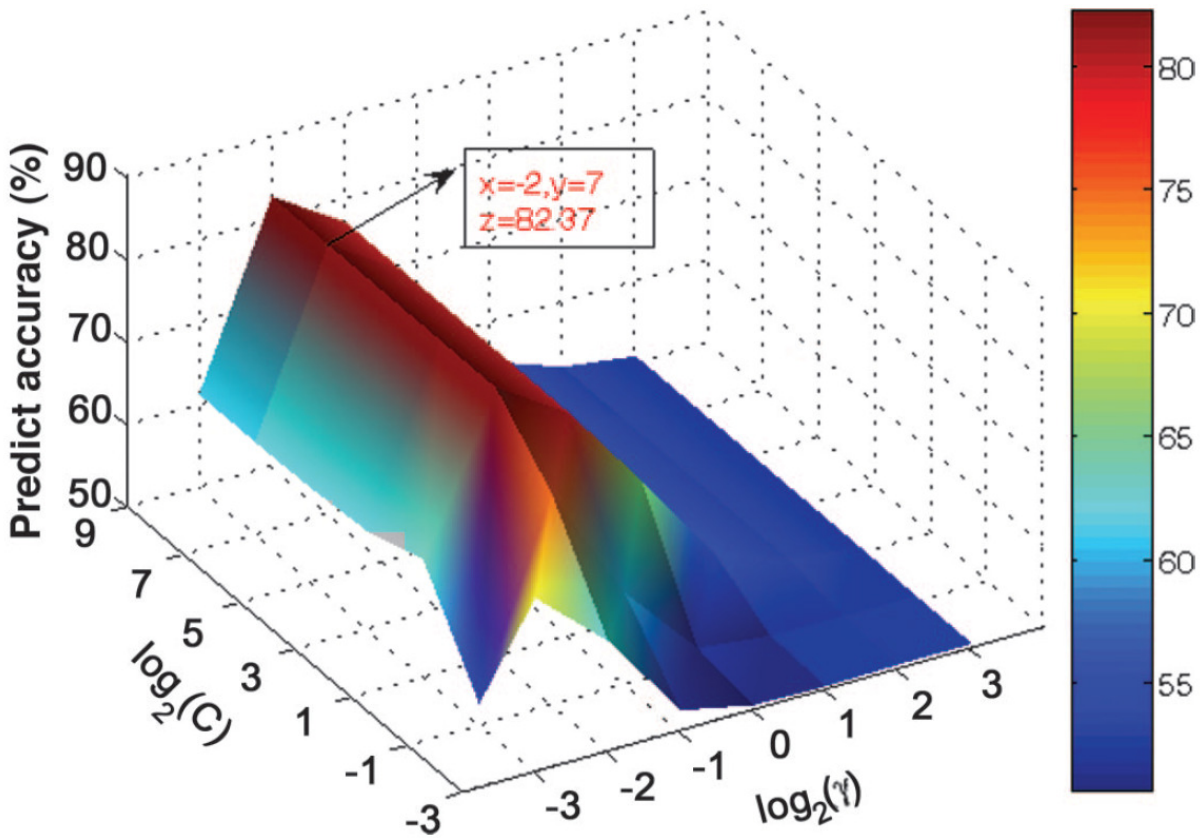
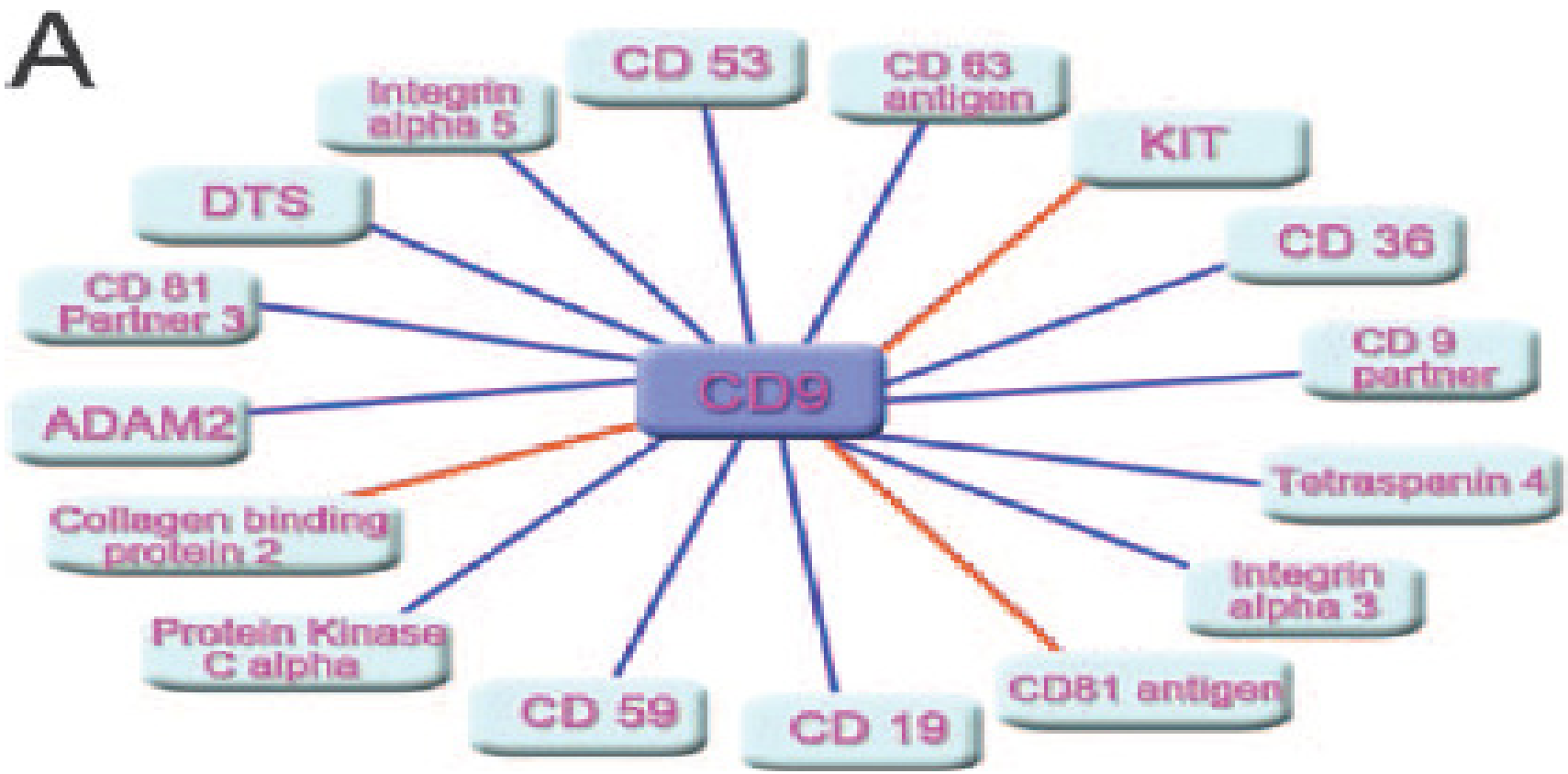


Fig. 1. Accuracy surface of threefold crossover validation on training set versus the variations of parameters C and γ.

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

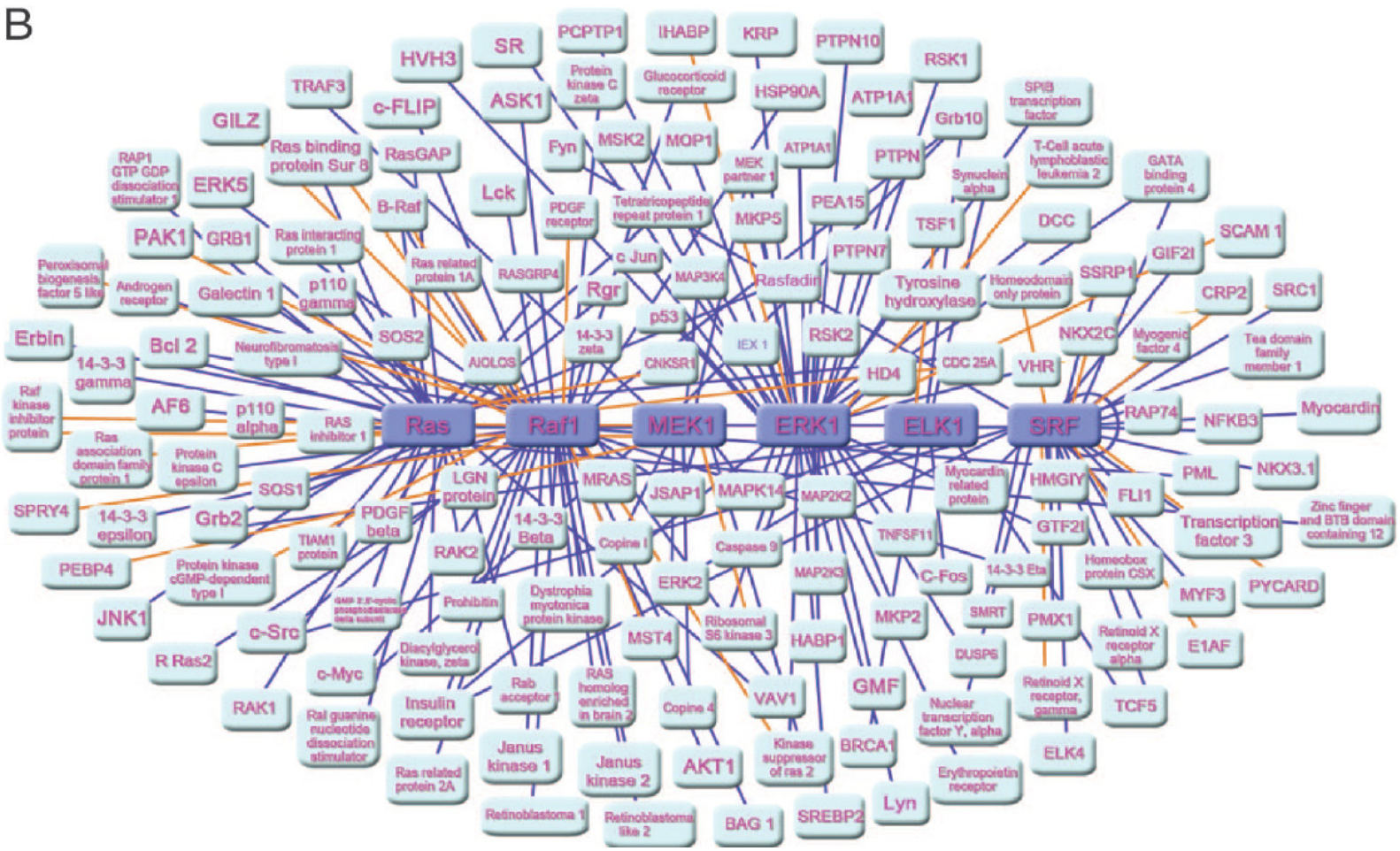
Network Prediction



One-core network

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

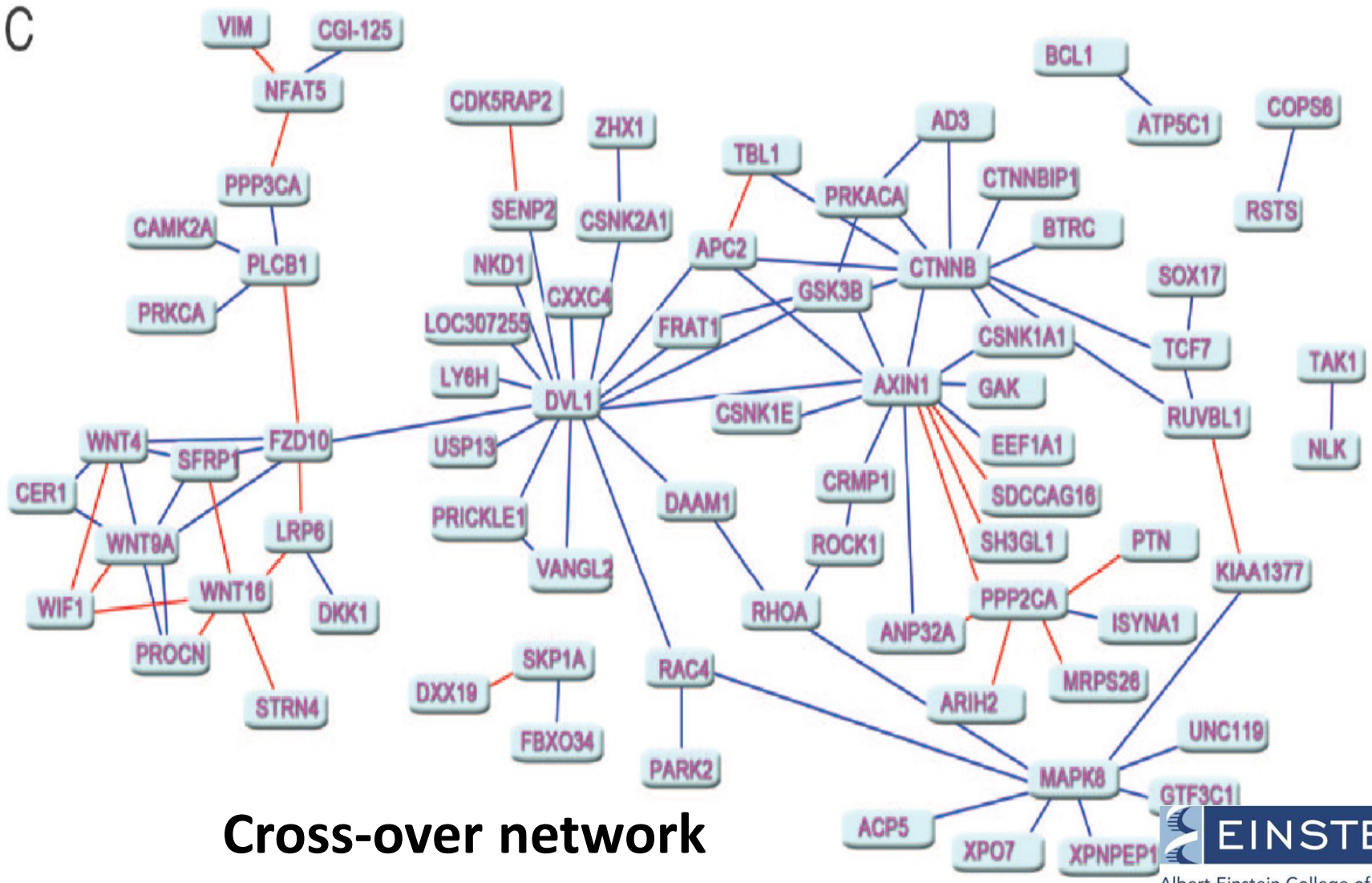
Network Prediction



Multi-core network

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Network Prediction



Cross-over network

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

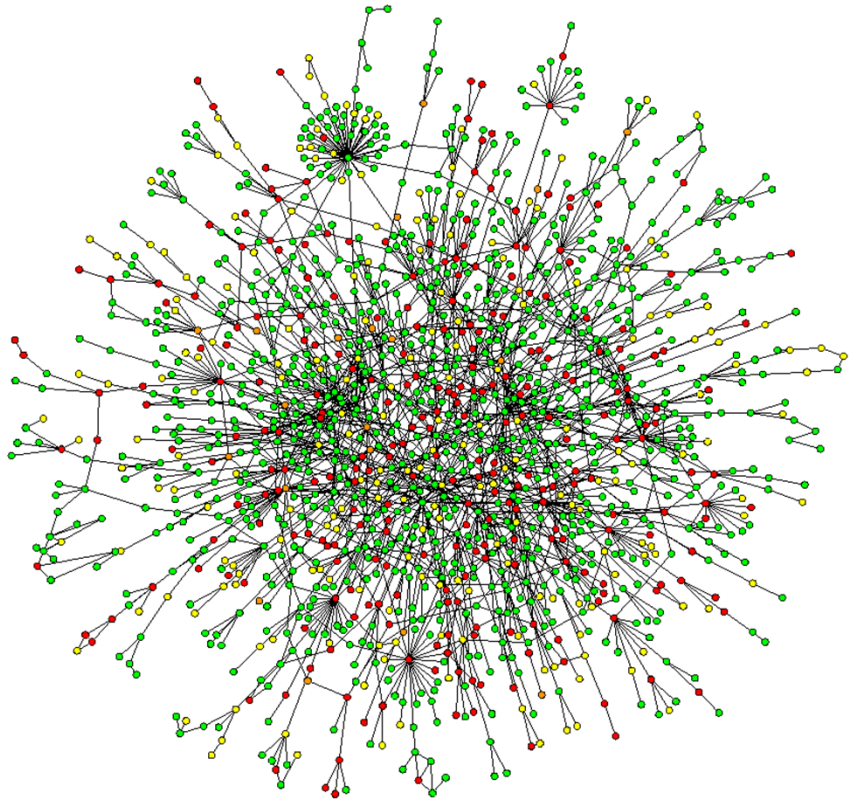
Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Protein-Protein Interaction Networks?

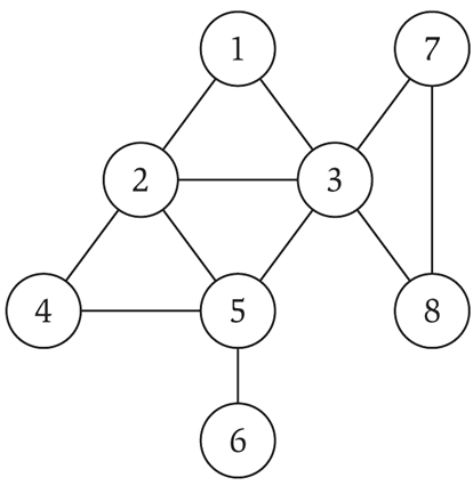
- Protein are nodes
- Interactions are edges



Yeast PPI ne

Introduction to graph theory

- **Graph** – mathematical object consisting of a set of:
  - V = **nodes** (vertices, points).
  - E = **edges** (links, arcs) between pairs of nodes.
  - Denoted by $G = (V, E)$.
  - Captures pairwise relationship between objects.
  - **Graph size** parameters: $n = |V|$, $m = |E|$.



V = { 1, 2, 3, 4, 5, 6, 7, 8 }

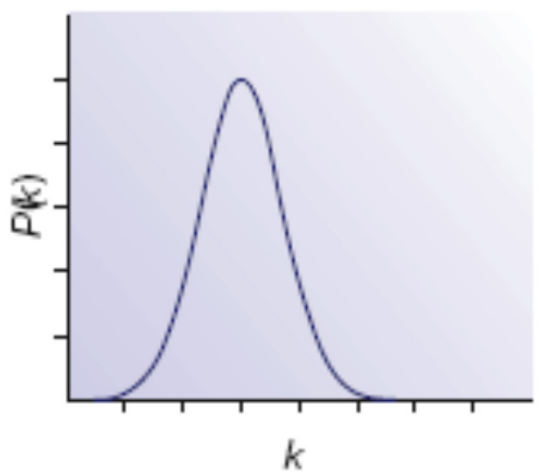
E = { {1,2}, {1,3}, {2,3}, {2,4}, {2,5}, {3,5}, {3,7}, {3,8}, {4,5}, {5,6} }

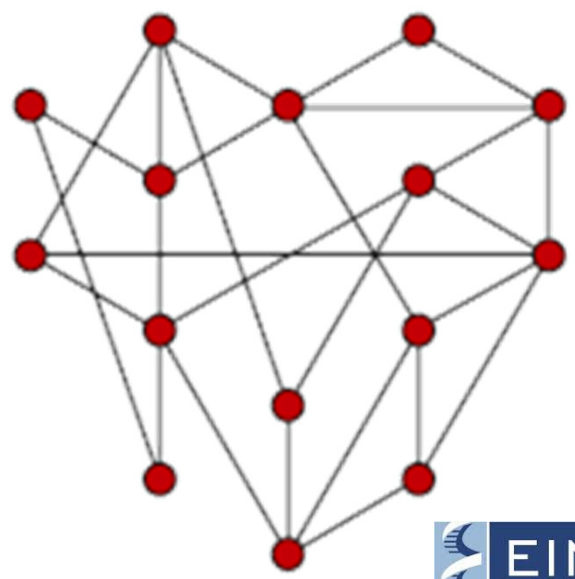
n = 8

m = 11

Random network

- Connect each pair of node with prob p
- Expect value of edge is $pN(N\text{-}1)/2$
- Poisson distribution
  - The node with high degree is rare

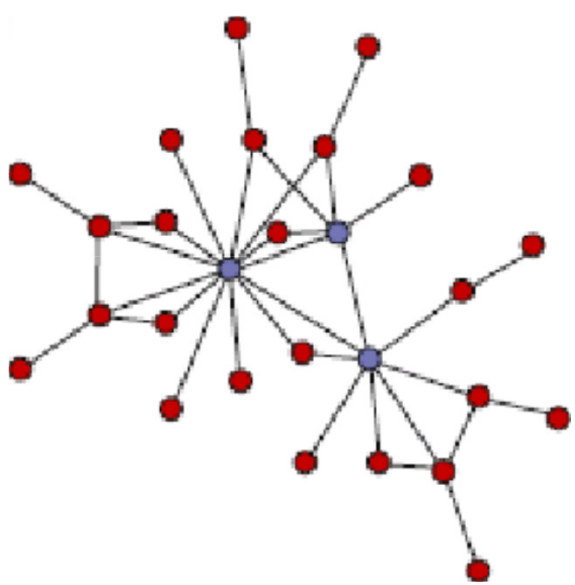


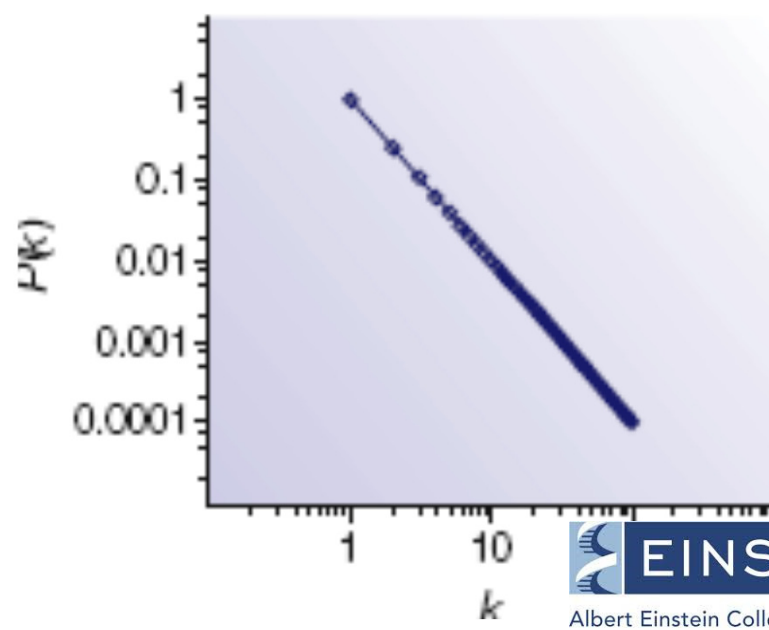


EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Scale-free network

- Power-law degree distribution
- Hubs and nodes
- When a node add into network, it prefer to link to hubs

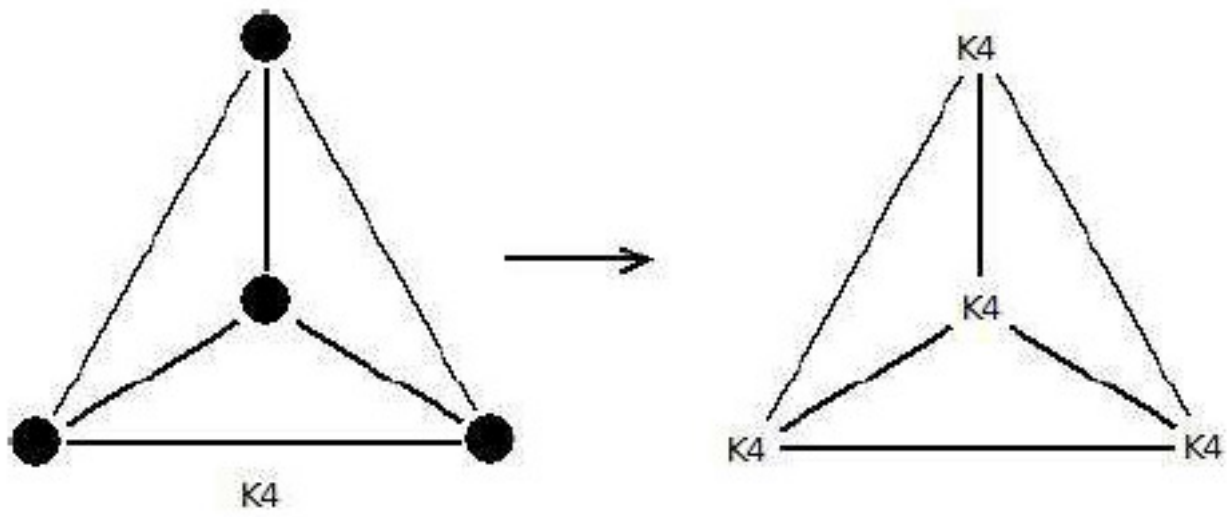




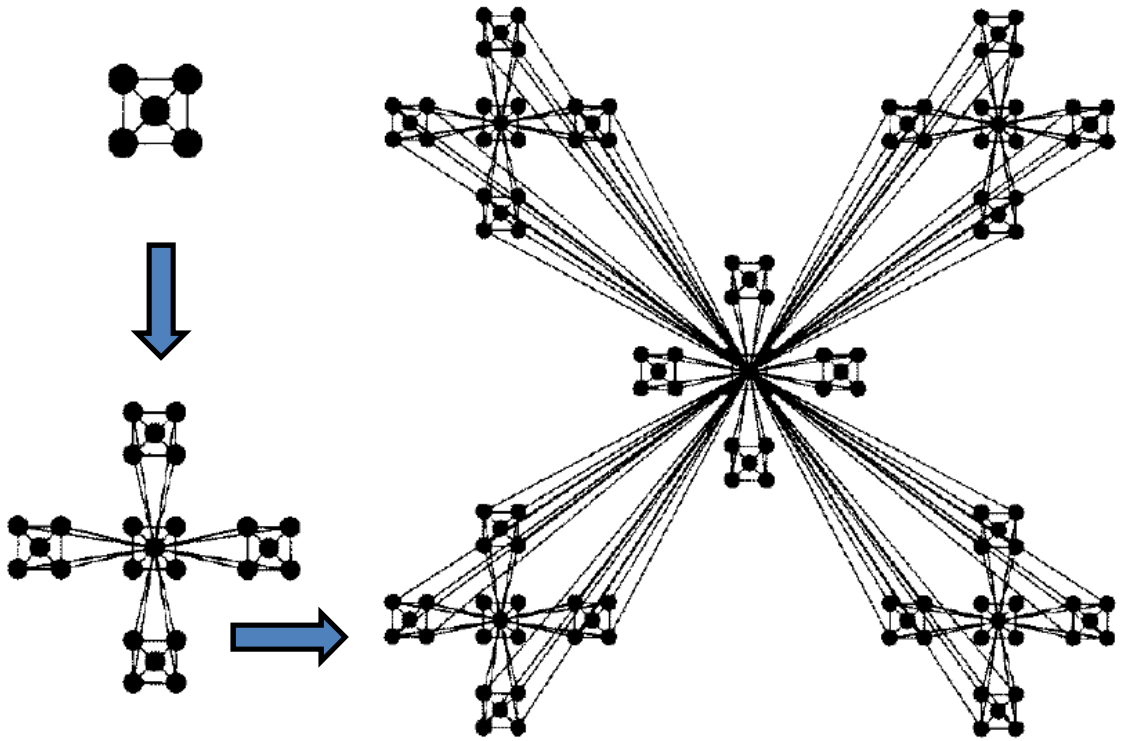
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Hierarchical network

- Preserves network “modularity” via a fractal-like generation of the network



Hierarchical network



EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Types of Network Comparisons

- 3 types (modes) of comparative methods:
  1. **Network alignment**
  2. **Network integration**
  3. **Network querying**

Types of Network Comparisons

1. Network alignment:

   - The process of comparison of two or more networks of <u>the same type</u> to identify regions of similarity and dissimilarity

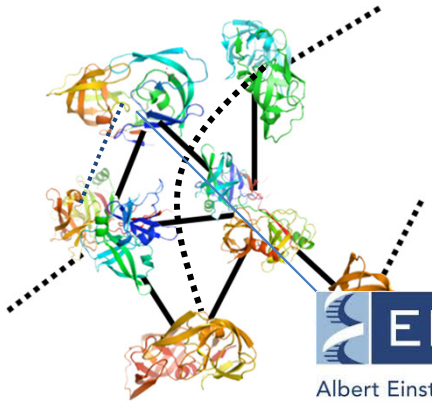
   - Commonly applied to detect subnetworks that are conserved across species and hence likely to present true functional modules

Types of Network Comparisons

2. Network integration:

- The process of combining networks encompassing <u>interactions of different types</u> over <u>the same set of elements</u> (e.g., PPI and genetic interactions) to study their interrelations

- Can assist in uncovering protein modules supported by interactions of different types



EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Types of Network Comparisons

- A grand challenge:

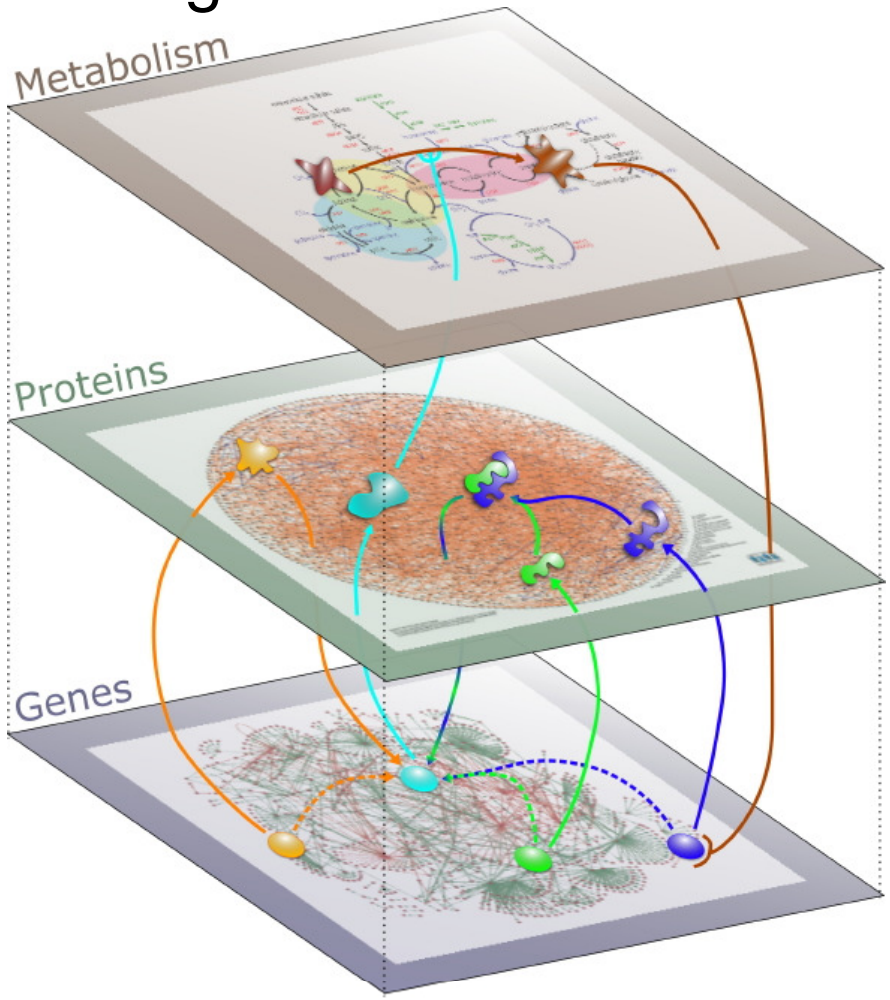


Image from: http://www-dsv.cea.fr/en/institutes/institute-of-biology-and-technology-saclay-il

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Types of Network Comparisons

3. Network querying:

- A given network is searched for subnetworks that are similar to a subnetwork query of interest

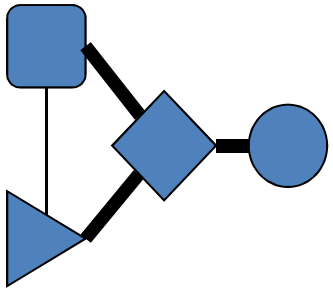
- This basic database search operation is aimed at transferring biological knowledge within and across species

- Currently limited to very sparse graphs, e.g., trees

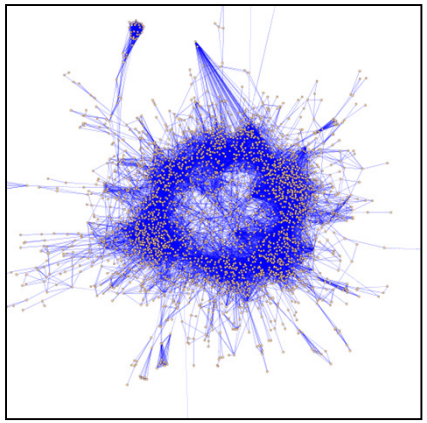
Types of Network Comparisons

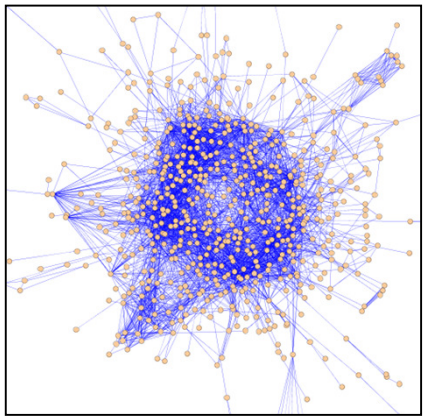
3. Network querying
   - Useful application for biologists: given a candidate module, align to a database of networks ("query-to-database")

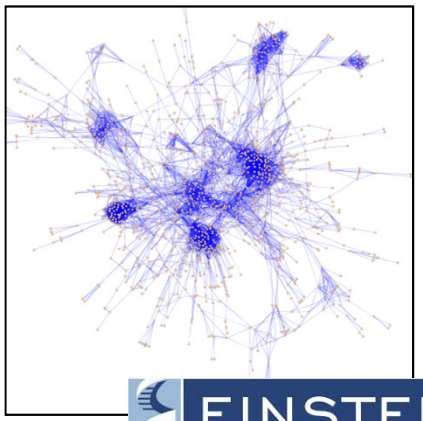
Query:



Database:







EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Types of Network Comparisons

Summary

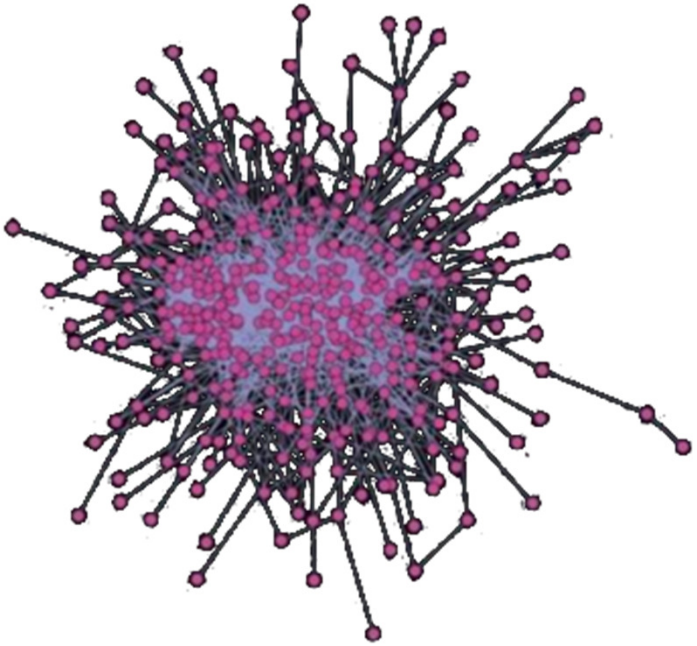
Table 1 Modes of network comparison

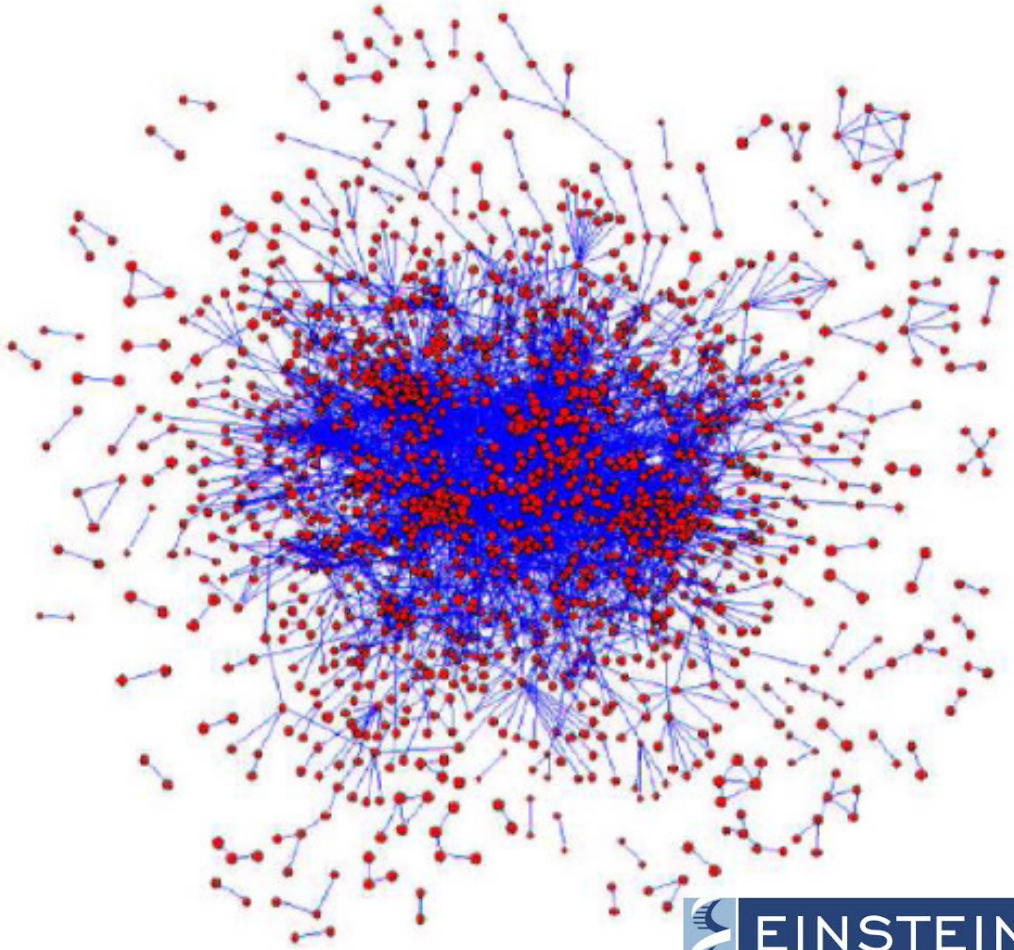
Mode	Common application	Main goals	Some current limitations
Alignment	At least two networks of the same type across species	Identification of functional (conserved) protein modules; study of network evolution; interaction prediction	Limited to few (five or fewer) species
Integration	At least two networks of different types for the same species	Identification of modules (supported by several networks); study of interrelations between data types; interaction prediction	No agreed-upon way to combine scores over different networks
Querying	Subnetwork module versus a network	Identification of duplicated/conserved instances of the module; knowledge transfer	Query is limited to a tree topology

Sharan and Ideker (2006) *Nature Biotechnology* **24**(4): 427-433

Network Alignment

- Finding structural similarities between two networks





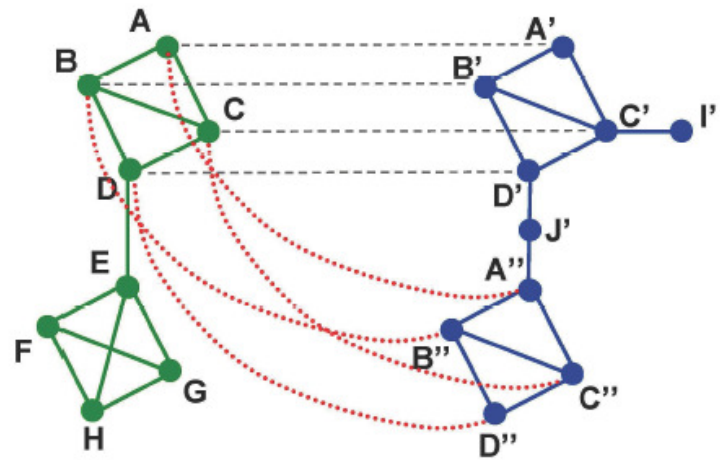
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Network Alignment

- Methods vary in these aspects:
  A. Global vs. local
  B. Pairwise vs. multiple
  C. Functional vs. topological information

Network Alignment

- Methods vary in these aspects:
  - **A. Global vs. local**
  - B. Pairwise vs. multiple
  - C. Functional vs. topological information



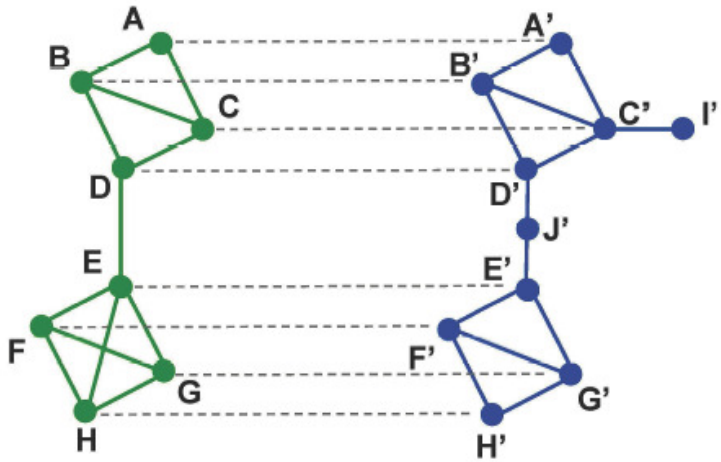
A. <u>Local alignment</u>:

- Mappings are chosen independently for each region of similarity
- Can be ambiguous, with one node having different pairings in different local alignments
- Example algorithms:

  PathBLAST, NetworkBLAST, MaWISh, Graemlin

Network Alignment

- Methods vary in these aspects:
  - **A. Global vs. local**
  - B. Pairwise vs. multiple
  - C. Functional vs. topological information



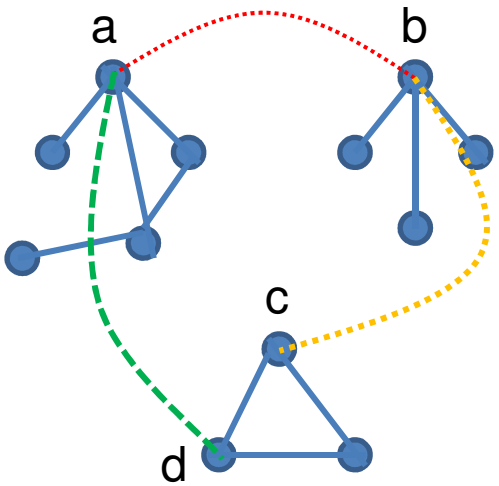
A. <u>Global alignment</u>:

- Provides a unique alignment from every node in the smaller network to exactly one node in the larger network
- May lead to inoptimal matchings in some local regions
- Example algorithms:

  IsoRank, IsoRankN, Graemlin 2, GRAAL, H-GRAAL

Network Alignment

- Methods vary in these aspects:
  A. Global vs. local
  B. Pairwise vs. multiple
  C. Functional vs. topological information



B. <u>Pairwise alignment</u>:

- Two networks aligned
- Example algorithms:
  GRAAL, H-GRAAL, PathBLAST, MaWISh, IsoRank

<u>Multiple alignment</u>:

- More than two networks aligned
- Computationally more difficult than pairwise alignment
- Example algorithms:
  Greamlin, Extended PathBLAST, Extended IsoRank

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Network Alignment

- Methods vary in these aspects:
  A. Global vs. local
  B. Pairwise vs. multiple
  C. Functional vs. topological information

C. <u>Functional information</u>

- Information external to network topology (e.g., protein sequence) used to define "similarity" between nodes
- Careful: mixing different biological data types, that might agree or contradict

<u>Topological information</u>

- Only network topology used to define node "similarity"
- Good – since it answers how much and what type of biological information can be extracted from topology only

Network Alignment

- In general, the network alignment problem is computationally hard (generalizing subgraph isomorphism)
- Hence, heuristic approaches are devised

- For now, let us assume that we have a heuristic algorithm for network alignment
- How do we measure the quality of its resulting alignments?

Network Alignment

- **<u>Key algorithmic components</u>** of network alignment algorithms:
  - Node similarity measure
  - Rapid identification of high-scoring alignments from among the exponentially large set of possible alignments

Network Alignment

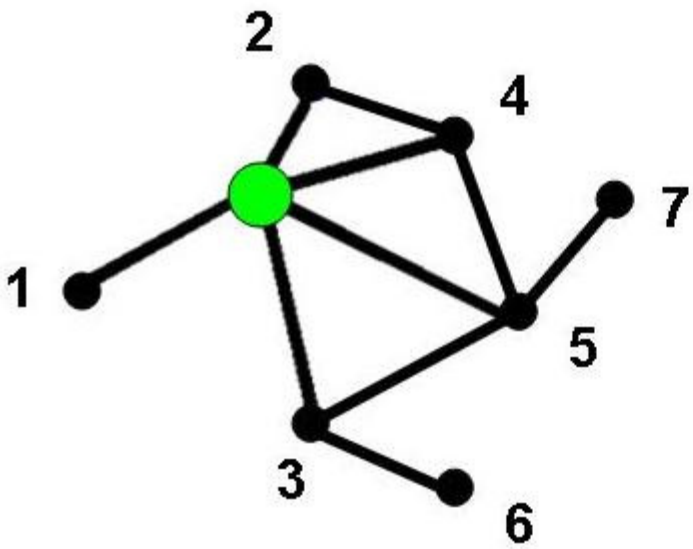
- How is **“similarity” between nodes** defined?
  - Using information external to network topology, e.g., the sequence alignment score
    - Homology, E-values, sequence similarity vs. sequence identity…
  - Using only network topology, e.g., node degree,
  - Using a combination of the two

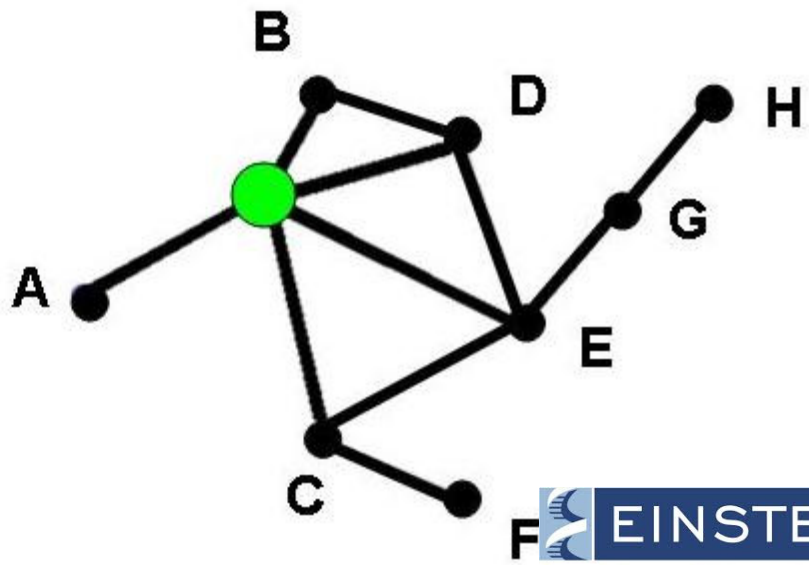
Network Alignment

- How to identify <u>high-scoring alignments</u>?
  - Idea: seeded alignment
    - Inspired by seeded sequence alignment (BLAST)
    - Identify regions of network in which “the best” alignments likely to be found

Network Alignment

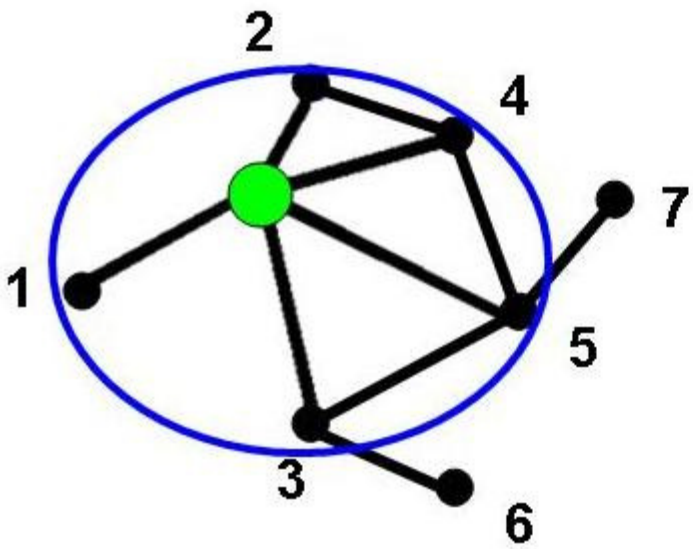
- How to identify <u>high-scoring alignments</u>?
  - Greedy ***seed and extend*** approaches
    - Use the most "similar" nodes across the two networks as "anchors" or **"seed nodes"**
    - "Extend around" the seed nodes in a greedy fashion

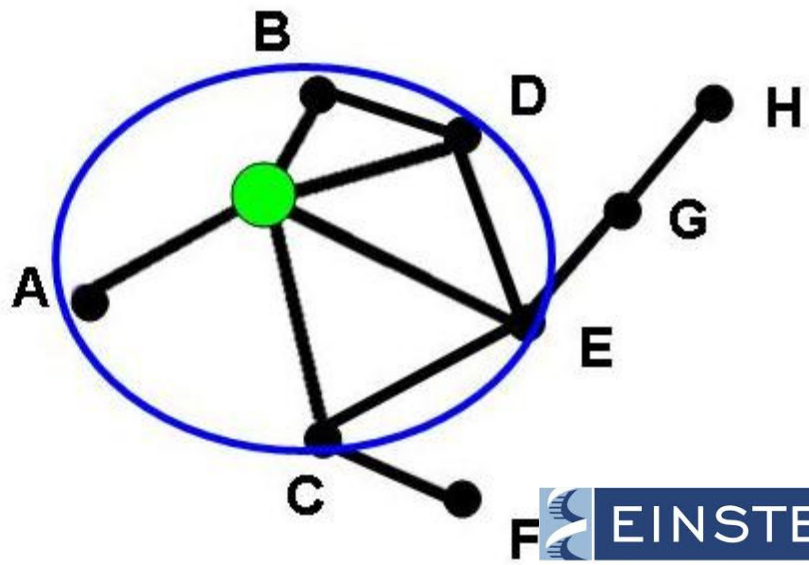




Network Alignment

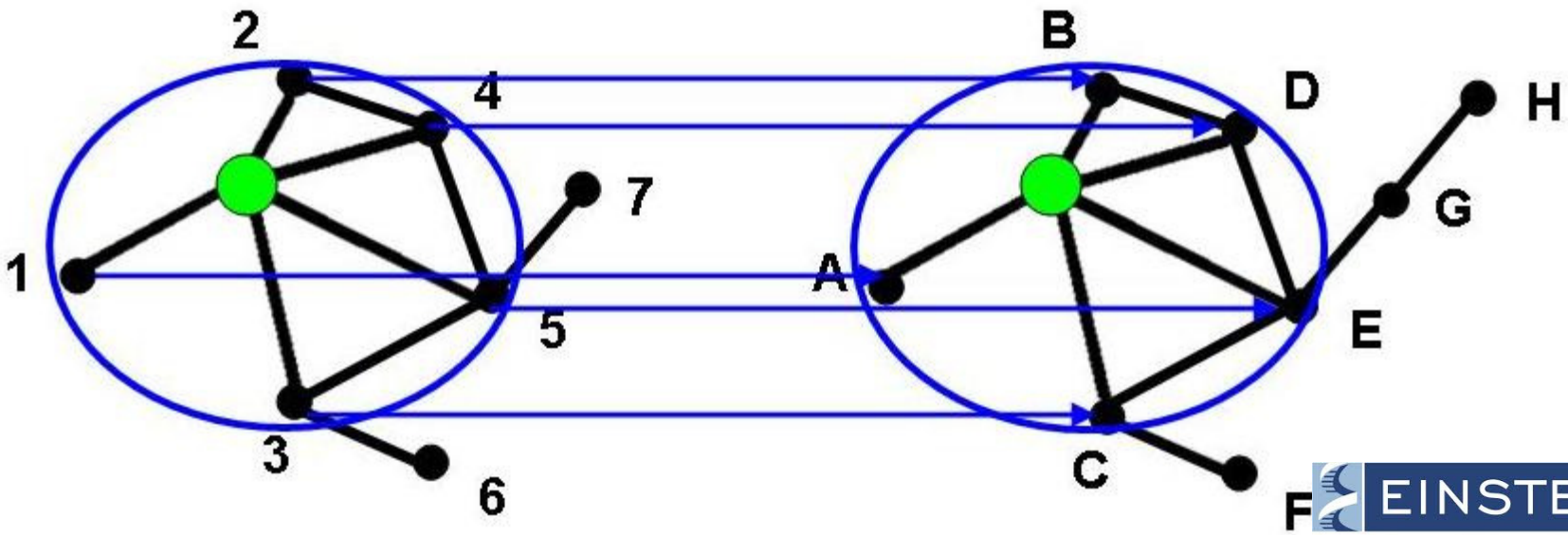
- How to identify <u>high-scoring alignments</u>?
  - Greedy ***seed and extend*** approaches
    - Use the most "similar" nodes across the two networks as "anchors" or **"seed nodes"**
    - "Extend around" the seed nodes in a greedy fashion





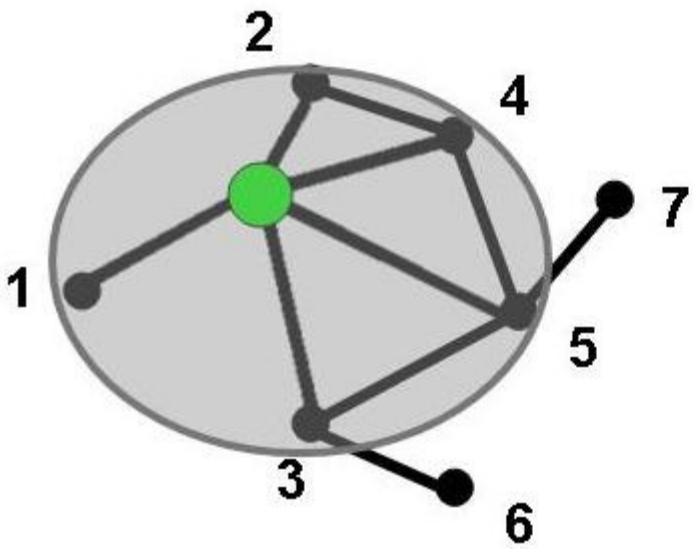
Network Alignment

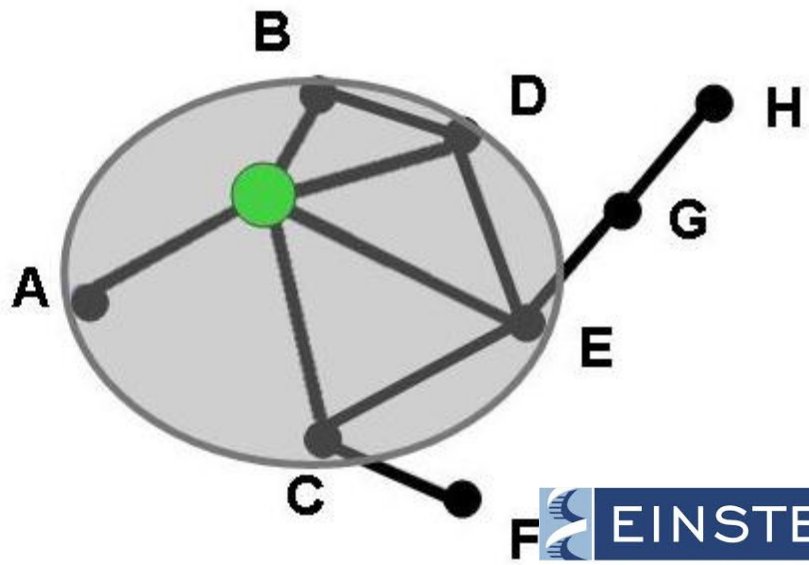
- How to identify <u>high-scoring alignments</u>?
  - Greedy ***seed and extend*** approaches
    - Use the most "similar" nodes across the two networks as "anchors" or **"seed nodes"**
    - "Extend around" the seed nodes in a greedy fashion



Network Alignment

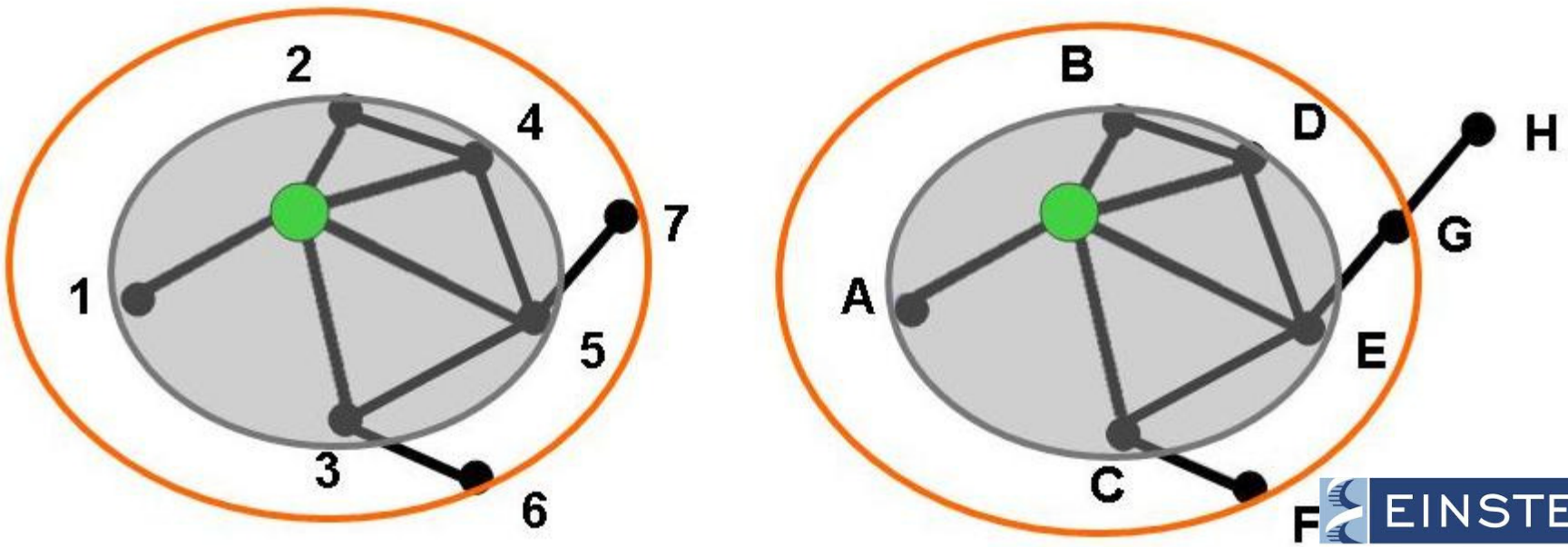
- How to identify <u>high-scoring alignments</u>?
  - Greedy ***seed and extend*** approaches
    - Use the most "similar" nodes across the two networks as "anchors" or **"seed nodes"**
    - "Extend around" the seed nodes in a greedy fashion





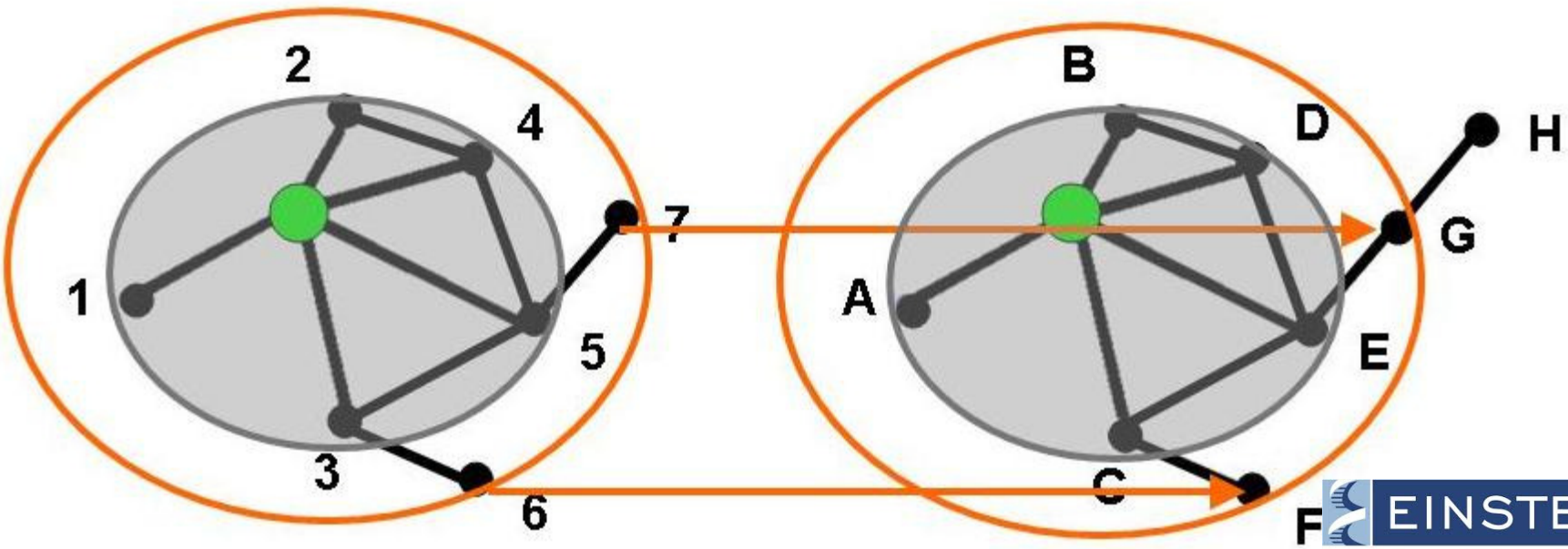
Network Alignment

- How to identify <u>high-scoring alignments</u>?
  - Greedy ***seed and extend*** approaches
    - Use the most "similar" nodes across the two networks as "anchors" or **"seed nodes"**
    - "Extend around" the seed nodes in a greedy fashion



Network Alignment

- How to identify <u>high-scoring alignments</u>?
  - Greedy ***seed and extend*** approaches
    - Use the most "similar" nodes across the two networks as "anchors" or **"seed nodes"**
    - "Extend around" the seed nodes in a greedy fashion



EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Take home message

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
  - Different topologies of network
  - Different type of network comparison
  - Basic ideas of network alignment
- Structural modeling of PPI
- Physical properties of PPI

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
  - Protein-protein docking
  - Template-based modeling
- Physical properties of PPI

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

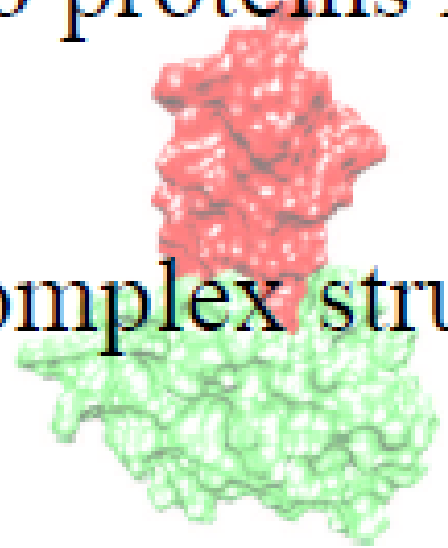
Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
  - Protein-protein docking
  - Template-based modeling
- Physical properties of PPI

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

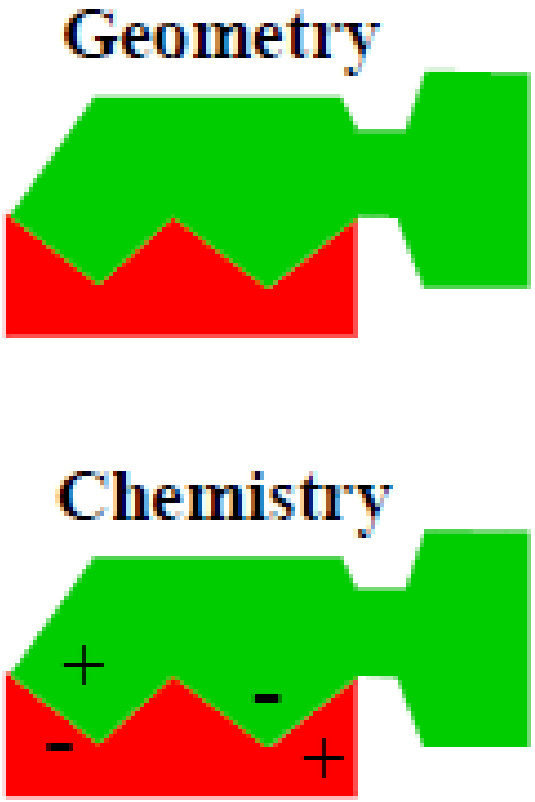
Protein-Protein Docking

- Given two proteins **A** and **B**



- Predict complex structure **AB**

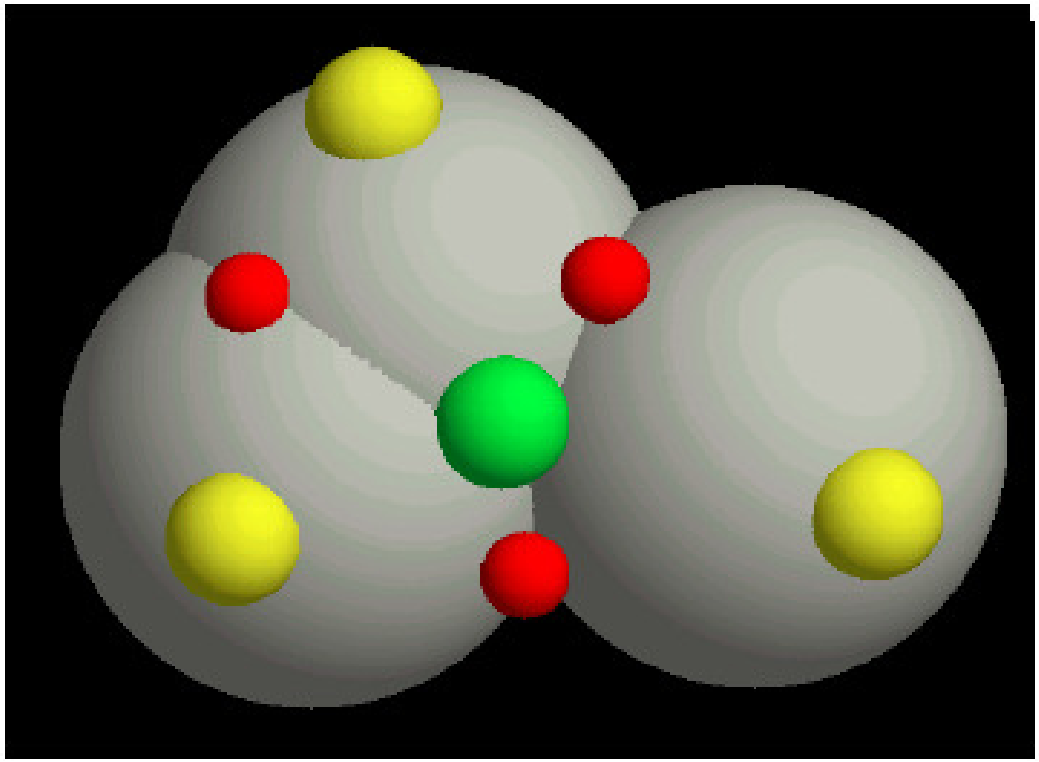
Lock-and-Key Principle



Docking Algorithm Scheme

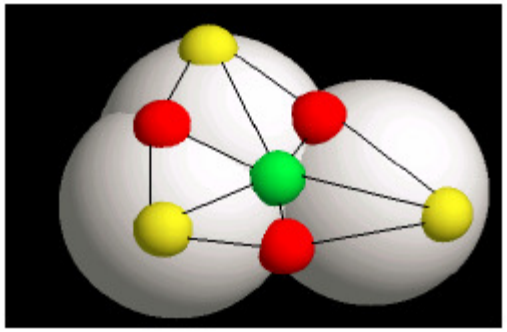
- Part 1: **Molecular surface representation**
- Part 2: Features selection
- Part 3: Matching of critical features
- Part 4: Filtering and scoring of candidate transformations

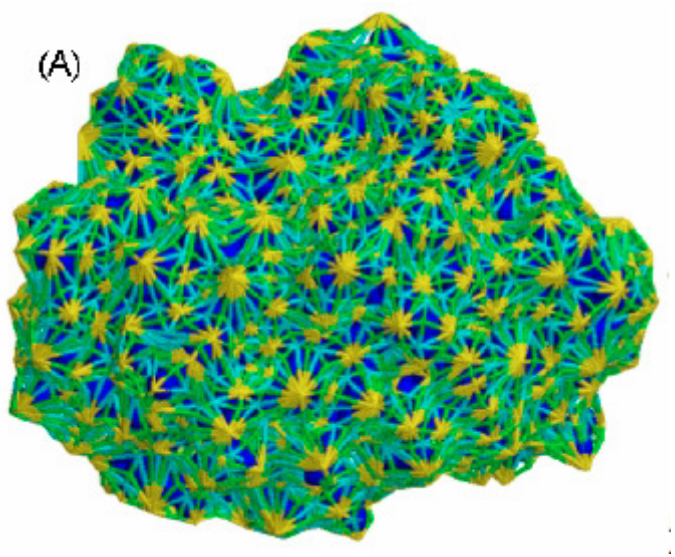
1. Surface Representation



Sparse Surface Graph – G_{top}

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





- G_{top} – Surface topology graph:

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Docking Algorithm Scheme

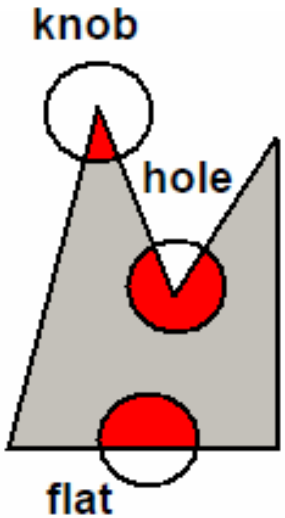
- Part 1: Molecular surface representation
- **Part 2: Features selection**
- Part 3: Matching of critical features
- Part 4: Filtering and scoring of candidate transformations

2.1 Coarse Curvature calculation
2.2 Division to surface patches of similar curvature

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 (70%).

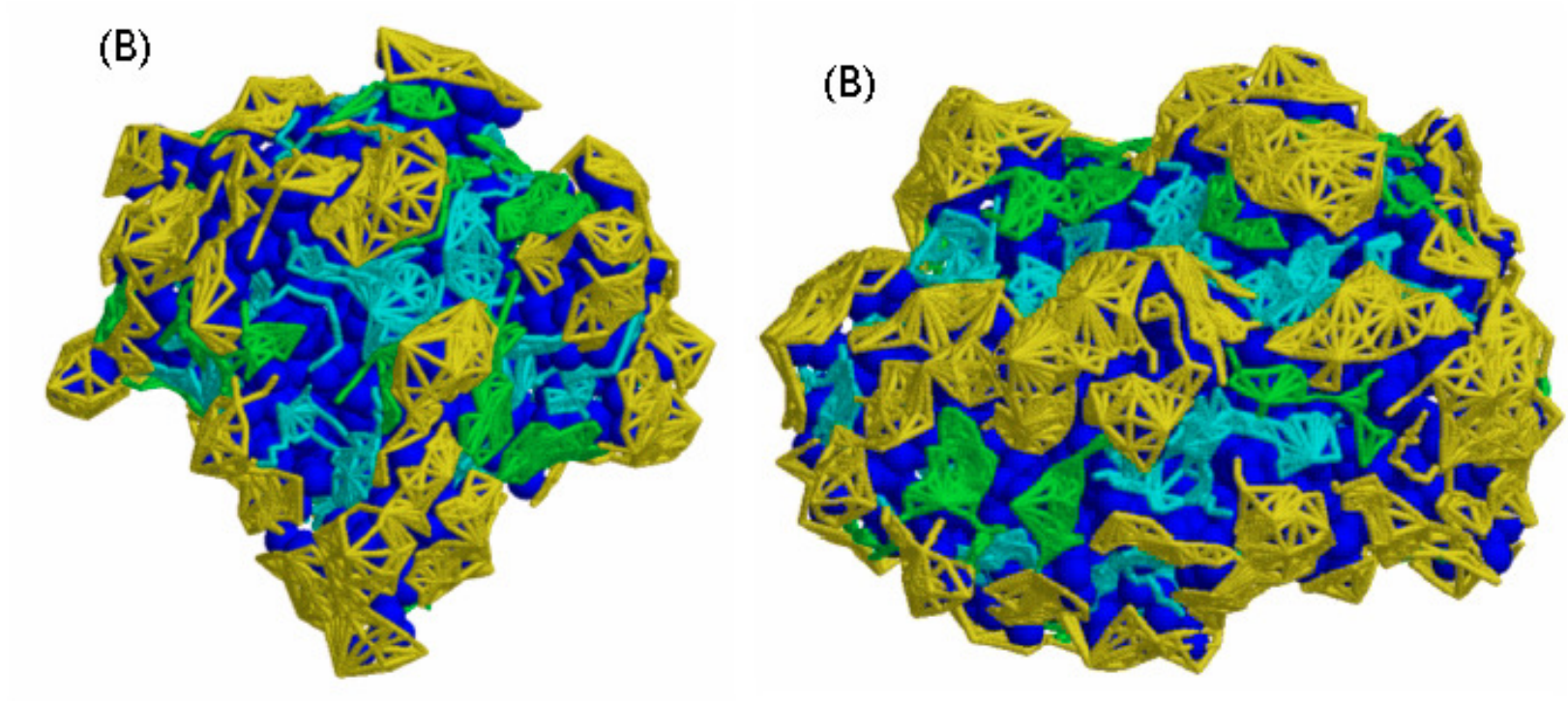


2.2 Patch Detection

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

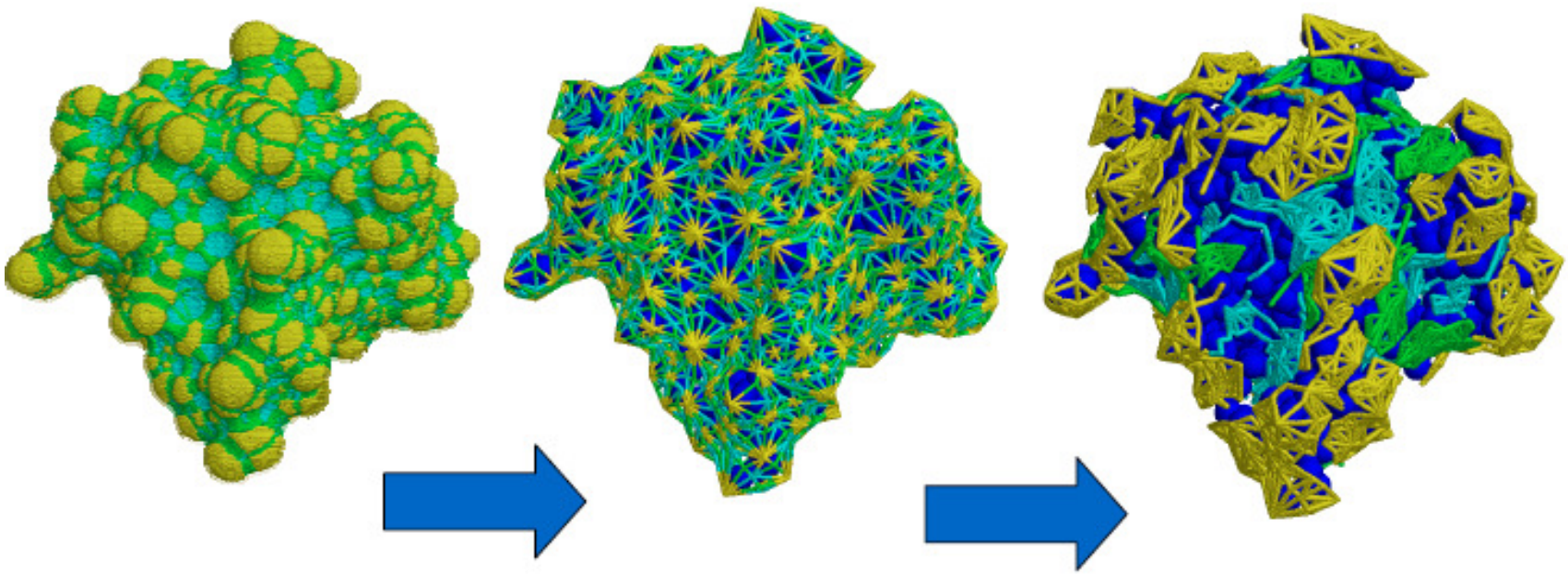
- **connected** – the points of the patch correspond to a connected sub-graph of G_{top}.
- **similar curvature** – all the points of the patch correspond to only one type: knobs, flats or holes.
- **equal sized** – to assure better matching we want shape features of almost the same size.

Examples of Patches for trypsin and trypsin inhibitor



Yellow – knob patches, cyan – hole patches, green – flat patches

Shape Representation Part



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

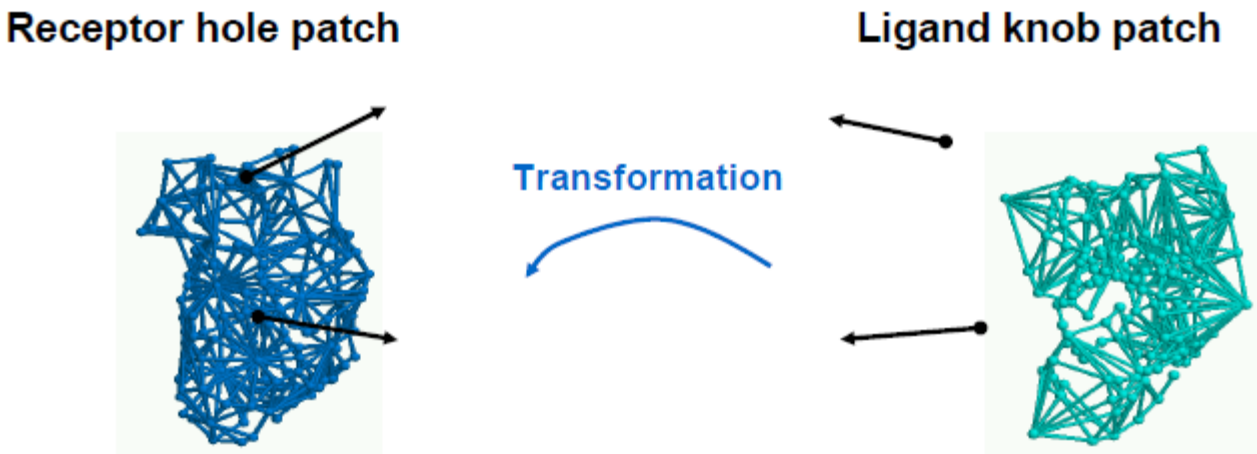
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

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:

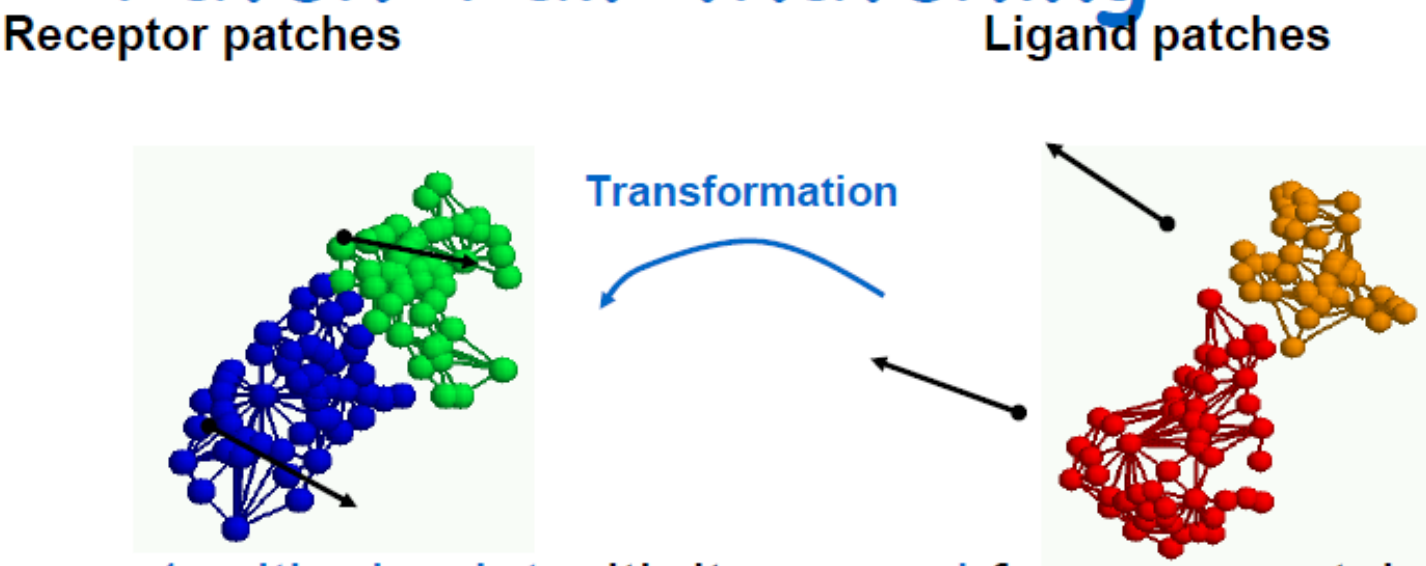
- **Single Patch Matching** – one patch from the receptor is matched with one patch from the ligand. Used in protein-drug cases.
- **Patch-Pair Matching** – two patches from the receptor are matched with two patches from the ligand. Used in protein-protein cases.

Single Patch Matching



- **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.

Patch-Pair Matching

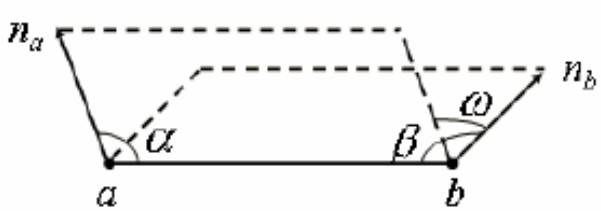


- **Base:** 1 critical point with its normal from one patch and 1 critical point with its normal from a neighboring patch.
- Match every base from the receptor patches with all the bases from complementary ligand patches.
- Compute the transformation for each pair of matched bases.

Base Compatibility

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



dE, dG, α, β, ω

Two bases are compatible if their signatures match

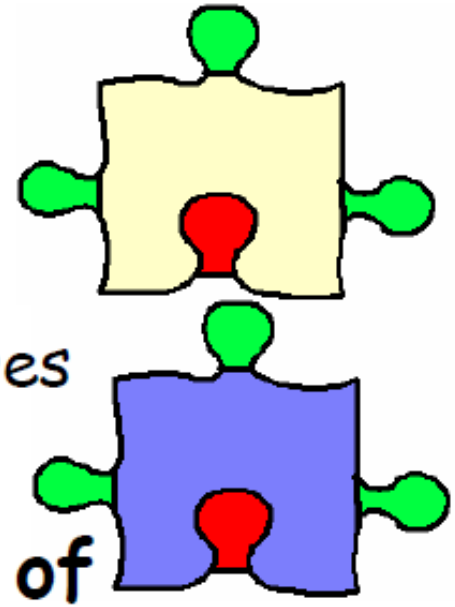
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Geometric Hashing

- **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: Features 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.

Scoring Shape Complementarity

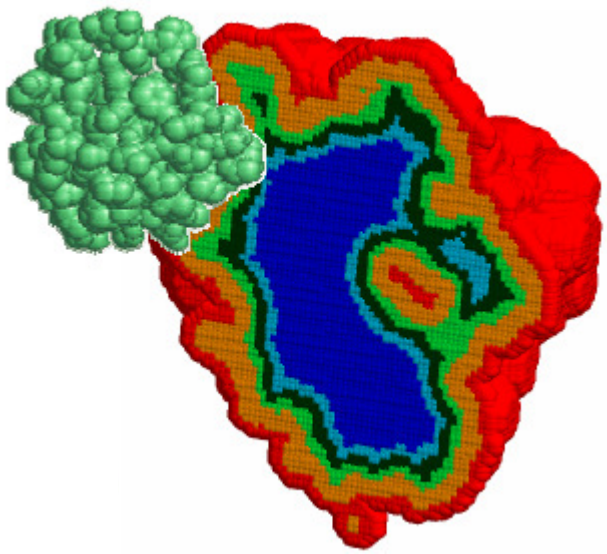
- The scoring is necessary to rank the remaining solutions.
- The surface of the receptor is divided into five shells according to the distance function: **S1 - S5**

[-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:



EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

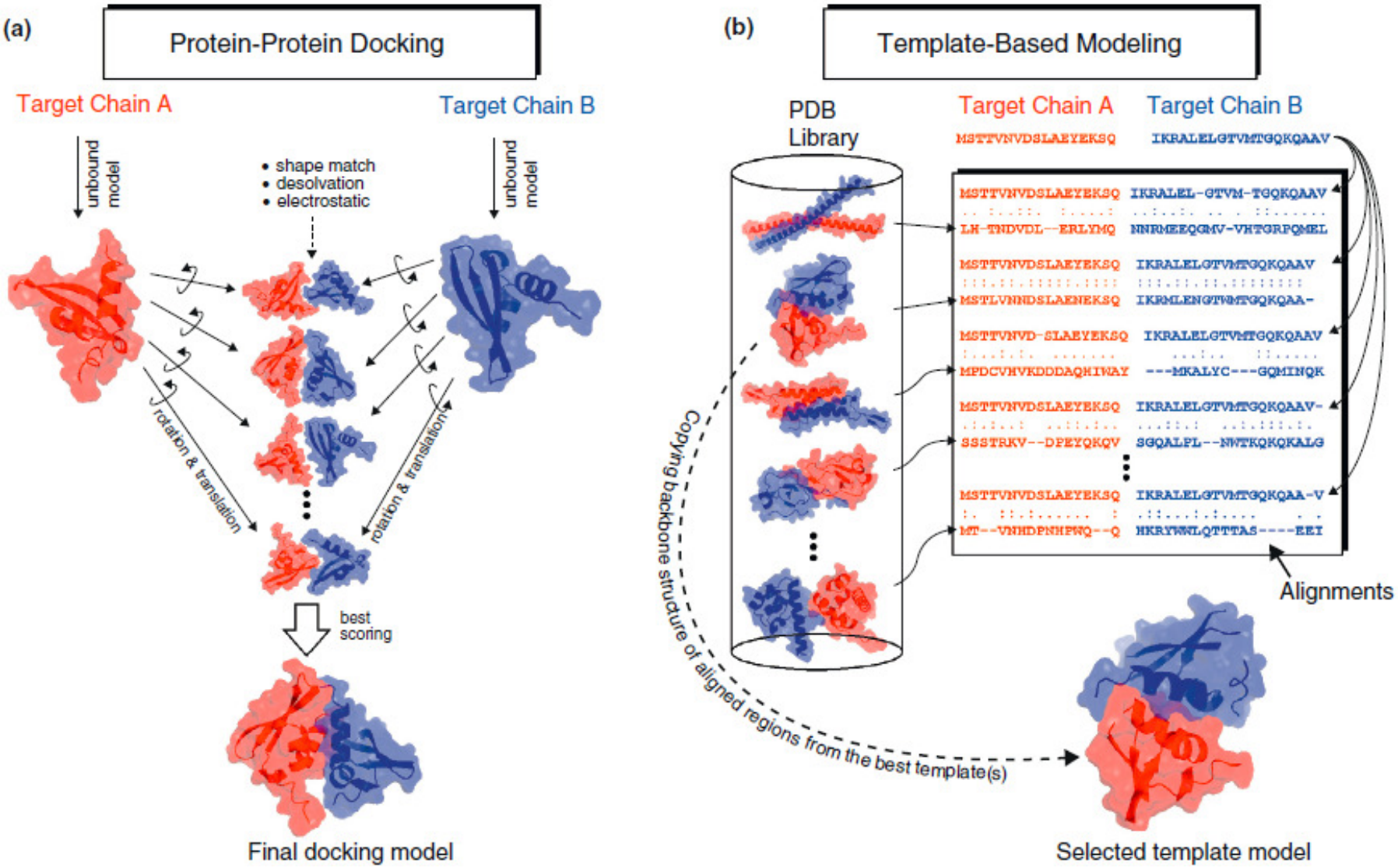
Flexible Docking - general methodology

- Rigid subpart docking :
  - Split the flexible molecule into rigid subparts.
  - Dock independently each subpart.
  - Pair the top hypotheses for each subpart to detect hinge consistency.
- Anchor fragment method :
  - Position a ‘preferred’ anchor fragment.
  - Rotate sequentially the flexible bonds to position the other fragments.

Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
  - Protein-protein docking
  - Template-based modeling
- Physical properties of PPI

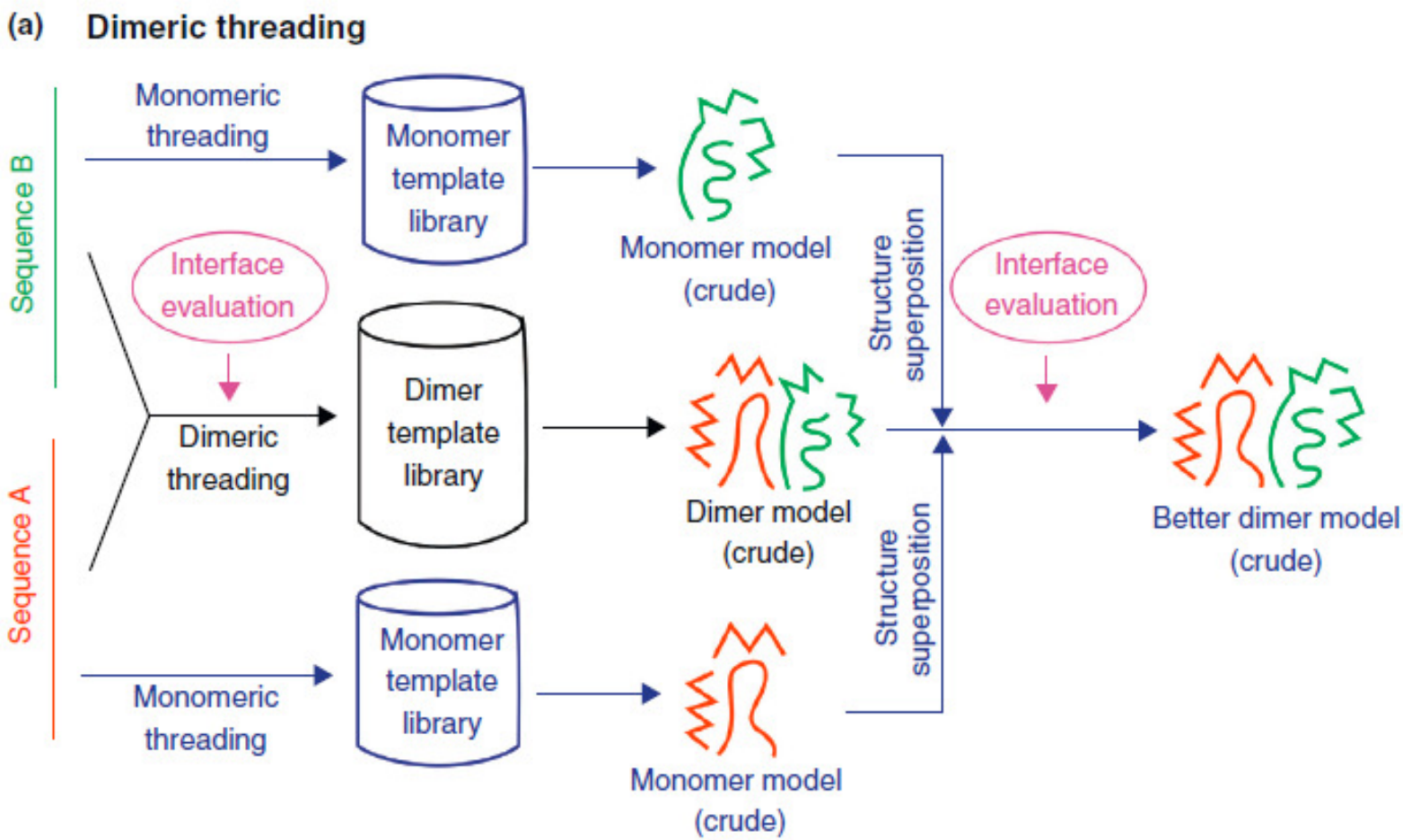
EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

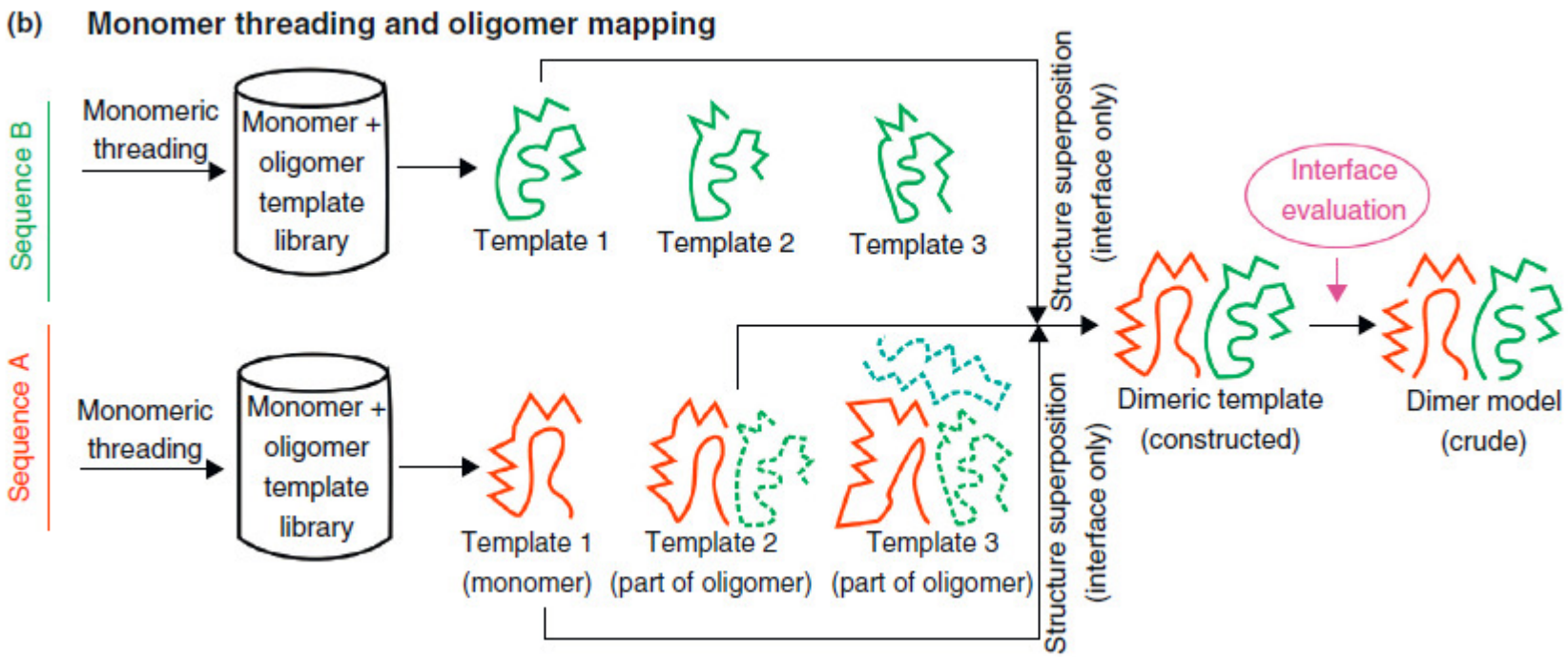


EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

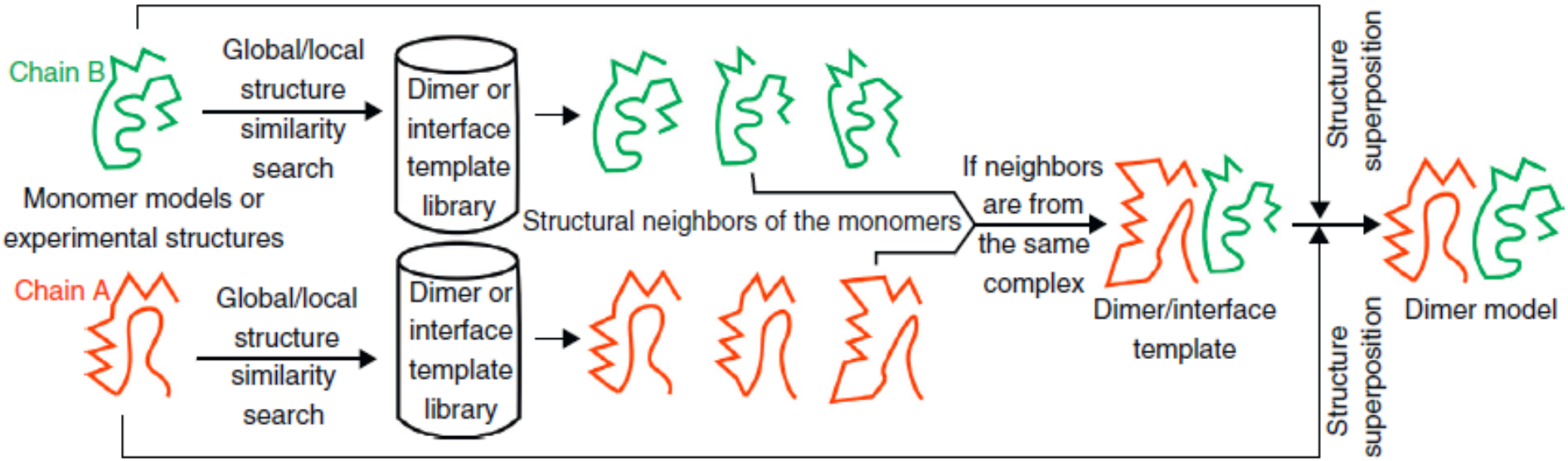
Template-based modeling: general methodology

- Dimeric threading
- Monomer threading and oligomer mapping
- Template-based docking





(c) **Template-based docking**



Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI
  - Kinetic rates
  - Binding affinity

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI
  - Kinetic rates
  - Binding affinity

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Kinetic parameters

The speed at which a complex AB dissociates is determined by its **dissociation rate constant** k_{dissoc} (s^{-1}):

$$k_{\text{dissoc}}\ [\text{AB}] = \frac{-d[\text{AB}]}{dt}$$

The speed at which a complex AB forms is determined by its **association rate constant** k_{assoc} ($M^{-1}s^{-1}$):

$$k_{\text{assoc}}[\text{A}][\text{B}] = \frac{+d[\text{AB}]}{dt}$$

At equilibrium: $d[AB]/dt = 0$

$$k_{dissoc}\,[AB] = k_{ass}\,[A][B]$$

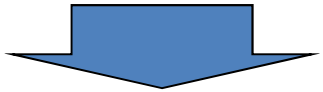
$$k_{dissoc}/k_{assoc} = [A][B]/[AB] = K_d$$

Surface plasmon resonance (SPR)

- SPR measures the change of the refractive index at the backside of a metal film when protein A binds to protein B immobilized on this film
- Using SPR, one can determine K_d, k_{assoc} and k_{dissoc}

Brownian Dynamics (BD)

- The dynamic contributions of the solvent are incorporated as a dissipative random force (Einstein's derivation on 1905). Therefore, water molecules are not treated explicitly.
- Since BD algorithm is derived under the conditions that solvent damping is large and the inertial memory is lost in a very short time, longer time-steps can be used.



- BD method is suitable for long time simulation.

Algorithm of BD

The Langevin equation can be expressed as

$$m_i \frac{\mathrm{d}^2 \mathbf{r}_i}{\mathrm{d}t^2} = -\zeta_i \frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} + \mathbf{F}_i + \mathbf{R}_i \qquad \textbf{(1)}$$

Here, $\mathbf{r}_i$ and m_i represent the position and mass of atom i, respectively. ζ_i is a frictional coefficient and is determined by the Stokes' law, that is, $\zeta_i = 6\pi a_i^{\mathrm{Stokes}}\eta$ in which a_i^{Stokes} is a Stokes radius of atom i and η is the viscosity of water. $\mathbf{F}_i$ is the systematic force on atom i. $\mathbf{R}_i$ is a random force on atom i having a zero mean $\langle \mathbf{R}_i(t) \rangle = 0$ and a variance $\langle \mathbf{R}_i(t)\mathbf{R}_j(t) \rangle = 6\zeta_i kT\delta_{ij}\delta(t)$; this derives from the effects of solvent.

For the overdamped limit, we set the left of eq.1 to zero,

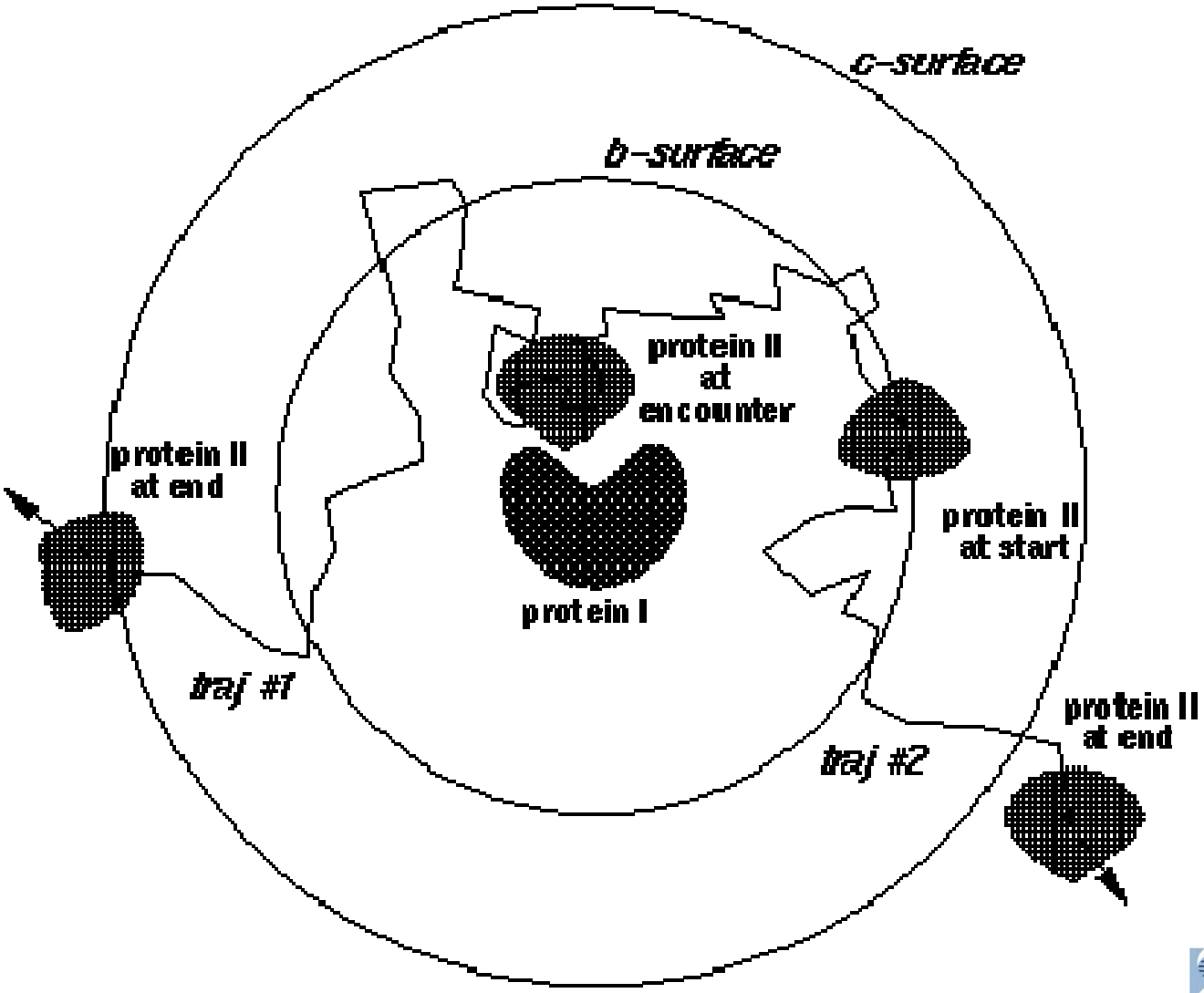
$$\zeta_i \frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} = \mathbf{F}_i + \mathbf{R}_i \qquad \textbf{(2)}$$

The integrated equation of eq. 8 is called Brownian dynamics;

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \frac{\mathbf{F}_i(t)}{\zeta_i}\Delta t + \sqrt{\frac{2k_{\mathrm{B}}T}{\zeta_i}\Delta t}\,\boldsymbol{\omega}_i \qquad \textbf{(3)}$$

where Δt is a time step and $\boldsymbol{\omega}_i$ is a random noise vector obtained from Gaussian distribution.

Brownian dynamic simulation of protein association



EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Outline

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI
  - Kinetic rates
  - Binding affinity

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Equilibrium parameters

The strength of an interaction is usually given as the **equilibrium dissociation constant**, K_d:

$$K_d = \frac{[A][B]}{[AB]}$$

Relationship between K_d and Gibbs free energy change ΔG upon binding

$$\Delta G = \Delta G^0 + RT \ln [AB]/[A][B]$$

Under equilibrium conditions ($\Delta G = 0$):

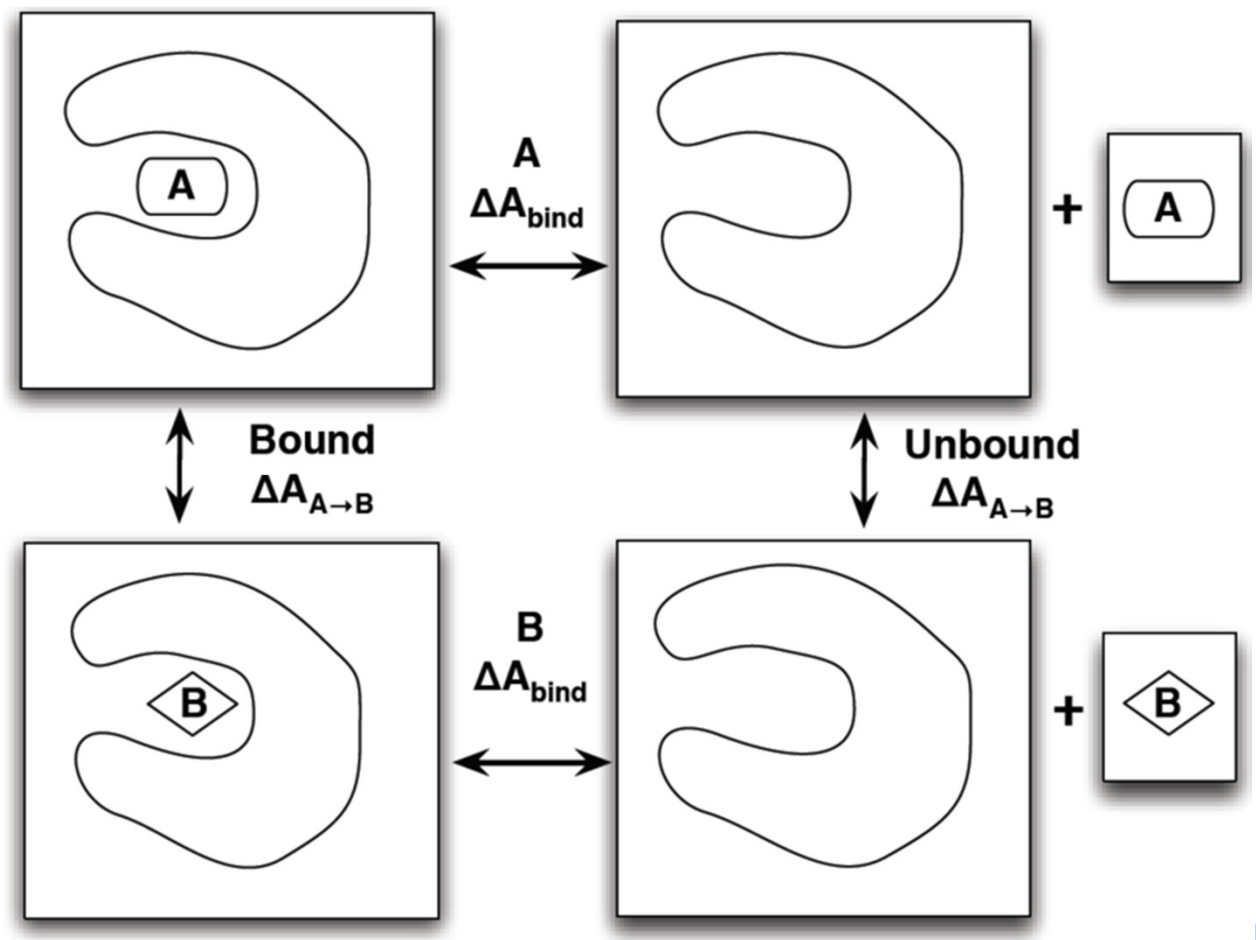
$$\Delta G^0 = -RT \ln \frac{[AB]}{[A][B]}$$

$$\Delta G^0 = -RT \ln K_a = -RT \ln\left(\frac{1}{K_d}\right) = RT \ln K_d$$

Isothermal titration calorimetry (ITC)

- ITC measures ΔH, which is the heat that is released or absorbed when the complex AB associates from A and B
- Using ΔH as the binding signal, one can determine K_d, the reaction stoichiometry (n), and the reaction entropy (ΔS)

Computational simulation of binding affinity: thermodynamic cycles



EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Summary

- Binary prediction of Protein-protein Interaction (PPI)
- Analysis of PPI networks
- Structural modeling of PPI
- Physical properties of PPI