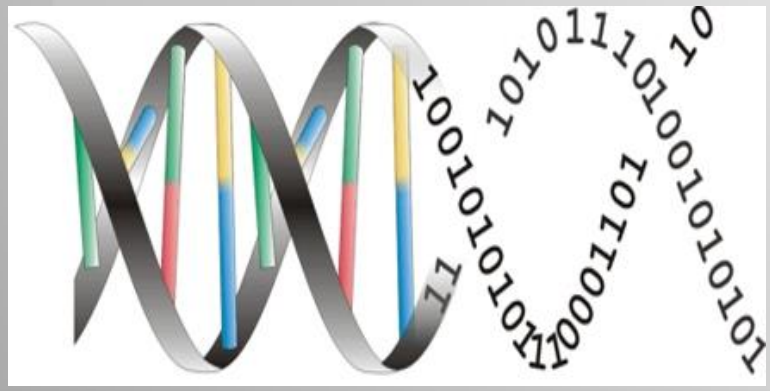
The University of Georgia

The University of Georgia

Programming for Computational & Systems Biology

Instructor: Paul Xie

Tue. & Thr. 9:35~10:50



Course Format

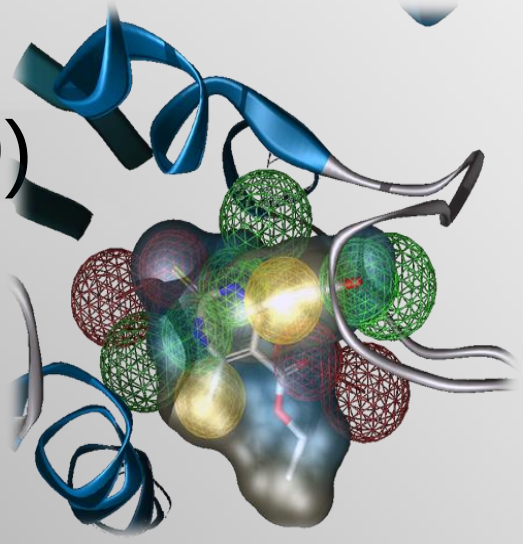
The University of Georgia

- Computer labs + lectures (3 hours/week)
- Coding concepts (some knowledge about biological data)
- 6-8 assignments, published on eLC, (30-40%)
  - Please upload them by the due days
- Paper review (10-20%)
- 1 term paper (50-60%)

The University of Georgia

Docking/Simulation

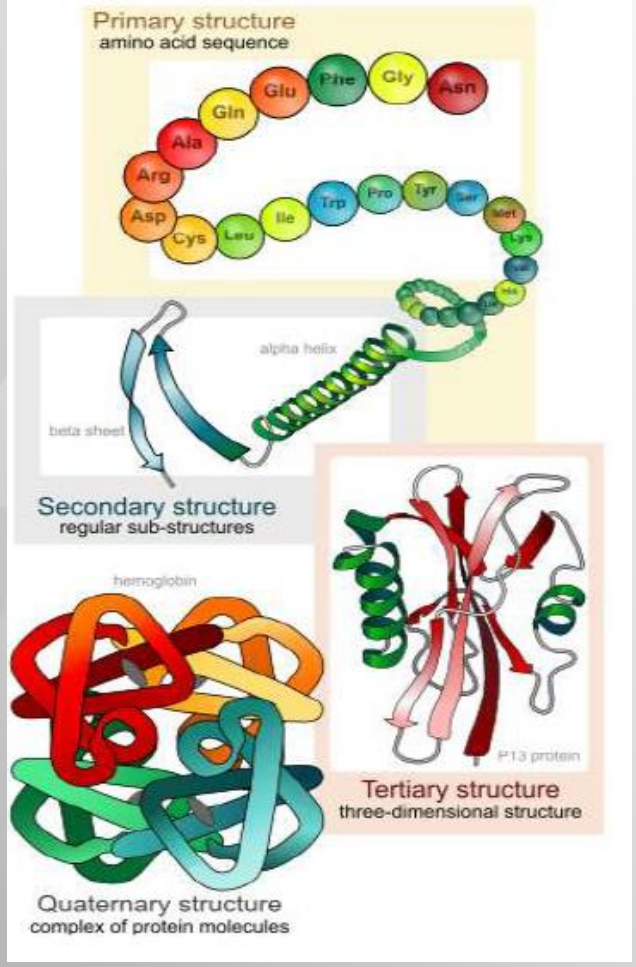
Instructor: Paul Xie (9)



Last Week

The University of Georgia

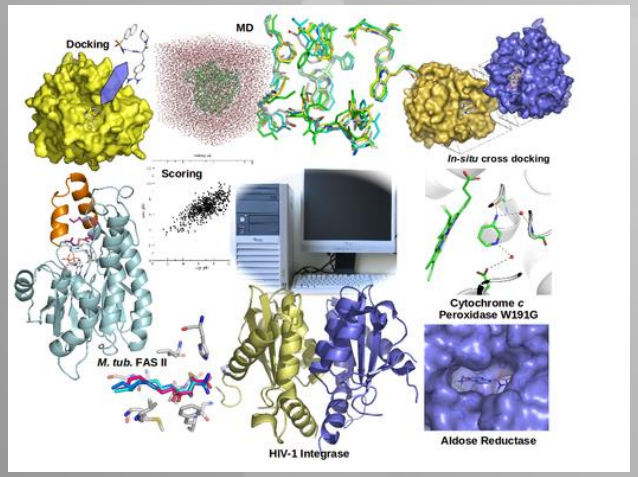
- Protein
- List & List Operation



This Week

The University of Georgia

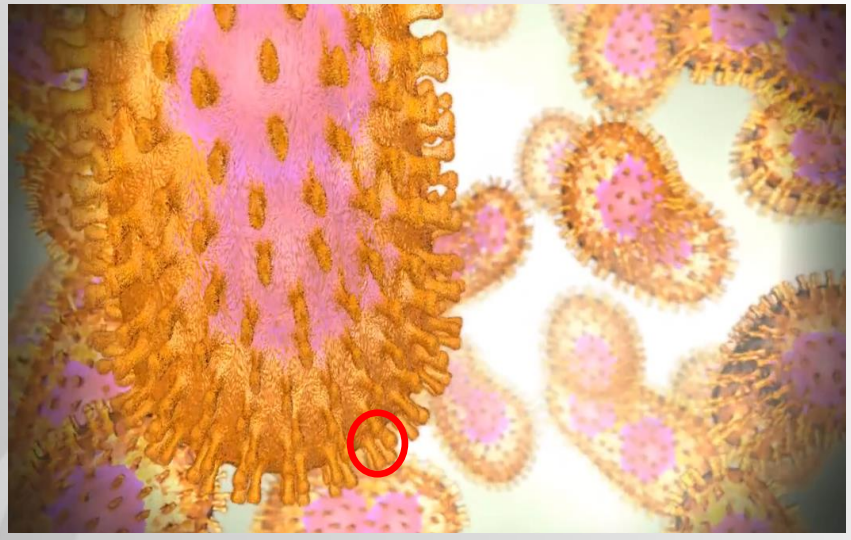
- Molecular docking
  - Virtual drug discovery
- List Operation → Modeling

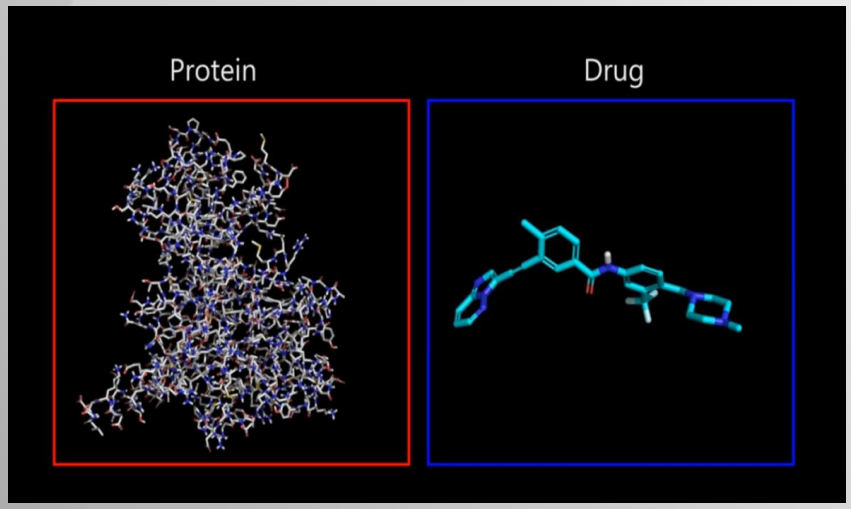


Drug (Medicine) Discovery

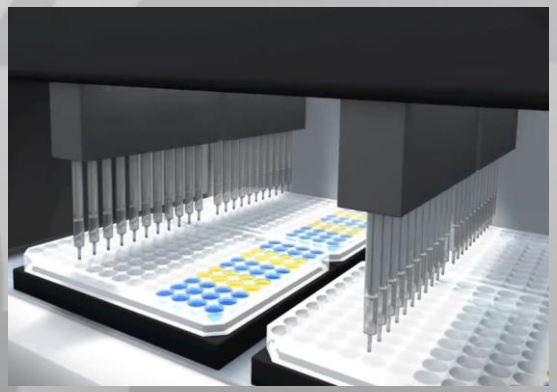
The University of Georgia

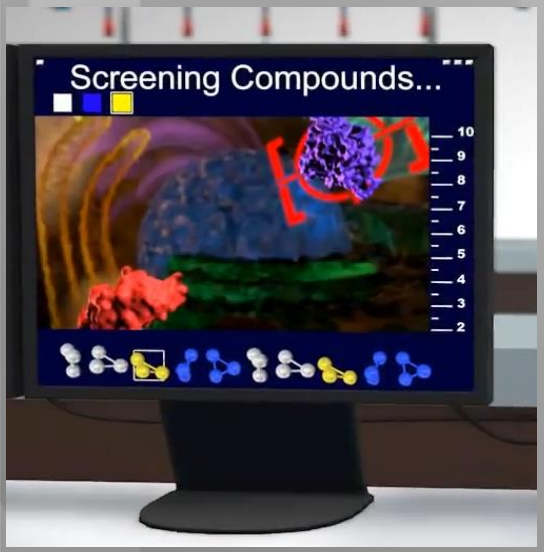
- How to discover a drug?
  - Identify Targets
    - Proteins (or DNA/RNA)
  - Screen Drugs
    - Molecules can bind and regulate the activity of the target protein





Conventional vs. Computational Drug Discovery



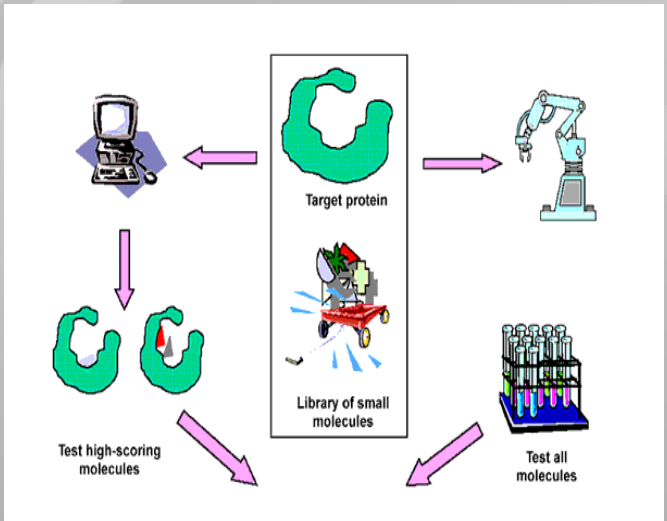


- Challenges of new drug discovery
- Comparison

	Conventional	Computational
Venue	Labs	CPUs
Target Proteins	Culture, purify, ...	Download
Millions of compounds	Purchase, deliver, store,...	Download
Drug binding	Bio-essays	Software
Operators	Labors or machine	CPU & software

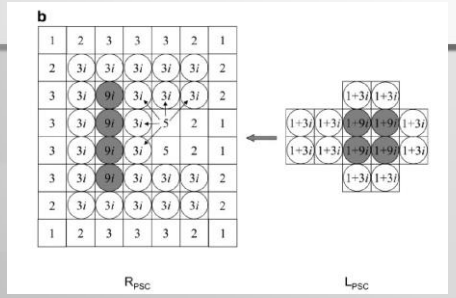
- Advantages of Computational Drug Discovery
  - Faster
  - Cheaper
  - More

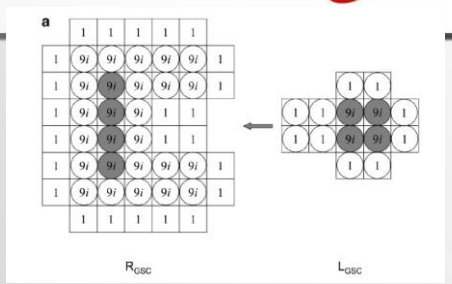
Docking



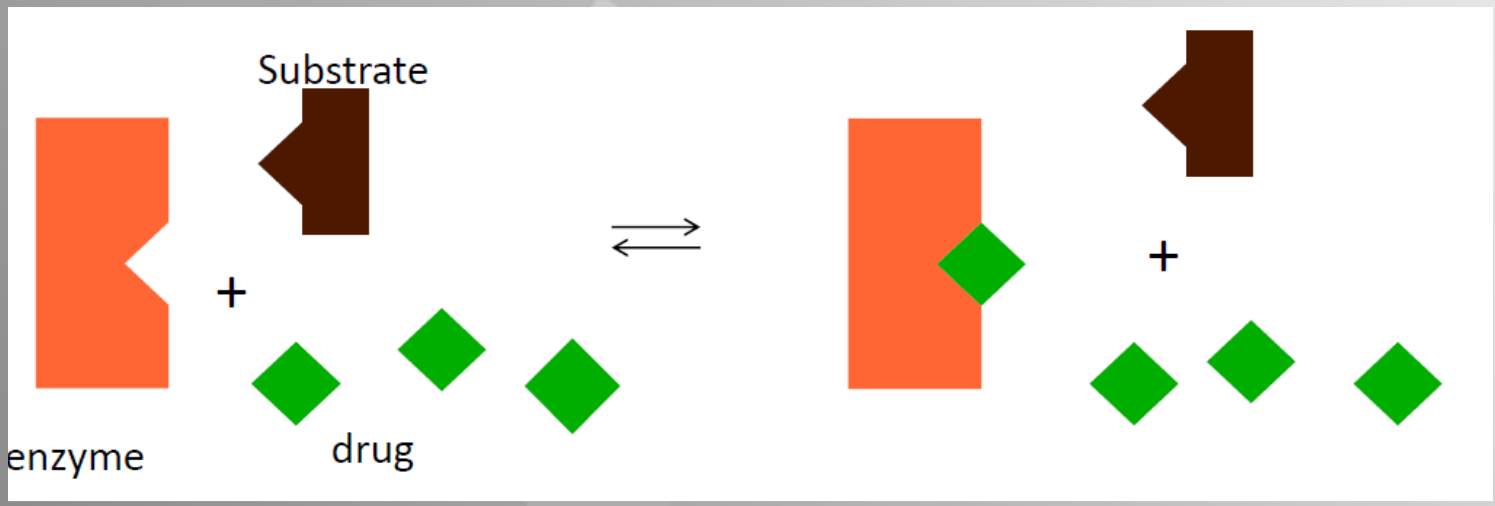
- To predict whether and how well one molecule bound to a second one to form a stable complex
- Application on medicinal chemistry & molecular biology
- Virtual screening

Basic Concept





- Shape complementarity
- Thermodynamic equilibrium $K_a = \frac{[A][B]}{[AB]} = \frac{[Protein][Ligand]}{[Complex]}$
- Affinity: $\Delta G = -RTlnK_a$



What are Key Components of Docking?

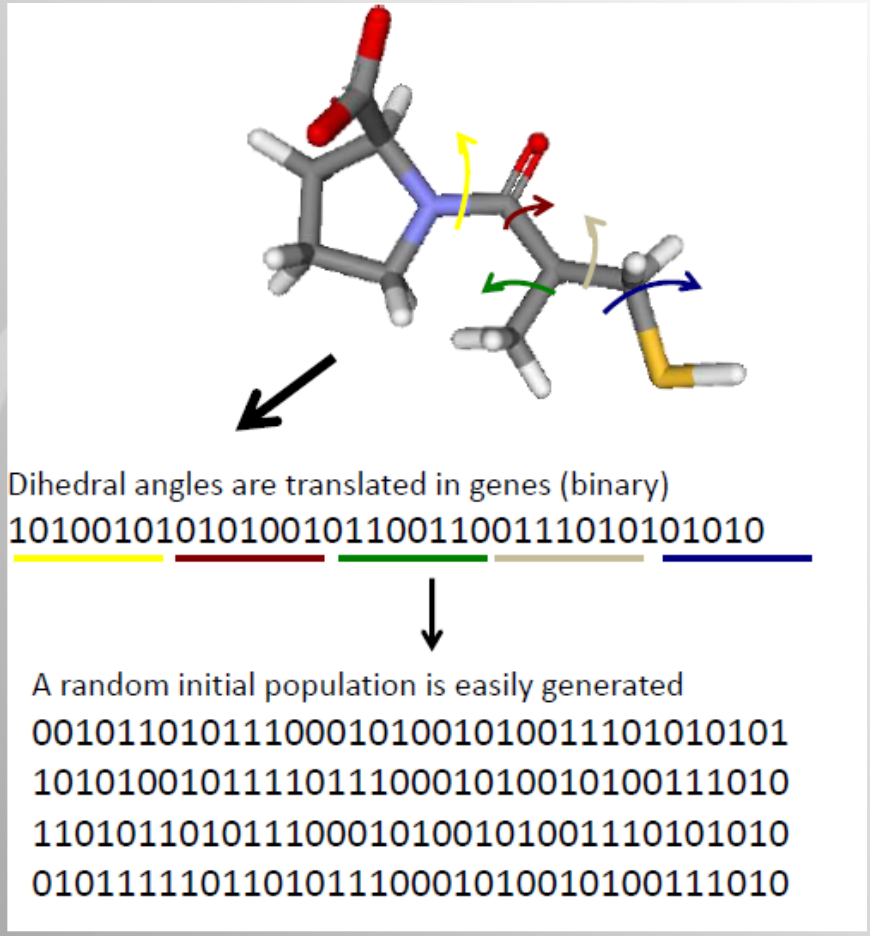
Scoring Functions of Docking

- Two components: Sampling and Scoring
  1. Generates a large number of poses of a molecule in the binding site
  2. Calculates a score or binding affinity for a particular pose
- Energy-based scoring functions:
  1. Physics-based: force-field or empirical
  2. Knowledge-based: statistical data to approximate binding energy

Genetic Algorithm

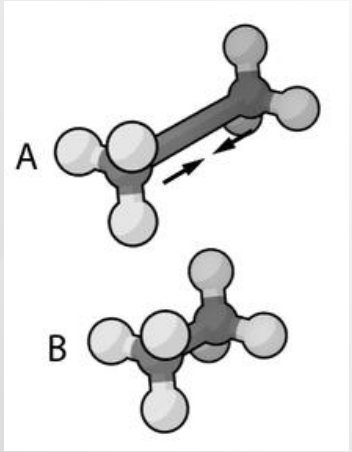
The University of Georgia

- Optimization
- To mimic evolution process
- Genes, individuals, and population
- Pose: location, orientation, conformation



Force Field

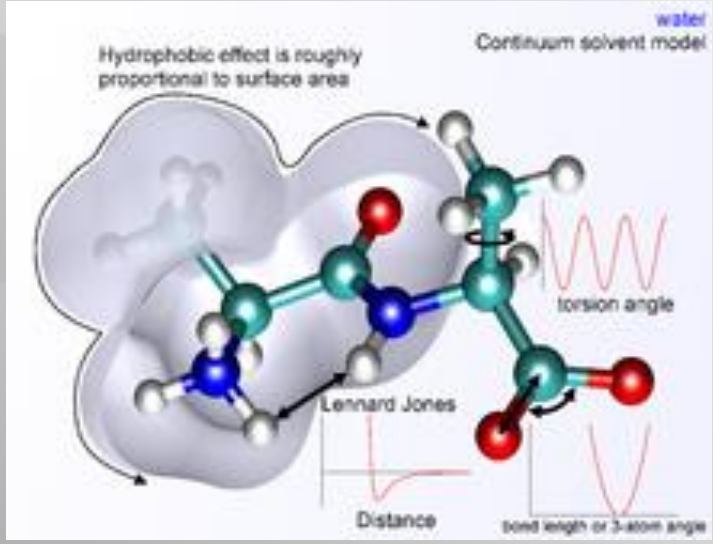
The University of Georgia



- Force-Field
  - a physical- or statistical model describing the interaction energy
  - $E_{total} = E_{bonded} + E_{nonbonded}$

$$E_{\text{bonded}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}}$$

$$E_{\text{nonbonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$



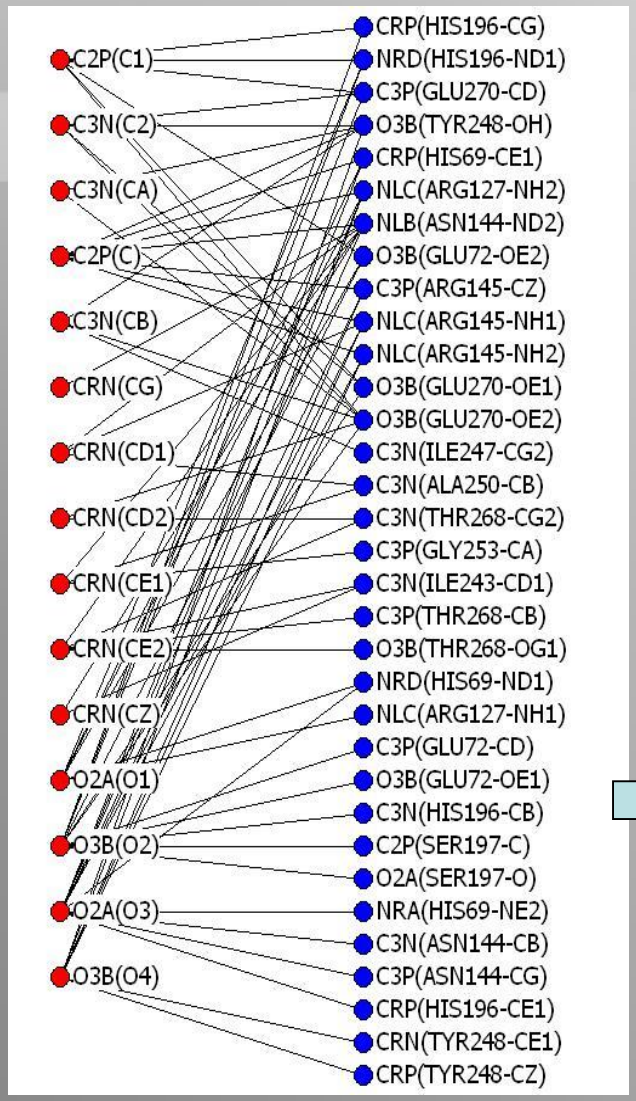
New Scoring Functions

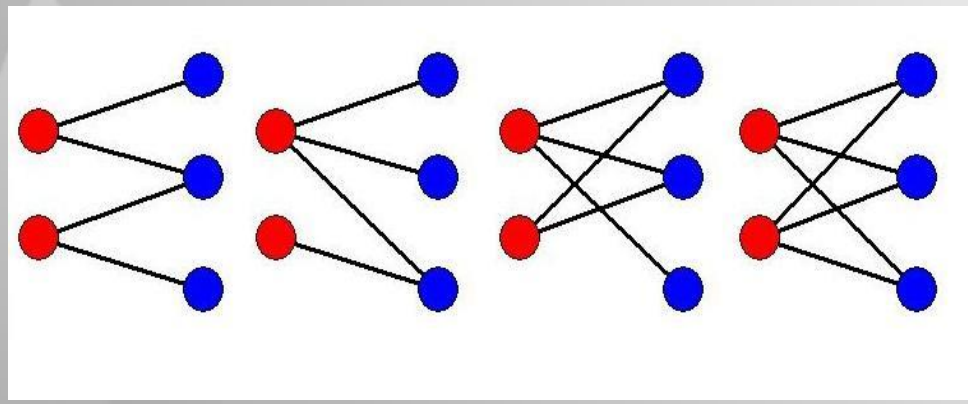
The University of Georgia

- How to create a new docking scoring function?
  - New approach
  - Overcome the difficulties, e.g. structure flexibility

From 3D Structures to Interaction Motifs

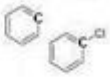
The University of Georgia



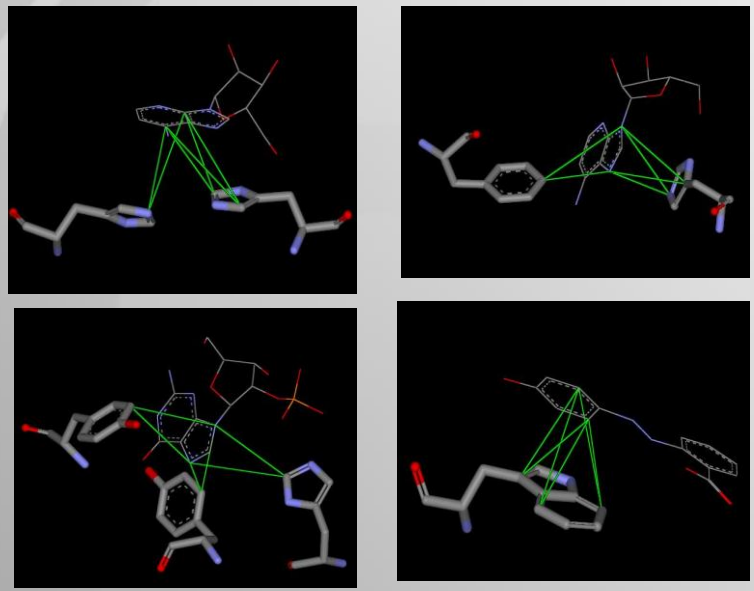


Xie & Hwang, *BMC Bioinformatics* 2010

MotifScore

Atom Type	Description	Example
C2N	C, SP2, normal/non-polar	—C=C—
C2P	C, SP2, polar	—C=O
C3N	C, SP3, normal/non-polar	—C—C—
C3P	C, SP3, polar	—C—OH
CRN	C, aromatic, normal/non-polar	
CRP	C, aromatic, polar	
O2A	O, SP2, hydrogen bond acceptor	>C=O
O3A	O, SP3, hydrogen bond acceptor	C—O—C
O3B	O, SP3, both	—OH
OLC	O, aliphatic, charge	--COO⁻
ORA	O, aromatic, hydrogen bond acceptor	
NLA	N, aliphatic, hydrogen bond acceptor	—NR_2
NLB	N, aliphatic, both	—$NH_{1\ or\ 2}$
NLC	N, aliphatic, charge	—NH_3^+
NRA	N, aromatic, hydrogen bond acceptor	
NRD	N, aromatic, hydrogen bond donor	
NRE	N, aromatic, either H.B. donor or acceptor	
S3N	S, SP2, normal/non-polar	—SH
SRN	S, aromatic, hydrogen bond acceptor	
SLC	S, aliphatic, charge	—SO_4^{-2}
P	P	
ME	metal	
HAL	F, Cl, Br, I	

- Automatic atom type assignment, graph converting, motif searching, & scoring
- Significant motifs



- 1st code : atom name (e.g., C, N, O, ME, HAL…)
- 2nd code: 2-SP2; 3-SP3; 1-SP; R-aromatic; L-aliphatic
- 3rd code: A-HB acceptor; D-HB donor; B-Both; N-normal/neither, E-either, C-charge P-polar

Xie & Hwang, *BMC Bioinformatics* 2010

Scoring Function

The University of Georgia

- Motif Gain ➔ Relative abundance of a specific motif

$$SG_i = \log\left(\frac{OM_i/N_i}{\prod_{j=1}^{5}(OA_j/M_j)}\right)$$

SG: Gain score

OM, OA: occurrence of motif/atom

N, M: number of specific type of motifs/atoms

$$Gain = \sum_{k=1}^{n} SG_k$$

- Penalty Function

$$Penalty = W \times NC$$

W: weight; *NC*: number of clash

- Overall Scoring Function

$$MotifScore = \frac{Gain}{Penalty+1}$$

Xie & Hwang, *BMC Bioinformatics* 2010

Comparison with Other Scoring Functions

Scoring function	Success rates (%) under different rmsd criteria[a]		
	≦1 Å	≦2 Å	≦3 Å
DrugScore^{CSD}	83	87	*[b]
MotifScore	72	84	86
Cerius2/PLP	63	76	80
SYBYL/F-Score	56	74	77
Cerius2/LigScore	64	74	76
DrugScore^{PDB}	63	72	74
Cerius2/LUDI	43	67	67
X-Score	37	66	74
AutoDock	34	62	72
Cerius2/PMF	40	52	57
SYBYL/G-Score	24	42	56
SYBYL/ChemScore	12	35	40

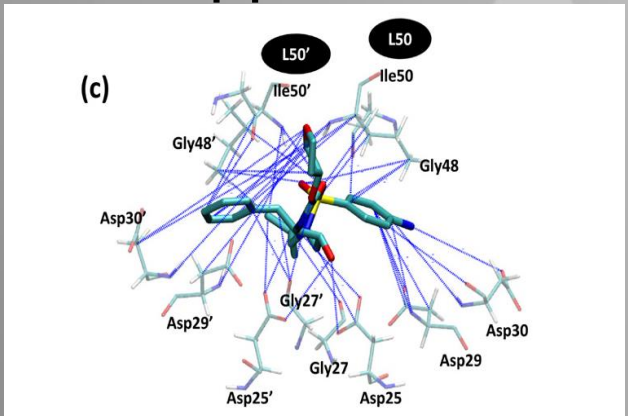
- Comparable to the best success rate of energy-based scoring functions
- To use the small molecule structure database, CSD
- Potential of Improvement

Xie Z.R.& Hwang M.J., *BMC Bioinformatics* 2010

New Scoring Function

The University of Georgia

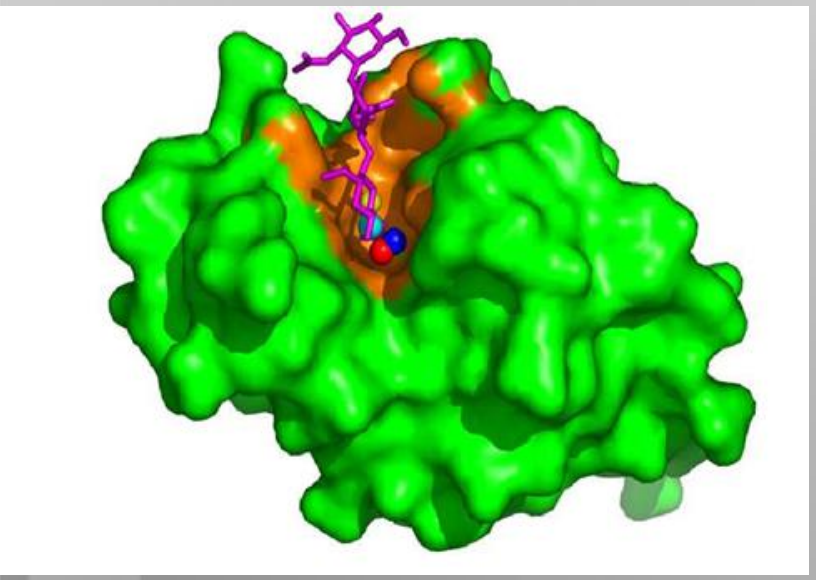
- Created a new docking scoring function
  - The first scoring function using network approach
  - More tolerant to structural flexibility
  - Good results compared with conventional approaches
  - Complementary with energy-based scoring functions
  - Has been applied to a study of drug resistance of HIV



Ghemitio, L. et al., *Health & Medical Informatics,* 2011

Before Docking

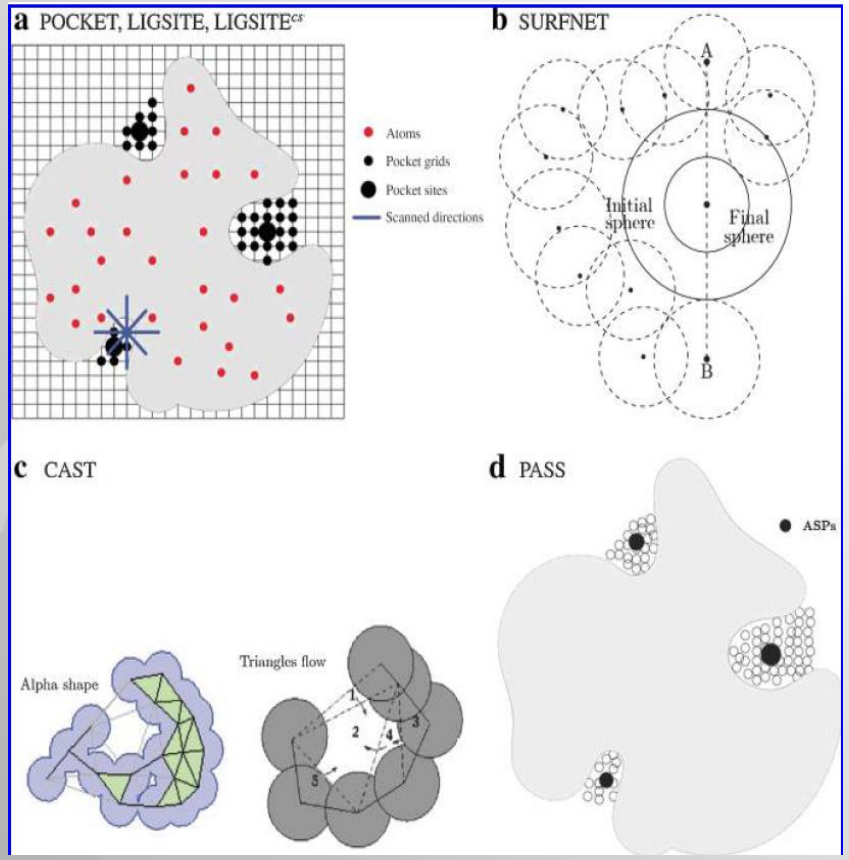
The University of Georgia



- Where is the binding site?
- Binding (Active) Site Prediction
  - To reduce computation load, and narrow down searching area on the protein
  - To predict protein functions

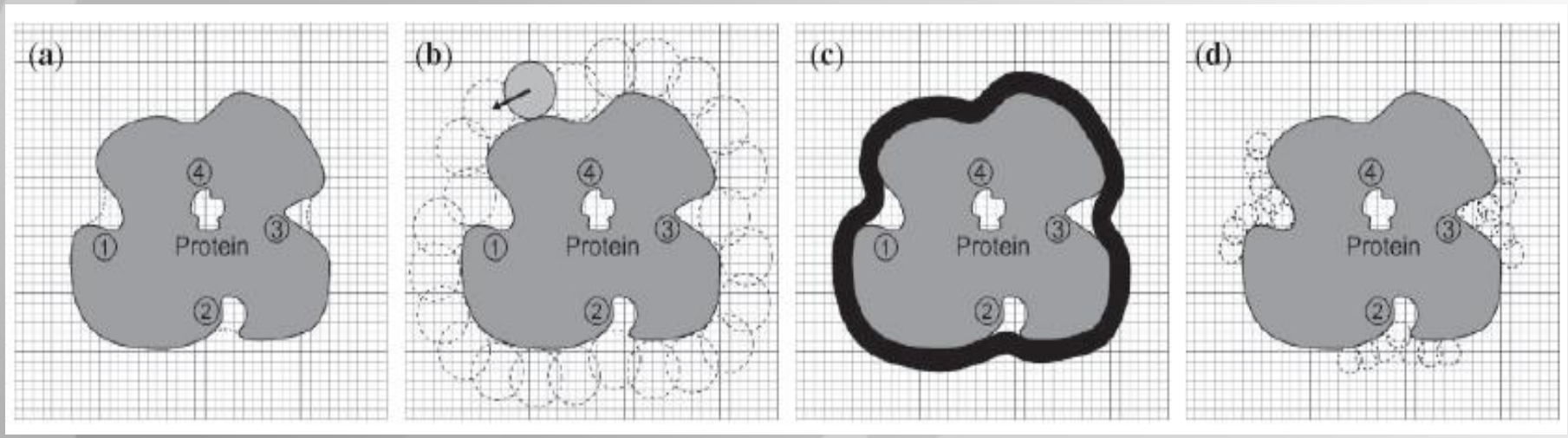
Huang, B. et al , *BMC Sturctural Biology* 2006

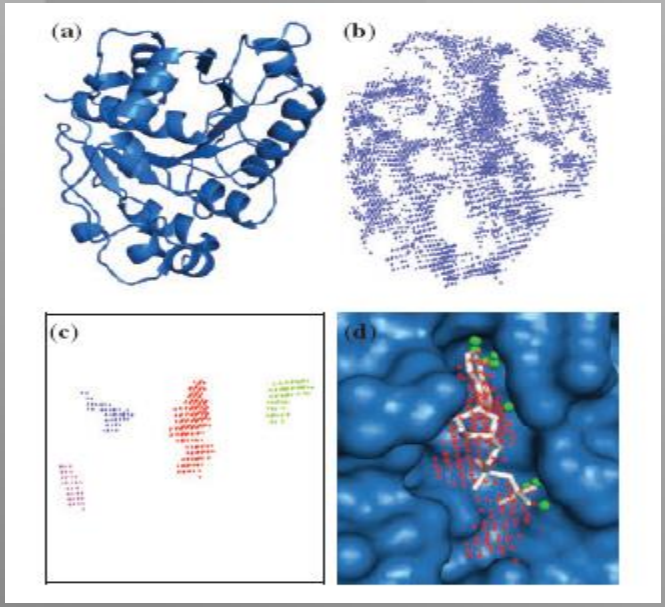
- From an observation
  - Small compounds often bind to largest cavities
  - New assumption
  - Geometry-based binding site prediction: to identify largest cavities
- Limitation: This assumption is often, but not always true



Liang, J., 1998, Brady, GP. et al, 2000, Huang, B. et al., 2006, Levitt, DG. et al.1992

POCASA





- Pocket: between protein & probe surfaces
- Diff. spheres create diff. cavities
- SPF: single point flag to remove noise

Yu, J. et al., *Bioinformatics* 2010

Limitation of Geometry-based Method

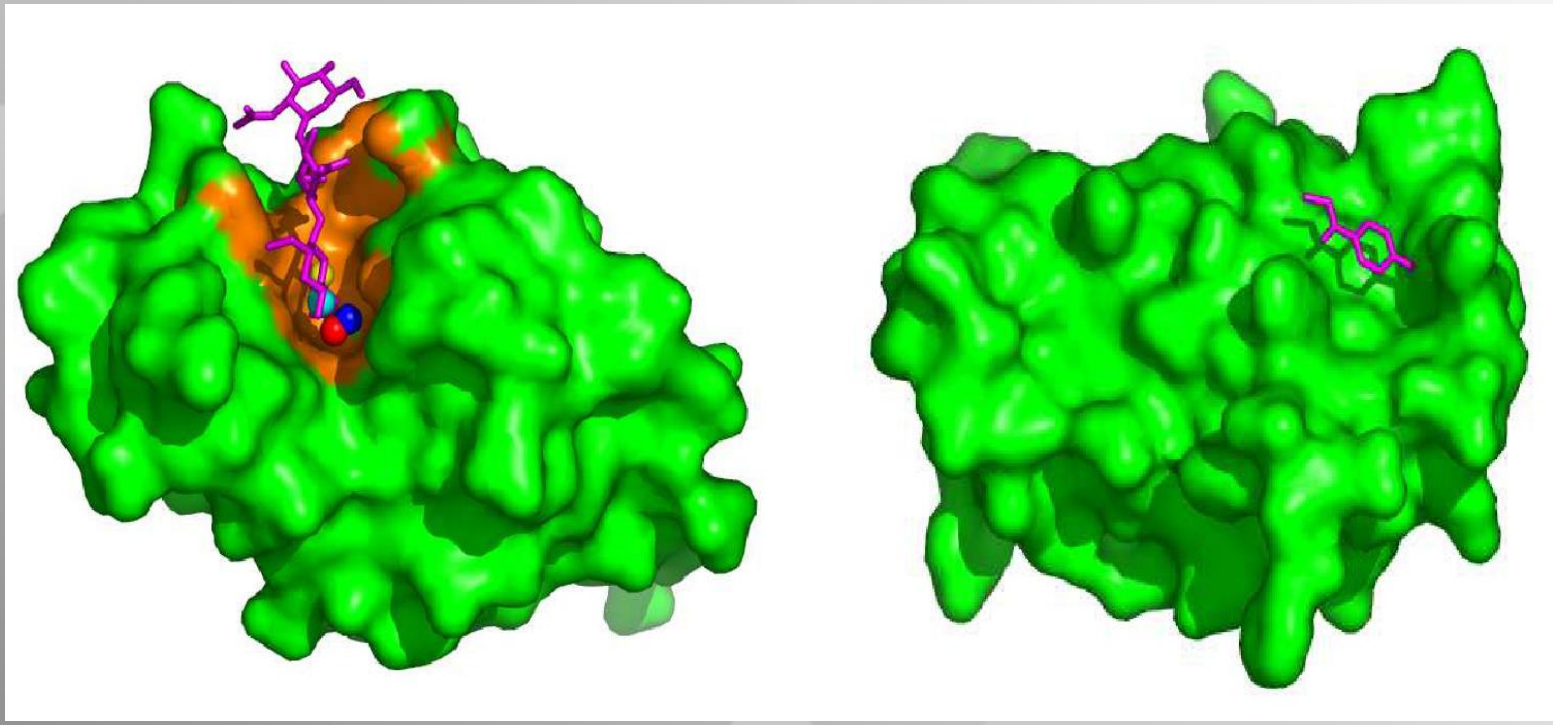


Table 3: Numbers of protein in each class for 210 bound structures.

Class	No. of proteins (as %)	Avg no. pocket points	Stdev
Class 1: Binding site in largest pocket	141/210 = 67%	209	185
Class 2: Binding site in second largest pocket	28/210 = 13%	66	64
Class 3: Binding site in third largest pocket	14/210 = 7%	40	41
Class 4: Binding site in none of above	27/210 = 13%		

Huang, B. et al , *BMC Sturctural Biology* 2006

Challenge

The University of Georgia

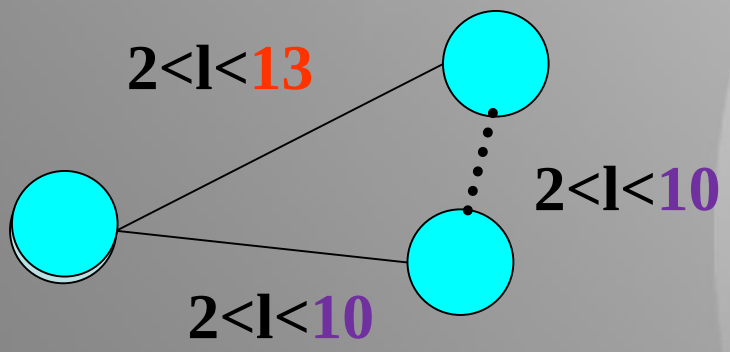
- How to create a new binding site prediction method?
  - New approach
  - Overcome the limitation

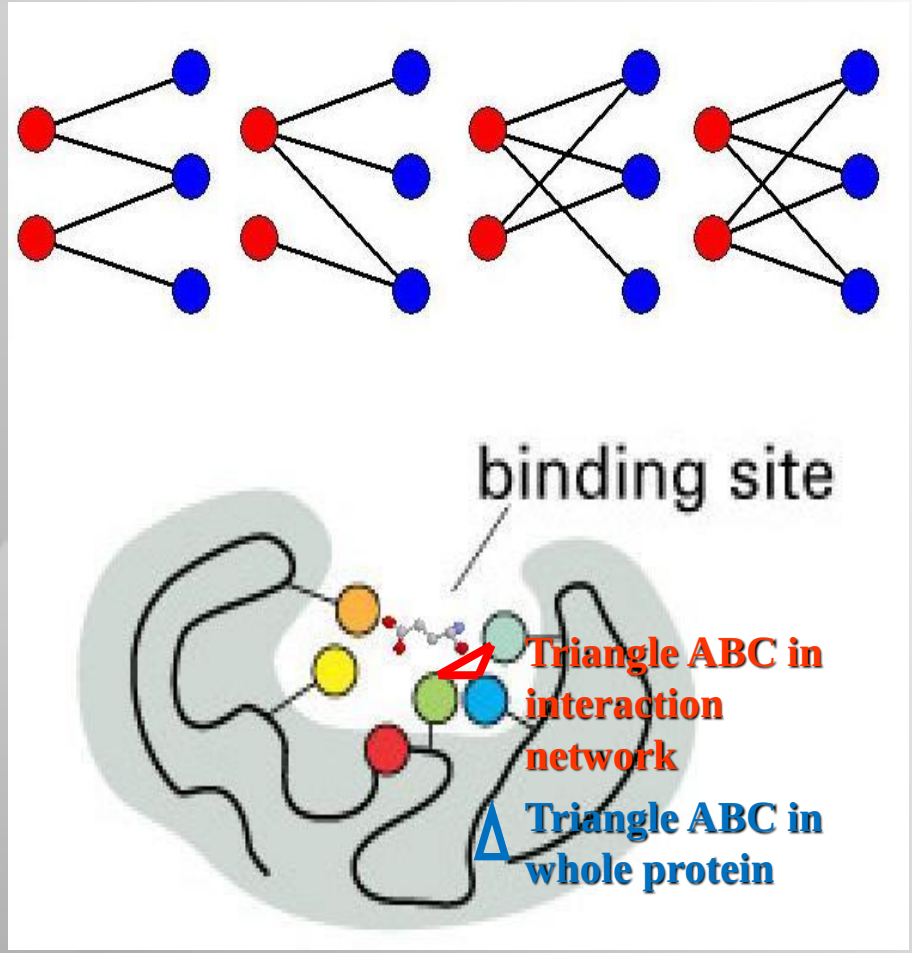
Using Graphs

The University of Georgia

- Enriched in Motifs
- Side length thresholds (10, 10, 13 angstrom)
- Propensities of triangles

$$P_i = \frac{n_i / \sum_{i=1}^{k} n_i}{N_i / \sum_{i=1}^{k} N_i}$$



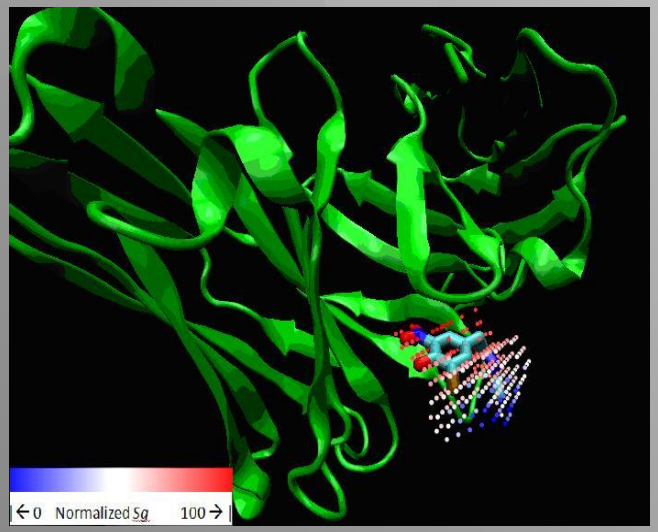


Xie Z.R.& Hwang M.J., *Bioinformatics* 2012

My Algorithm: LISE

The University of Georgia

- Statistical data: Propensities
- Identify protein triangles and assign the scores
- Cluster high scored grids

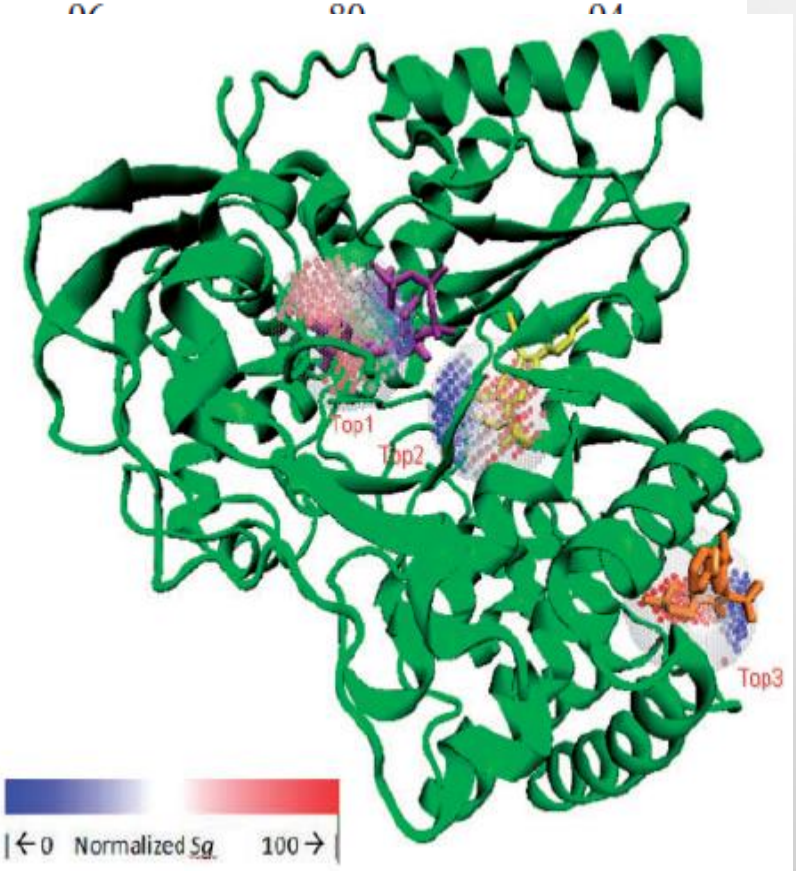


Rank	Triangle	F_b
1	NLC-NLC-NLC	70.6
2	CRP-CRP-CRP	61.9
3	CRP-NLC-NLC	47.2
4	CRP-CRP-NLC	44.3
5	CRP-CRP-OLC	38.1
6	NLC-NLC-NRE	32.8
7	CRP-CRP-NRE	31.7
8	CRP-CRP-O3B	29.8
9	NLC-NLC-O3B	29.6
10	CRP-CRP-NLB	26.2
11	CRN-CRP-CRP	26.2
12	CRP-OLC-OLC	25.8
13	NLB-NLC-NLC	24.6
14	CRP-NRE-OLC	23.5
15	CRP-NLC-O3B	23.5
16	CRP-NLC-NRE	23.0
17	NRE-OLC-OLC	19.9
18	CRP-NRE-NRE	19.7
19	CRP-NLB-NLC	19.2
20	NLC-O3B-O3B	18.6

Xie Z.R.& Hwang M.J., *Bioinformatics* 2012

Results & Comparison (the Best Results So Far)

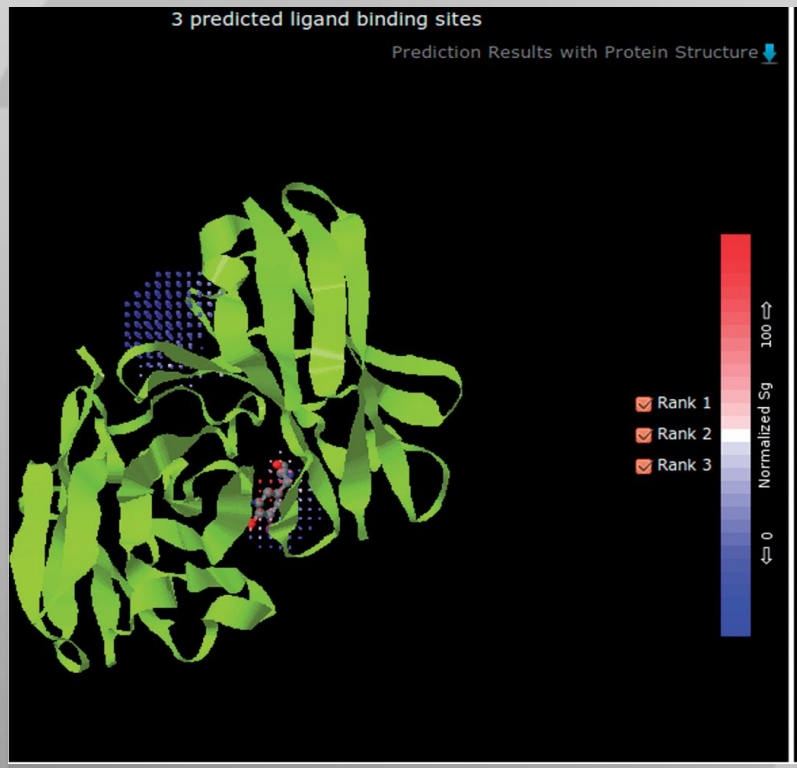
Methods	Bound		Unbound	
	Top1 (%)	Top3 (%)	Top1 (%)	Top3 (%)
LISE (this work)[b]	92	96	81	92
MPK2 (Zhang *et al.*, 2011)	85			
VICE (Tripathi and Kellogg, 2010)[c]	85			
MPK1 (Huang, 2009)	83			
DoGSite (Volkamer *et al.*, 2010)	83			
Fpocket (Le Guilloux *et al.*, 2009)	83			
LIGSITE[cs] (Huang and Schroeder, 2006)	81			
LIGSITE[csc] (Huang and Schroeder, 2006)	79			
MSPocket (Zhu and Pisabarro, 2010)	77			
POCASA (Yu *et al.*, 2010)	77			
Q-SiteFinder (Laurie and Jackson, 2005)[c]	75			
PocketPicker (Weisel *et al.*, 2007)	72			
CAST (Liang *et al.*, 1998)[c]	67			
PASS (Brady and Stouten, 2000)[c]	63			
SURFNET (Laskowski, 1995)[c]	54			

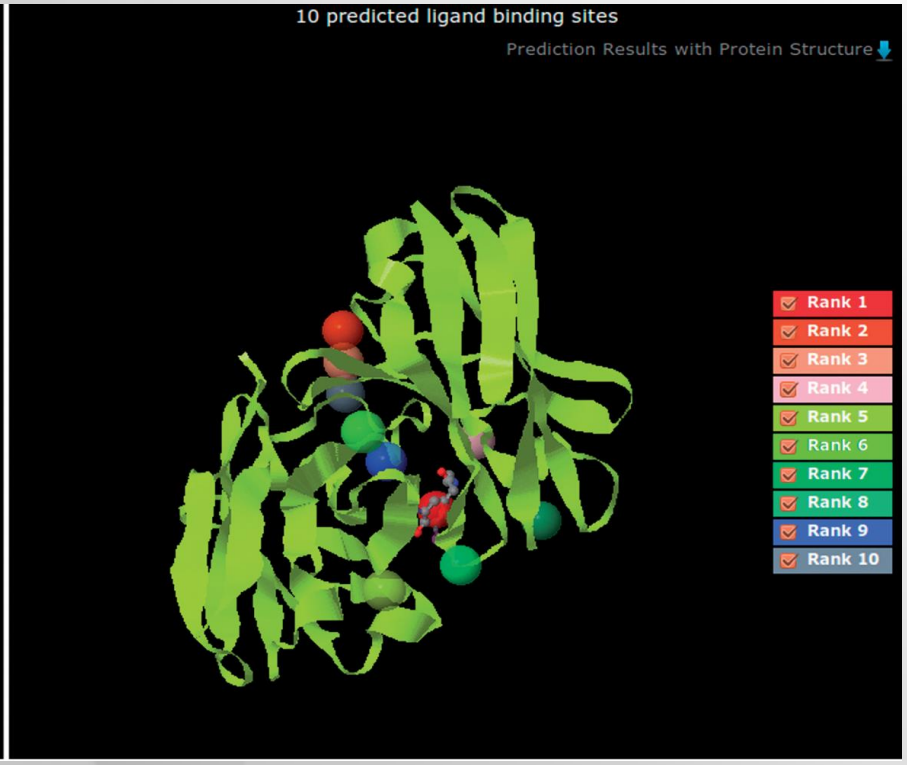


Xie Z.R.& Hwang M.J., *Bioinformatics* 2012

“LISE” Web Server

The University of Georgia





- A web server established and published on *NAR* 2013
- http://lise.ibms.sinica.edu.tw

Xie Z.R. et. al., ***NAR*** 2013

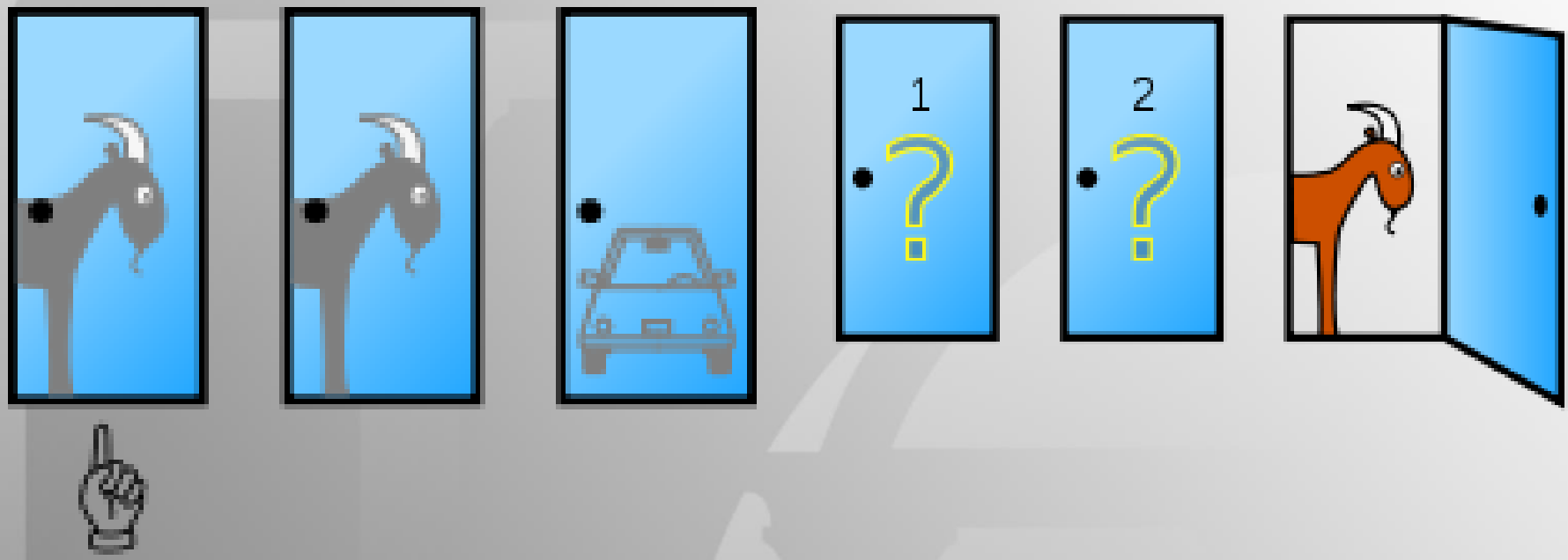
Summary

The University of Georgia

- Created a new docking scoring function
  - The first scoring function using network approach
  - More tolerant to structural flexibility
  - Good results compared with conventional approaches
  - Complementary with energy-based scoring functions
  - Has been applied to a study of drug resistance of HIV

Monte Hall Problem

The University of Georgia



- What is the best strategy?
- Python simulation model

Xie Z.R. et. al., ***NAR*** 2013

Python Time

The University of Georgia

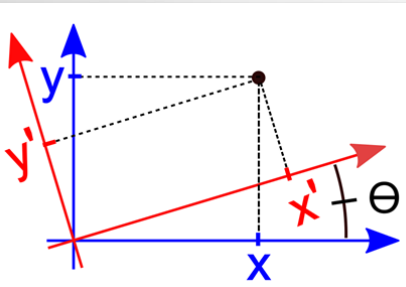
Rotation

The University of Georgia

2D rotation

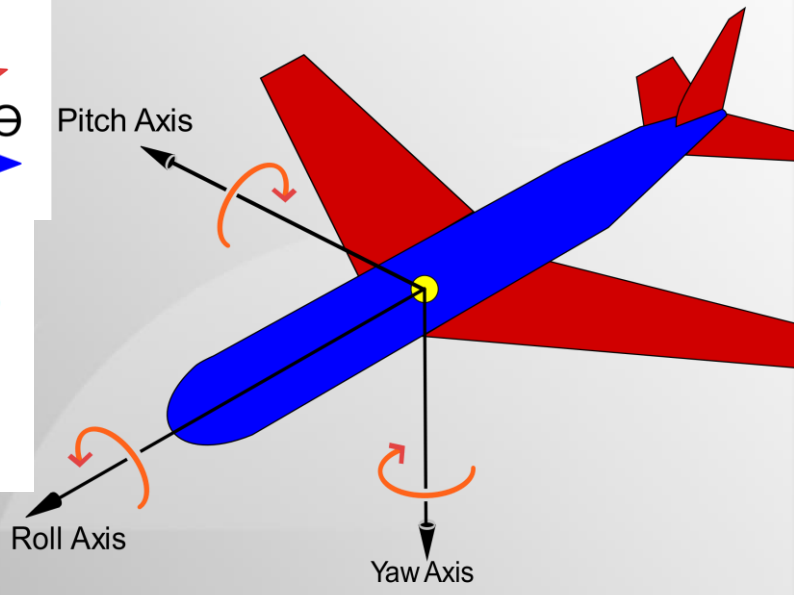
$$x' = x\cos\theta + y\sin\theta$$

$$y' = -x\sin\theta + y\cos\theta.$$[7]



Equations (5) and (6) can be represented in matrix form as

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$



3D rotation

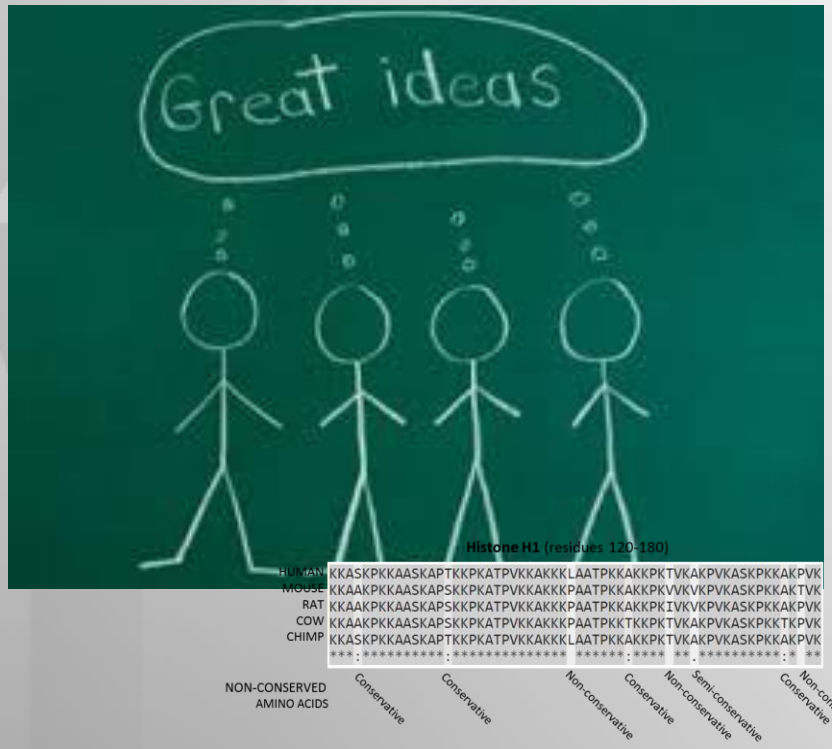
$$R = \begin{bmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \sin\alpha\cos\beta & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma \\ -\sin\beta & \cos\beta\sin\gamma & \cos\beta\cos\gamma \end{bmatrix}$$

Xie Z.R. et. al., ***NAR*** 2013


Thank
You

Group Discussion

The University of Georgia




TIME
FOR A
BREAK