



Large scale data mining for binding patterns and drug discovery

CMSC5719

Prof. Leung, Kwong Sak

Professor of Computer Science and Engineering

Nov, 2012

Outline

❧ I. Introduction

❧ II. Protein-DNA Interactions

❧ III. Drug Discovery

❧ IV. Discussion and Conclusion

Introduction



Bio-/Medical Sciences

- Huge & noisy data
- Costly annotations
- Specific cases

- High Impacts

Bridging:

-Bioinformatics:

More and more crucial
in life sciences and
biomedical
applications for
analysis and new
discoveries



Informatics (e.g. CS)

- Well-organized schemes
- Automatic analysis
- Generalized knowledge

- Desire for Applications

I. Introduction to Bioinformatics

- ⌘ Research Areas
- ⌘ Biological Basics

Bioinformatics Research Areas

☞ Many (crossing) areas:

☞ (Genome-scale) Sequence Analysis

- ☞ Sequence alignments, motif discovery, genome-wide association (to study diseases such as cancers)

☞ Computational Evolutionary Biology

- ☞ Phylogenetics, evolution modeling

☞ Analysis of Gene Regulation

- ☞ Gene expression analysis, alternative splicing, protein-DNA interactions, gene regulatory networks

☞ Structural Biology

- ☞ Drug discovery, protein folding, protein-protein interactions

☞ Synthetic Biology

☞ Applications on High throughput Sequencing Data (NGS)

☞ ...

Our Bioinformatics Group



☞ Dept. of Computer Science & Engineering, CUHK

- ☞ Prof. Kwong-Sak LEUNG
- ☞ Prof. Kin-Hong LEE
- ☞ Prof. Man-Hon WONG
- ☞ Prof. Kevin YIP
- ☞ Dr. Cyrus Tak-Ming CHAN

☞ Research Partners from CUHK

- ☞ Prof. Stephen Kwok-Wing TSUI, Director of Hong Kong Bioinformatics Center, School of Biomedical Sciences
- ☞ Prof. Hsiang-Fu KUNG, Stanley Ho Centre for Emerging Infectious Diseases
- ☞ Prof. Marie Chia-Mi LIN, Department of Surgery, Prince of Wales Hospital
- ☞ Prof. Pang-Chui SHAW, School of Biomedical Sciences

☞ 10 Research Students/Staff (KS Group)

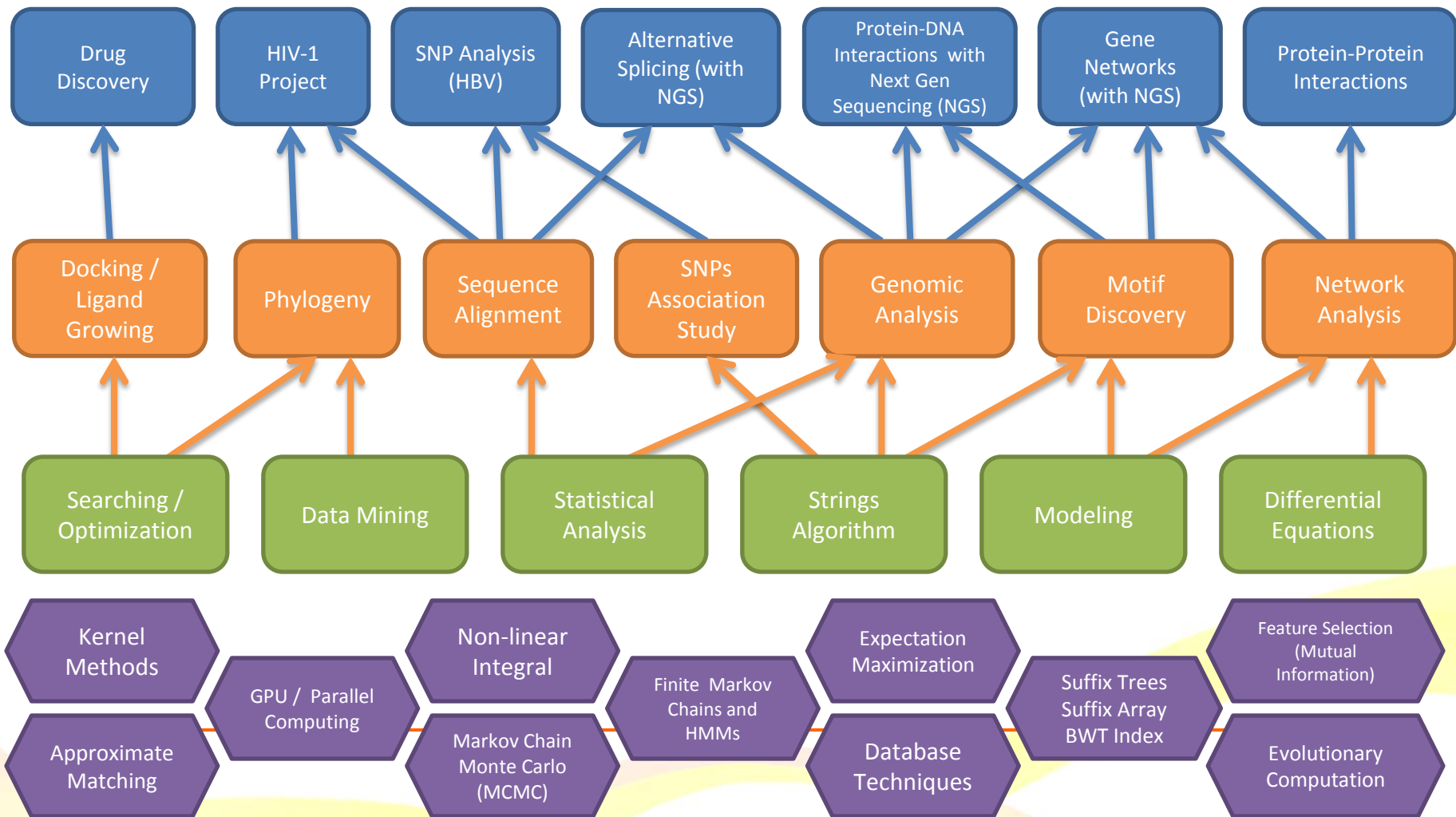
- ☞ 1 Postdoc
- ☞ 5 PhD Students
- ☞ 4 MPhil Students

Our Research Roadmap

Real-life Projects

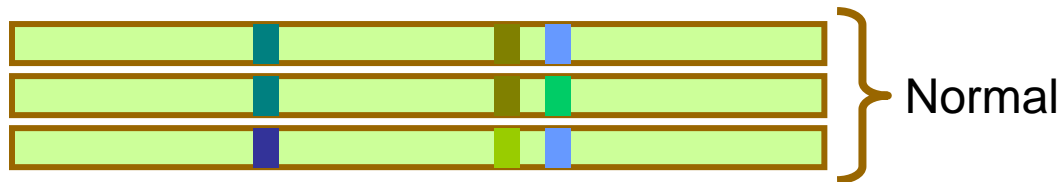
→ Related Bioinformatics Problems

→ Computer Techniques

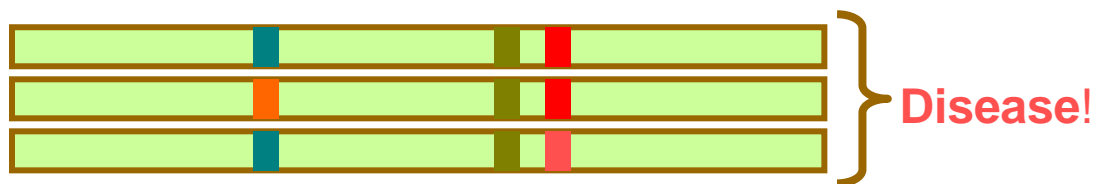


Genome-wide Association

Human DNA sequences



...



SNPs (single nucleotide polymorphism; >5% variations)

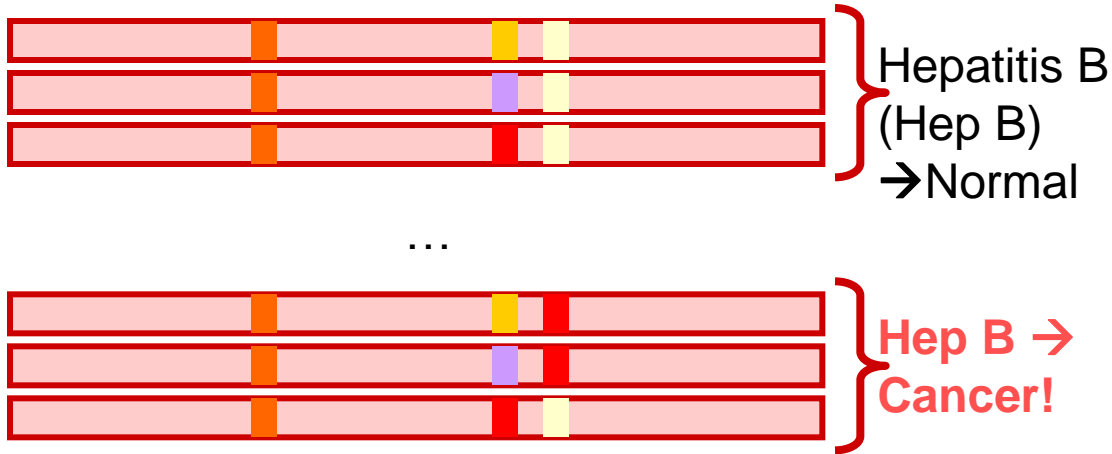


Targets: SNPs that are associated with genetic diseases; Diagnosis and healthcare for high-risk patient

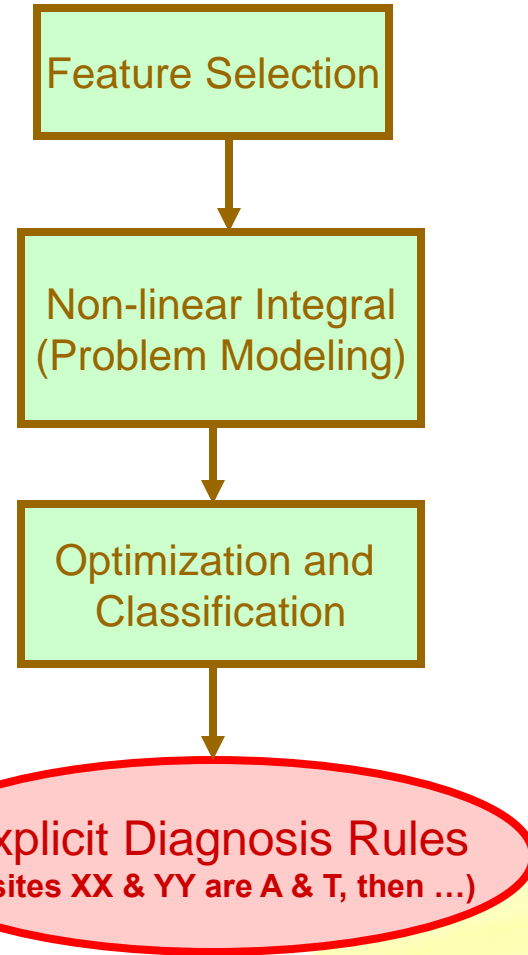
Methods: Feature selection; mutual information; non-linear integrals; Support Vector Machine (SVM);

HBV Project (Example)

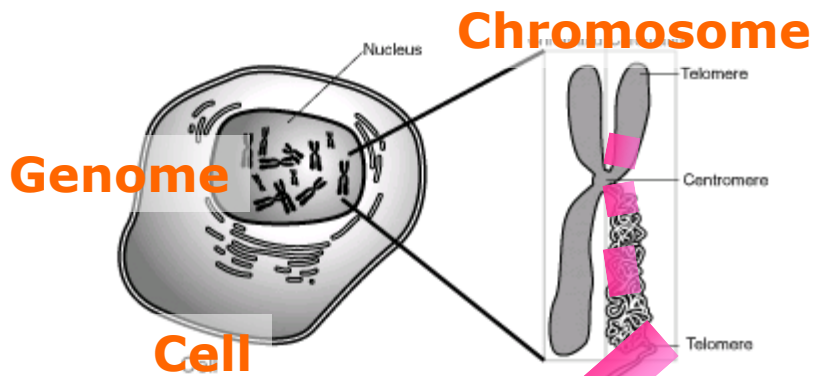
HBV sequences



SNPs are not known and to be discovered by alignments



Biological Basics



A string of amino acids
 $\Sigma = \{A, R, N, D, C, E, \dots\}$
 $|\Sigma| = 20$

Other functions:
 Protein-protein
 Protein-ligand



Regulatory functions

DNA Sequence

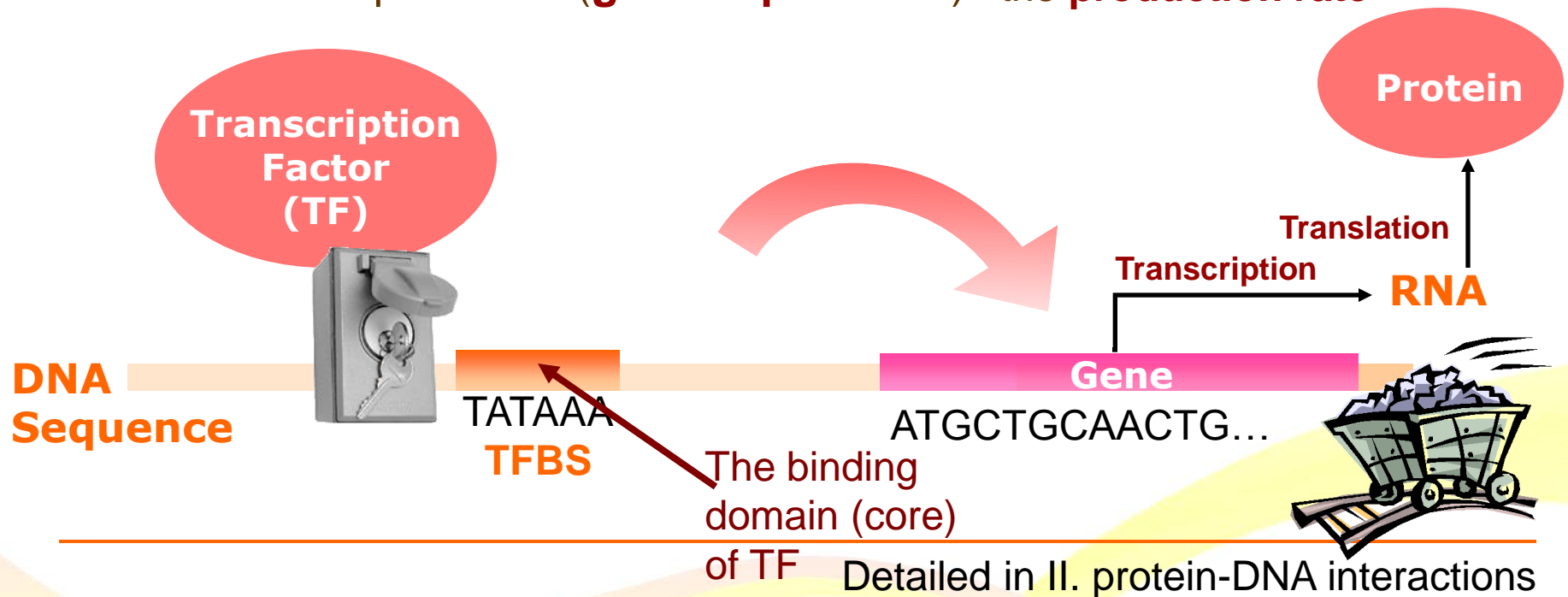
5' ... AGACTGCGGA ... 3'
 3' ... TCTGACGCCT ... 5'

... AGACTGCGGA ...
 A string with alphabet
 $\Sigma = \{A, C, G, T\}$

Transcriptional Regulation

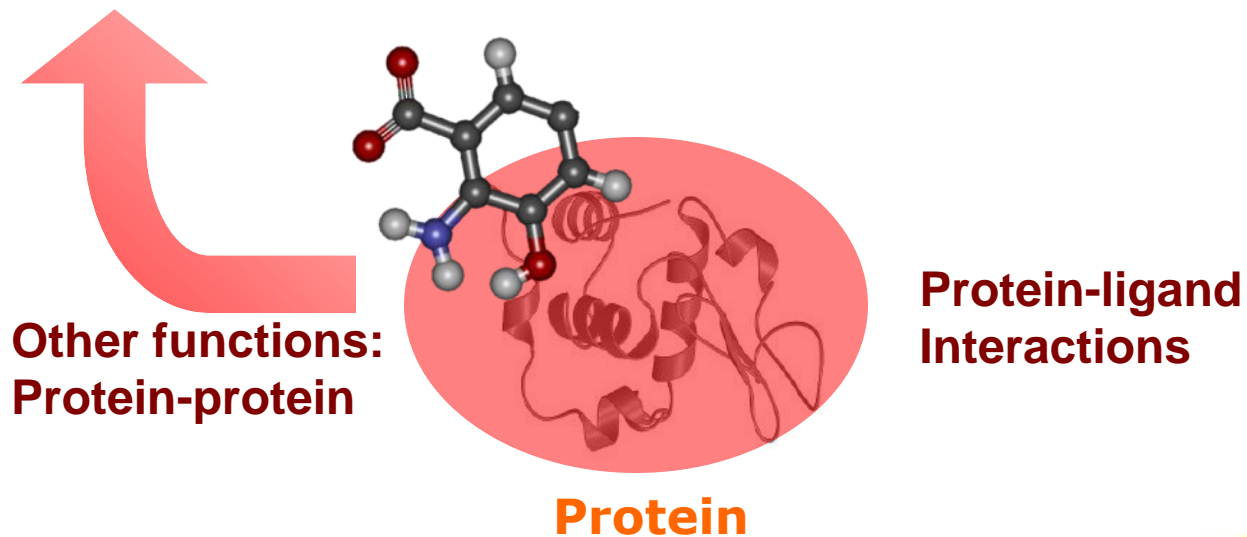
Binding for Transcriptional Regulation

- Transcription Factor (**TF**): the protein as the **key**
- TF Binding Site (**TFBS**): the DNA segment as the **key switch**
- Transcription rate (**gene expression**): the **production rate**



Protein-ligand Interactions

∞ Drug Discovery

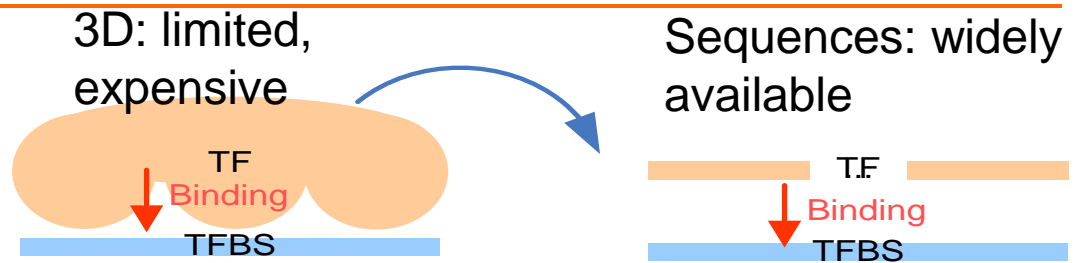


Detailed in III. drug discovery

II. Protein-DNA Interactions

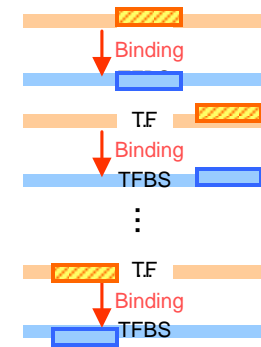
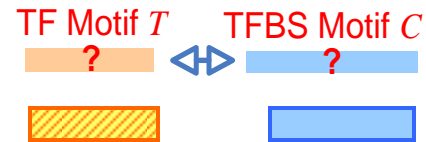
- ❧ Introduction
- ❧ Approximate TF-TFBS rule discovery
- ❧ Results and Analysis
- ❧ Discussion

Introduction



- ☞ We focus on TF-TFBS bindings which are primary protein-DNA interactions
- ☞ Discover TF-TFBS binding relationship to understand gene regulation
 - ☞ Experimental data: 3D structures of TF-TFBS bindings are **limited and expensive (Protein Data Bank PDB)**; TF-TFBS binding sequences are **widely available (Transfac DB)**
 - ☞ Further bioengineering or biomedical applications to manipulate or predict TFBS and/or TF (esp. cancer targets) given either side
- ☞ Existing Methods
 - ☞ **Motif discovery**: either on protein (TF) or DNA (TFBS) side. No linkage for direct TF-TFBS relationship
 - ☞ **One-one binding codes**: R-A, E-C, K-G, Y-T? No universal codes!
 - ☞ **Machine learning**: training limitation (limited 3D data) and not trivial to interpret or apply

Conservation



- ☞ TFBSs, Genes → merely A,C,G,Ts;
- ☞ The binding domains of TFs → merely amino acids (AAs)
 - ☞ What distinguish them from the others? **Conservation**
 - ☞ Functional sequences are less likely to change through evolution
 - similar **Patterns** across genes/species → Bioinformatics!

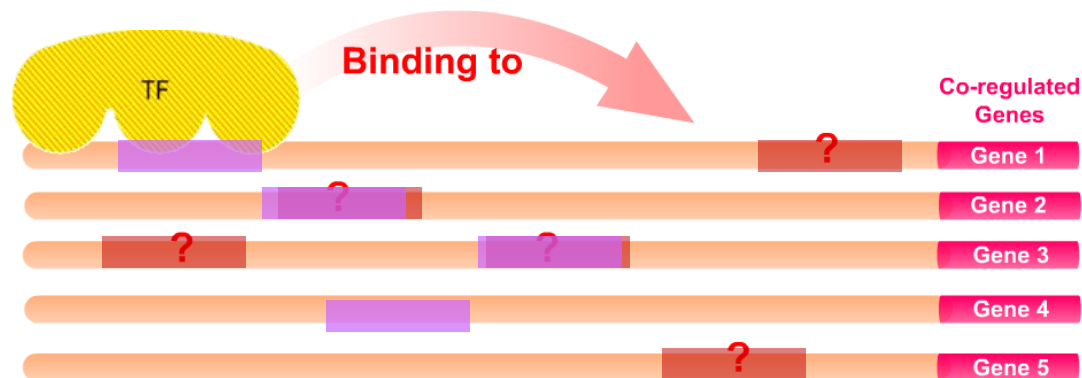
☞ Association rule mining

- ☞ Exploit the overrepresented and conserved sequence patterns (motifs) from large-scale protein-DNA interactions (TF-TFBS bindings) sequence data
- ☞ Promising initial results obtained with verifiable rules!
- ☞ Biological mutations and experimental noises exist!—**Approximate** rules

Motivations: overall

Finding motifs one-sided is challenging and difficult

e.g. TFBS Motif Discovery: Noises, variations through mutations, unknown locations—weak signals to be recovered



—Prediction —True TFBS

Motivations: overall

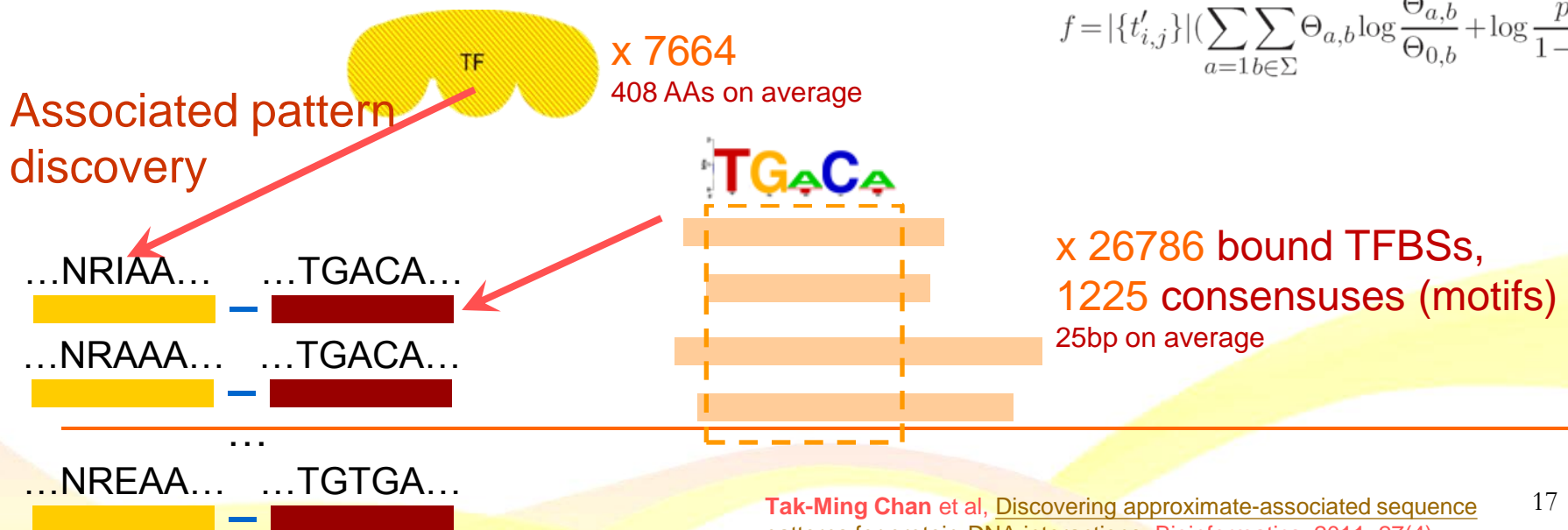
∞ Finding associated patterns on both sides is shown to be promising—when you have many diverse binding sequences (e.g. TRANSFAC)

∞ Associated TF-TFBS patterns found from sequences

∞ Grouping TRANSFAC data

∞ Developing a customized TF core motif discovery algorithm

$$f = |\{t'_{i,j}\}| \left(\sum_{a=1}^w \sum_{b \in \Sigma} \Theta_{a,b} \log \frac{\Theta_{a,b}}{\Theta_{0,b}} + \log \frac{p}{1-p} - 1 \right)$$

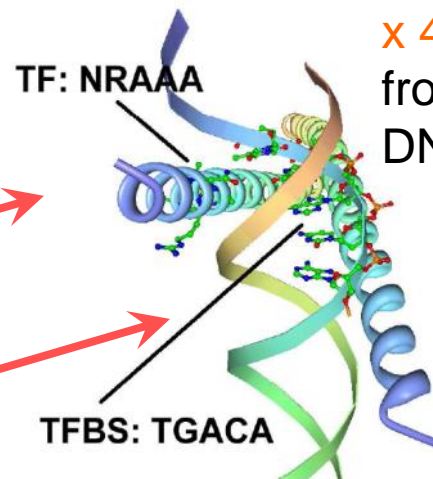


Motivations: overall

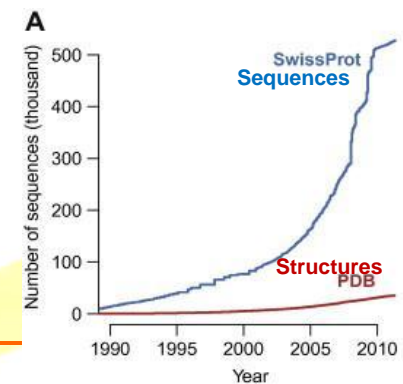
Finding associated patterns on both sides is shown to be promising—when you have many diverse binding sequences (e.g. TRANSFAC)

Associated TF-TFBS patterns found from sequences are verified on 3D structures to be binding cores!

Verified on 3D structures
(binding cores $<3.5\text{\AA}$)



x 40222 binding pairs
from 1290 PDB protein-DNA complexes



Problem Definition

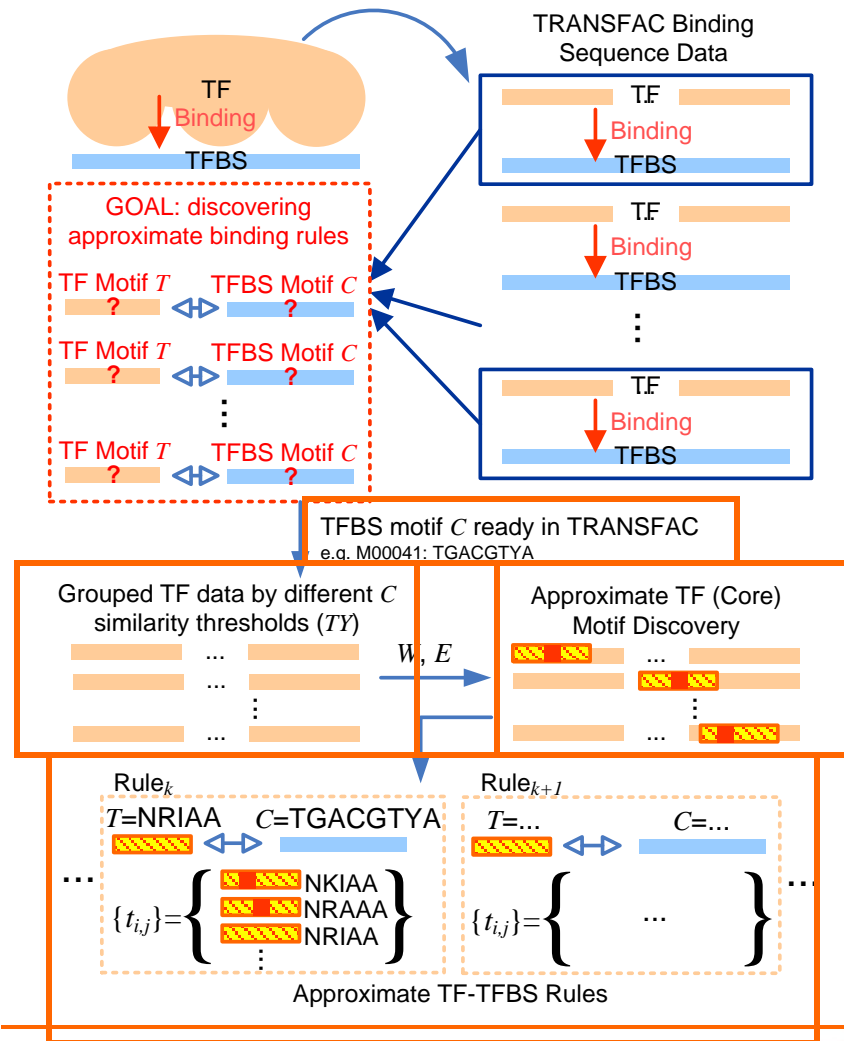
GOAL: discovering
approximate binding rules



- Input: given a set of TF-TFBS binding sequences** (TF: hundreds of AAs; TFBS: tens of bps depending on experiment resolution), discover the associated patterns of width w (potential interaction cores within binding distance)
- Output: Approximate associated TF-TFBS binding sequence patterns (TF-TFBS rules)**

 - given binding sequence data (Transfac) **ONLY**, predict short TF-TFBS pairs verifiable in real 3D structures of protein-DNA interactions (PDB)!

Overall Methodology



A progressive approach:

Use the available TFBS motifs C from Transfac DB—already **approximate** with ambiguity code representation—TFBS side done!

Group TF sequences corresponding to different TFBS consensus (motif) groups C with similarity thresholds $TY=0.0, 0.1, 0.3$

Approximate TF Core Motif Discovery for T (instance set $\{t_{i,j}\}$) give W and E —TF side done

Associating T ($\{t_{i,j}\}$) with C

TF Side: Core TF Motif Discovery

☞ The customized algorithm

- ☞ Input: width W and (substitution) error E , TF Sequences S
- ☞ Find W -patterns (at least 1 hydrophilic amino acid) and their E approximate matches
- ☞ Iteratively find the optimal match set $\{t_{i,j}\}$ based on the Bayesian scoring function f for motif discovery:

$$f = |\{t'_{i,j}\}| \left(\sum_{a=1}^w \sum_{b \in \Sigma} \Theta_{a,b} \log \frac{\Theta_{a,b}}{\Theta_{0,b}} + \log \frac{p}{1-p} - 1 \right)$$

$p = |\{t'_{i,j}\}| / |S|$ is the abundance ratio
Overrepresented

position weight matrix (PWM) Θ
Conserved

- ☞ Top $K=10$ motifs are output, each with its instance set $\{t_{i,j}\}$

Results and Analysis

Verification

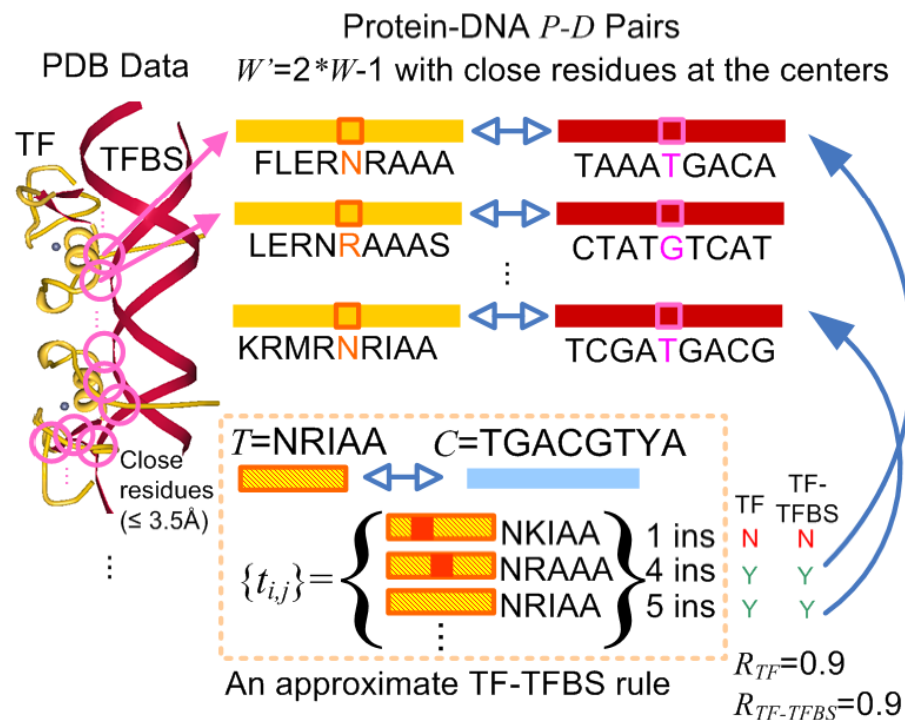
on Protein Data Bank

(PDB)

Most representative database of **experimentally determined** protein-DNA 3D structure data

* expensive and time consuming

* most accurate evidence for verification



Check the approximate TF-TFBS rules $T(\{t_{i,j}\})-C$

Approximate appearance in binding pairs from PDB 3D structure data : width W bounded by error E

TF side (R_{TF}): instance oriented— $\{t_{i,j}\}$ evaluated

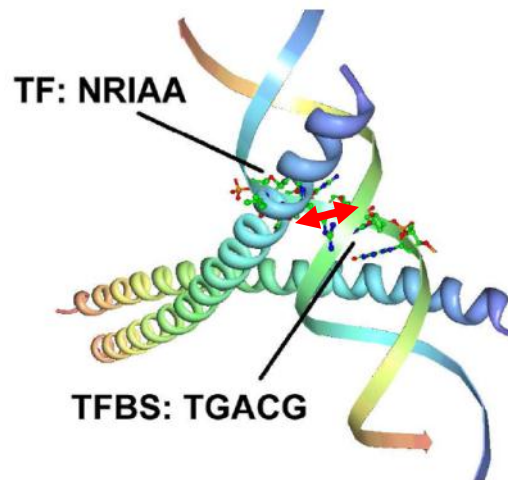
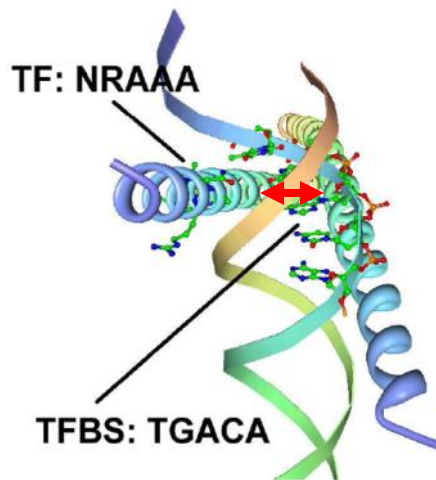
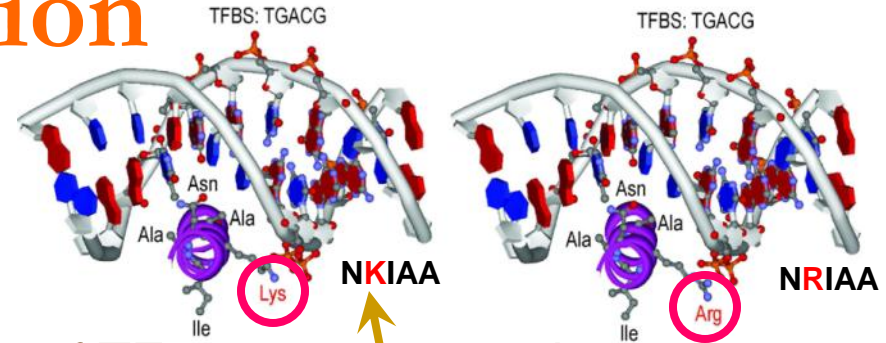
TFBS side ($R_{TF-TFBS}$): pattern oriented— C evaluated

R : verification ratio
[0,1] higher the better

Biological verification

Recall the challenge

- Given sequence datasets of tens of TF sequences, each **hundreds of AA in length**, grouped by TFBS consensus C (5~20bp),
- Predict $W(=5,6)$ substrings ($\{t_{i,j}\}$) associated with C



Which can be verified in actual 3D TF-TFBS binding structures as well as homology modeling (by bio experts)!

PDB Verified examples in Rule NRIAA(NKIAA; NRAAA; NREAA; NRIAA)-TGACGTYA

Results and Analysis

One more verified example

1NKP:

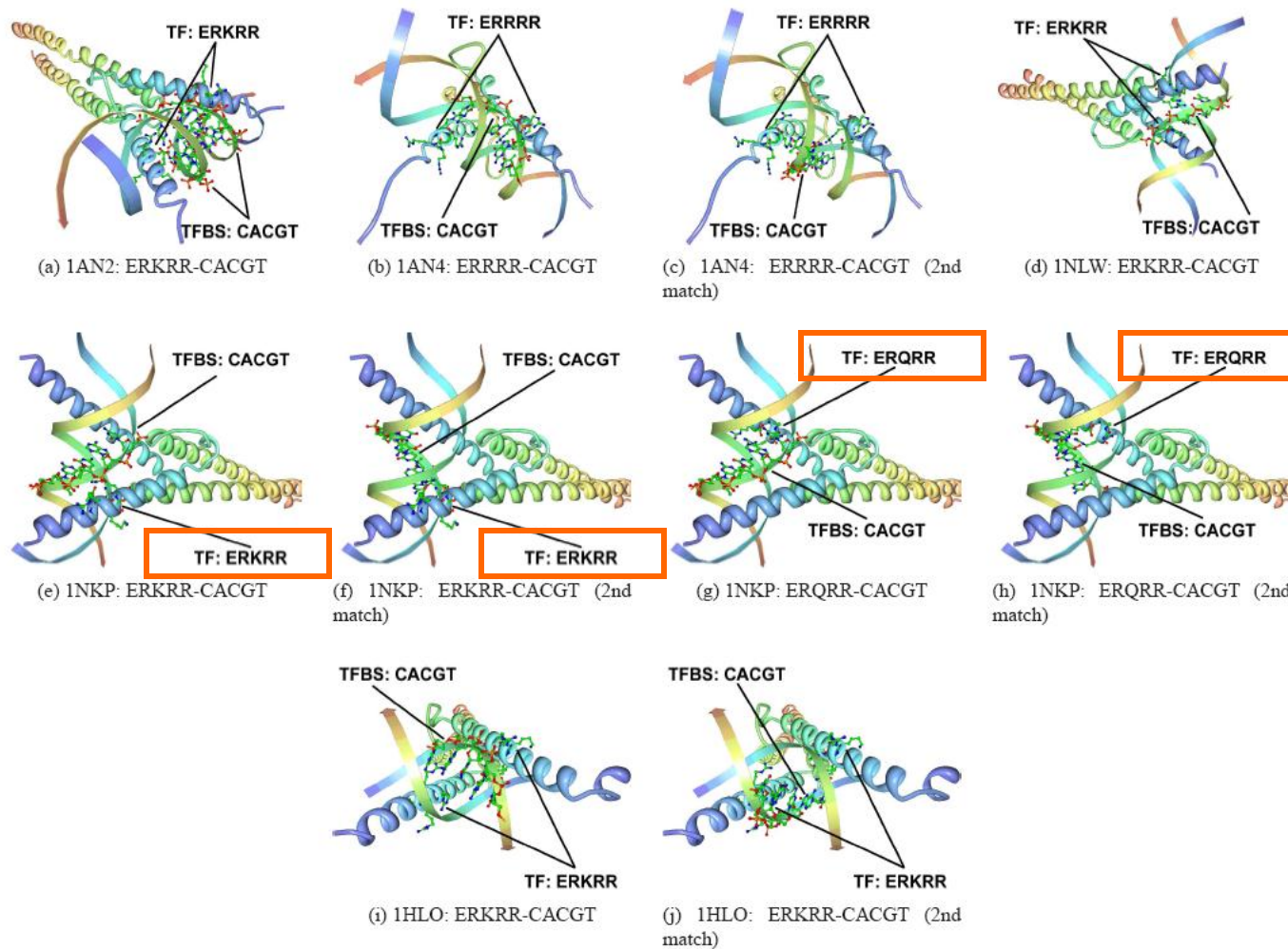


Figure 4. PDB verifications for rule M00217: ERKRR(ERKRR; ERQRR; ERRRR)-CACGTG for $W = 5$, $E = 1$, $TY = 0.1$ using ProteinWorkshop.

M00217: ERKRR(ERKRR; ERQRR; ERRRR)-CACGTG

Results and Analysis

Quantitative Comparisons with Exact Rules

TY	$W = 5, E = 0$				$W = 5, E = 0$				$W = 5, E = 1$					
	Exact rules [52]		0.0		0.1		0.3		0.0		0.1		0.3	
	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS
AVG R_*	0.57	0.44	0.74	0.64	0.78	0.70	0.82	0.73	0.57	0.56	0.63	0.62	0.69	0.68
$R_* > 0$	99	76	127	110	165	147	636	567	235	231	291	287	2101	2072
Rule No.	173	173	172	172	211	211	774	774	346	346	396	396	2559	2559
$R_* > 0$ Ratio	0.57	0.44	0.74	0.64	0.78	0.70	0.82	0.73	0.68	0.67	0.73	0.72	0.82	0.81
TY	$W = 6, E = 0$				$W = 6, E = 0$				$W = 6, E = 1$					
	Exact rules [52]		0.0		0.1		0.3		0.0		0.1		0.3	
	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS
AVG R_*	0.18	0.18	0.71	0.58	0.76	0.65	0.81	0.67	0.58	0.54	0.63	0.60	0.70	0.68
$R_* > 0$	6	6	108	88	143	121	448	370	181	169	234	222	1665	1618
Rule No.	34	34	153	153	187	187	555	555	271	271	319	319	1920	1920
$R_* > 0$ Ratio	0.18	0.18	0.71	0.58	0.76	0.65	0.81	0.67	0.67	0.62	0.73	0.70	0.87	0.84

More informative (verified) rules (**76 VS 110** $W=5$; **6 VS 88** $W=6$)

Improvement on exact ones (**AVG R_* 29%, 46%** better with $W=5$)

Results and Analysis

73%-262% improvement on AVG R_*
33%-84% improvement on $R_* > 0$ Ratio
 Customized TF core motif discovery is necessary

Comparisons with MEME as TF side discovery tool

MEME Results	$W = 5, E = 0$						$W = 5, E = 1$					
TY	0.0		0.1		0.3		0.0		0.1		0.3	
R_*	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS
AVG R_*	0.33	0.26	0.36	0.28	0.37	0.28	0.33	0.32	0.36	0.34	0.37	0.36
Ours better by	124%	144%	120%	146%	120%	160%	73%	74%	76%	79%	85%	91%
$R_* > 0$	143	123	179	151	1306	1071	143	142	179	175	1306	1262
Rule No.	298	298	342	342	2118	2118	298	298	342	342	2118	2118
$R_* > 0$ Ratio	0.48	0.41	0.52	0.44	0.62	0.51	0.48	0.48	0.52	0.51	0.62	0.60
Ours better by	54%	55%	49%	58%	33%	45%	42%	40%	40%	42%	33%	36%

MEME Results	$W = 6, E = 0$						$W = 6, E = 1$					
TY	0.0		0.1		0.3		0.0		0.1		0.3	
R_*	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS
AVG R_*	0.29	0.22	0.31	0.23	0.29	0.18	0.29	0.27	0.31	0.29	0.29	0.26
Ours better by	142%	163%	145%	181%	178%	262%	97%	96%	102%	104%	142%	157%
$R_* > 0$	127	96	163	121	1194	839	127	120	163	154	1194	1127
Rule No.	289	289	334	334	2170	2170	289	289	334	334	2170	2170
$R_* > 0$ Ratio	0.44	0.33	0.49	0.36	0.55	0.39	0.44	0.42	0.49	0.46	0.55	0.52
Ours better by	61%	73%	57%	79%	47%	72%	52%	50%	50%	51%	58%	62%

Discussion

- ☞ For the first time we generalize the exact TF-TFBS associated sequence patterns to approximate ones
- ☞ The discovered approximate TF-TFBS rules
 - ☞ Competitive performance with respect to verification ratios (R_*) on both TF and TF-TFBS aspects
 - ☞ Strong edge over exact rules and MEME results
 - ☞ Demonstration of the flexibility of specific positions TF-TFBS binding (further biological verification with NCBI independent protein records!)

Further Results

☞ We can go further with these promising associated TF-TFBS patterns

☞ Discovering and analyzing the binding variances (subtypes): e.g. 3rd E variation is associated with T, G variations on TFBS

Subtypes may

- Lead to changed binding preferences
- Distinguish conserved from flexible binding residues
- Reveal novel binding mechanisms



The result analysis positively supports all these!

Results more “biological”; check out the following if interested:

Tak-Ming Chan et al, Subtypes of Associated Protein-DNA (TF-TFBS) Patterns, Nucleic Acids Research, 2012, 40 (19): 9392-9403.

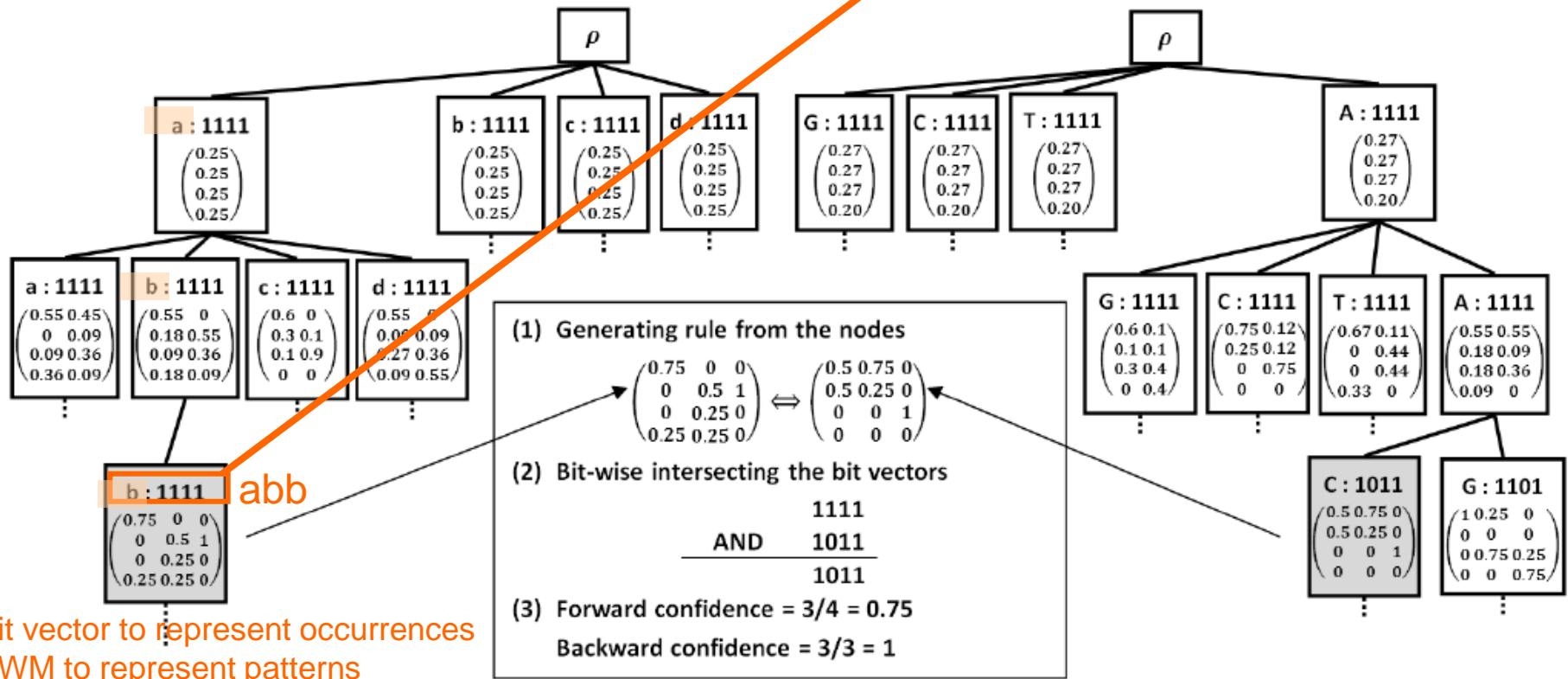
Further Results

ID	TF (simplified)	TFBS
1	abbccdac	TACGTG
2	ddbbedac	ACGTCA
3	dacadb	TACATC
4	bcdacb	GAACGT

A simple biological database

Frequency Sequence Class (FSC) Tree structures for efficiency

(ICDE 2012)



Bit vector to represent occurrences
PWM to represent patterns

Several orders of magnitude faster than Apriori (association rule mining) algorithm

Discussion

- ☞ Great and promising direction for further discovering protein-DNA interactions

- ☞ Future Work
 - ☞ Formal models for whole associated TF-TFBS rules
 - ☞ Advanced Search algorithms for motifs
 - ☞ Associating multiple short TF-TFBS rules
 - ☞ Handling uncertainty such as widths

- ☞ Applications
 - ☞ Generalization of TF-TFBS binding mechanisms
 - ☞ Subtype and phylogeny analysis
 - ☞ Genetic disease and regulation modification analysis

III. Drug Discovery

Background

idock: Protein-ligand docking

istar: Novel web platform

igrow: De novo ligand design

iview: HTML5 visualizer

Case study of influenza

Case study of cancers

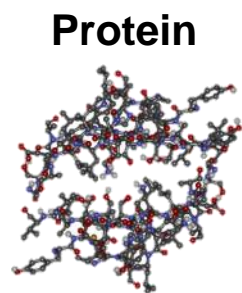
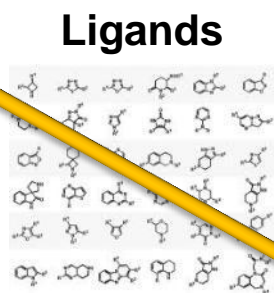
Drug Discovery

☞ Expensive and long-term business

☞ US\$1.8B over 13 years to develop a new drug



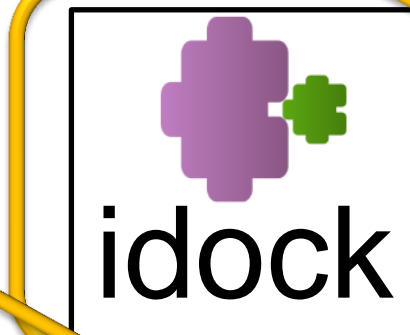
Our Contributions & Proposal



**Influenza A
CCRK-Related Cancers
Cancer Stem Cells**

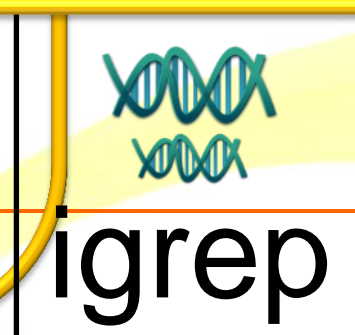
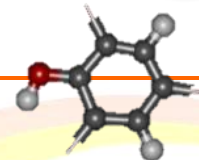
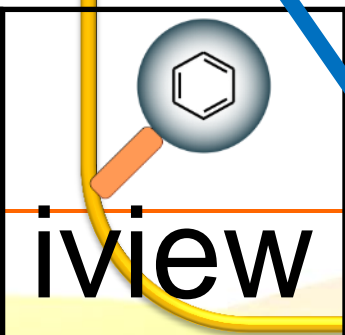
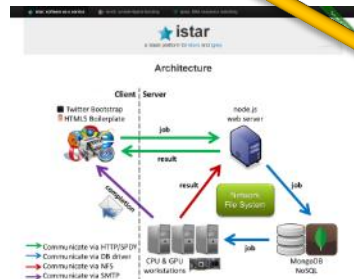


idock 3.0



idock 2.0

**General
GPU ready**



Our Progress

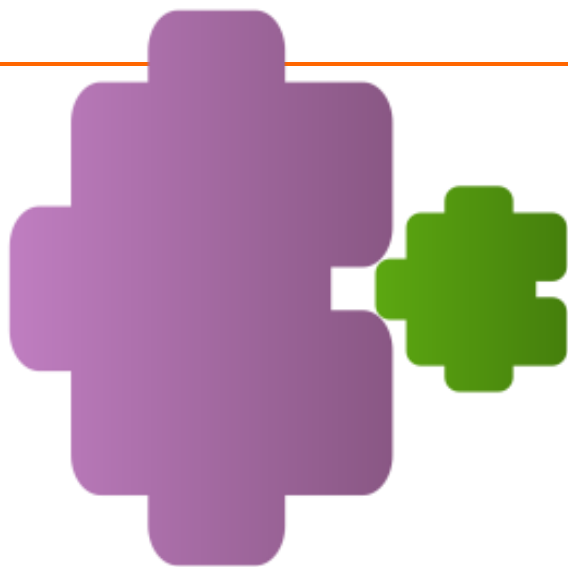
Projects / Case studies	Progress
idock 1.0: Protein-Ligand Docking	100%
idock 1.6: Protein-Ligand Docking	100%
istar: Software-as-a-Service Platform	100%
idock 2.0: GPU Acceleration	5%
idock 3.0: Ligand Synthesis	30%
iview: HTML5 Visualizer	30%
Case Study of Influenza A	90%
Case Study of CCRK-Related Cancers	90%
Case Study of Cancer Stem Cells	0%



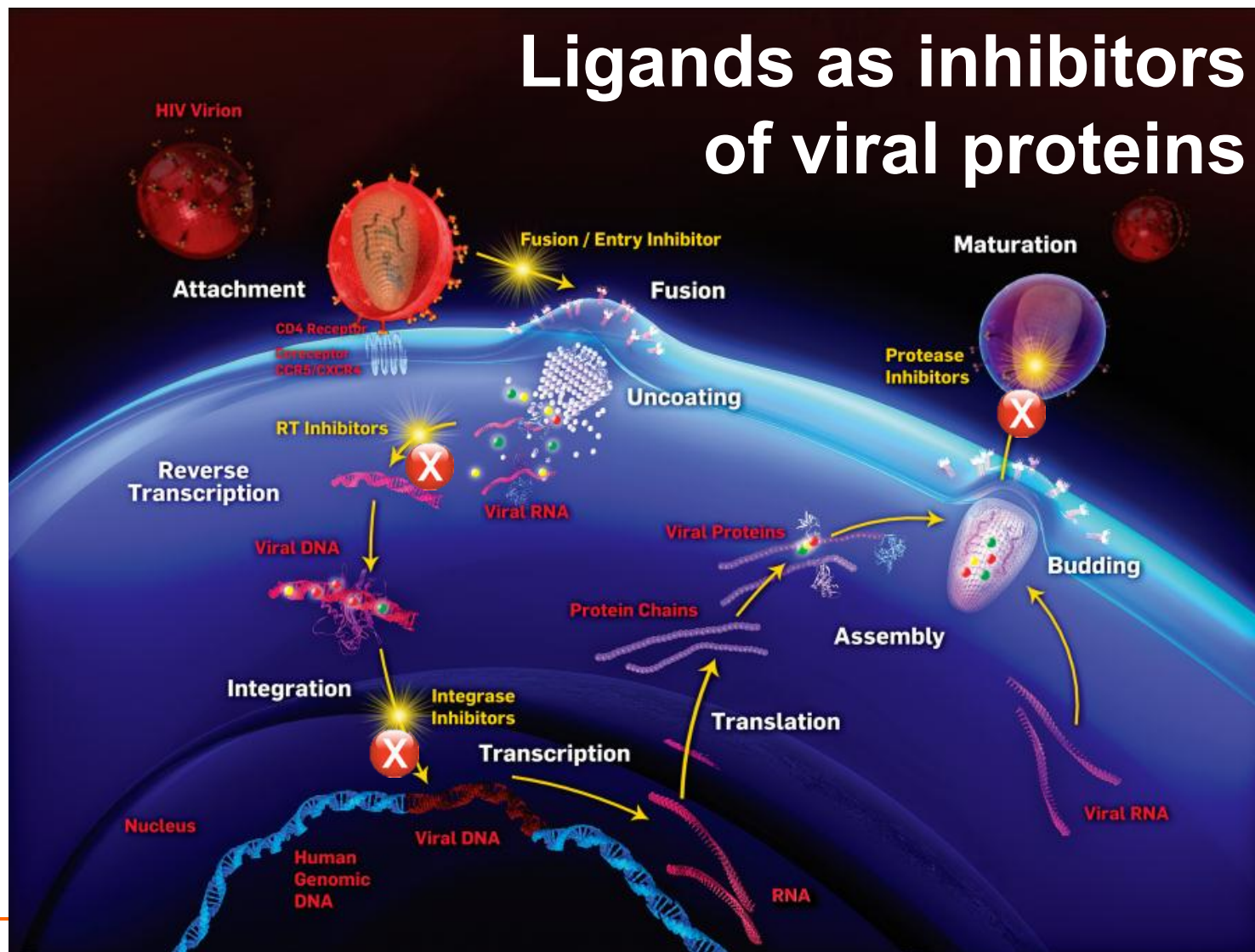


idock

Protein-Ligand Docking

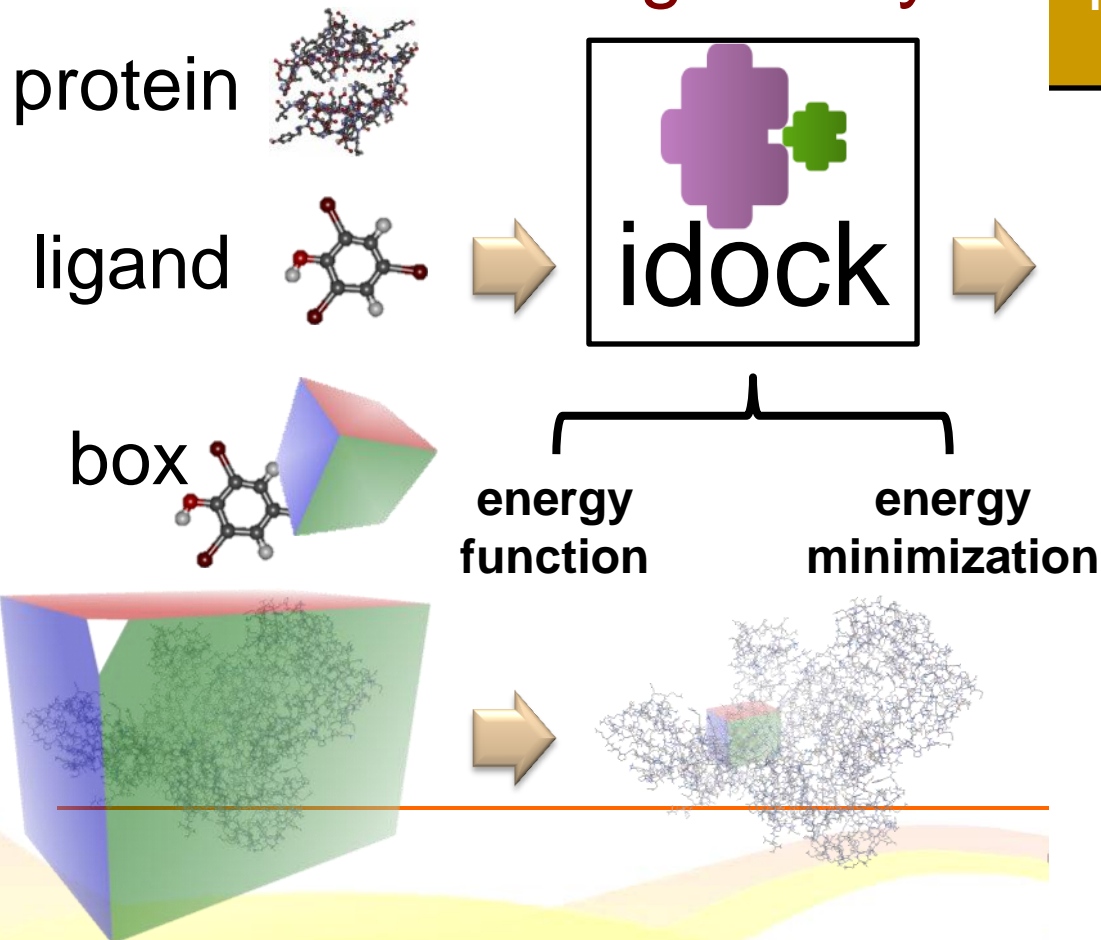



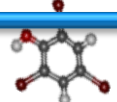


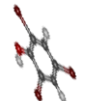
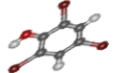
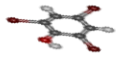
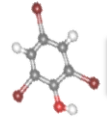
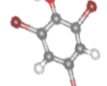
Replication Cycle of HIV/AIDS

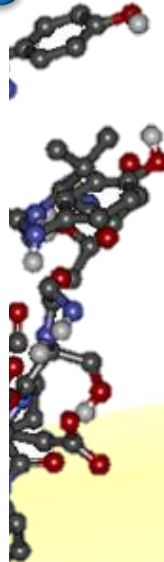


Input and Output

- Translate and rotate the ligand (ie. compound)
- Predict binding affinity



Rank	Conformation	Free energy (kcal/mol)
1		-7.0
2		-6.1
3		-6.0
4		-5.9
5		-5.9
6		-5.8
7		-5.8
8		-5.7
9		-5.6



Energy Function

$$\mathcal{E} = \sum_{i < j} \left(\begin{array}{l} (-0.035579) * \text{Gauss}_1(t_i, t_j, r_{ij}) + \\ (-0.005156) * \text{Gauss}_2(t_i, t_j, r_{ij}) + \\ (+0.840245) * \text{Repulsion}(t_i, t_j, r_{ij}) + \\ (-0.035069) * \text{Hydrophobic}(t_i, t_j, r_{ij}) + \\ (-0.587439) * \text{HBonding}(t_i, t_j, r_{ij}) \end{array} \right)$$

☞ Sum over all pairs of movable heavy atoms i and j

☞ r_{ij} : interatomic distance, cutoff $r_{ij} = 8 \text{ \AA}$

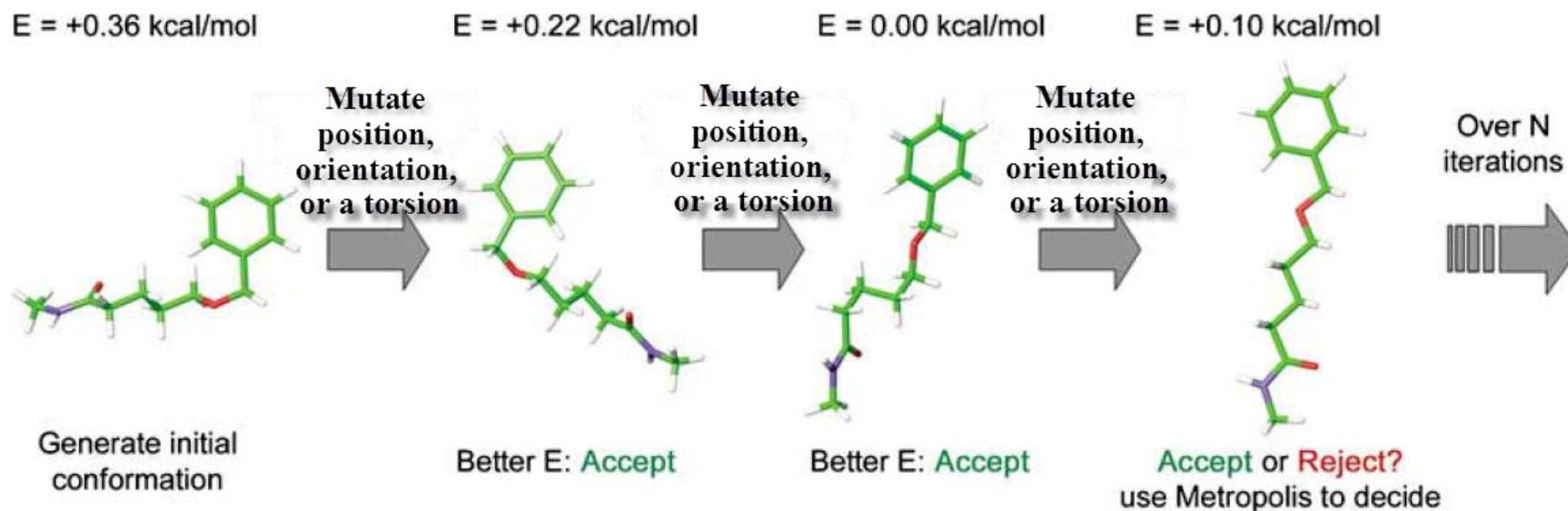
☞ t_i : atom type of i

☞ t_j : atom type of j

☞ Conformation = (position, orientation, torsions)

Energy Optimization Algorithm

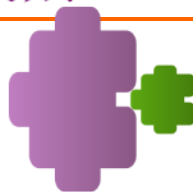
- Global optimization: Multithreaded Monte Carlo
- Local optimization: BFGS Quasi-Newton method



Multithreading via parallel tasks

$$\text{if } \frac{\exp(-E_n/RT)}{\exp(-E_{n-1}/RT)} > z, \text{ accept}$$

Our Tool idock



Based on AutoDock Vina

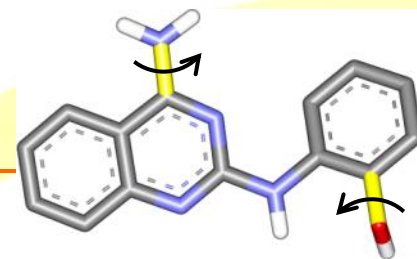
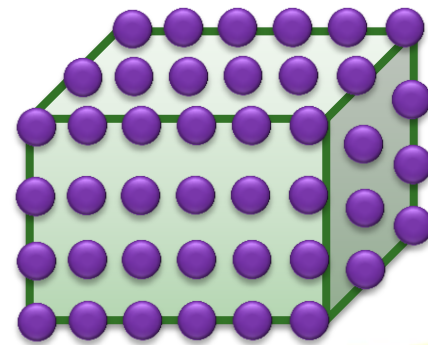
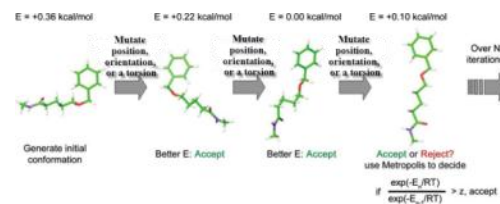
Same energy function

Same optimization algorithm

Our contributions

- NEW Support for virtual screening
- NEW Faster evaluation of scoring function
- NEW Thread pool for high CPU utilization
- NEW Auto deactivation of inactive torsions
- NEW Support for 25 chemical elements
- NEW Support for gzip/bzip2 ligands
- NEW Verbose output to PDBQT and CSV

$$e = \sum_{i < j} \begin{pmatrix} (-0.035579) * Gauss_1(t_i, t_j, r_{ij}) + \\ (-0.005156) * Gauss_2(t_i, t_j, r_{ij}) + \\ (+0.840245) * Repulsion(t_i, t_j, r_{ij}) + \\ (-0.035069) * Hydrophobic(t_i, t_j, r_{ij}) + \\ (-0.587439) * HBonding(t_i, t_j, r_{ij}) \end{pmatrix}$$



idock speedup over Vina

12 proteins

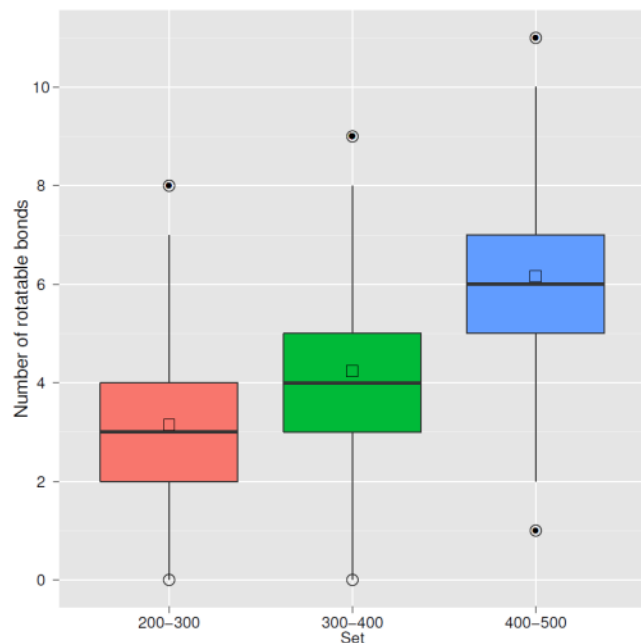
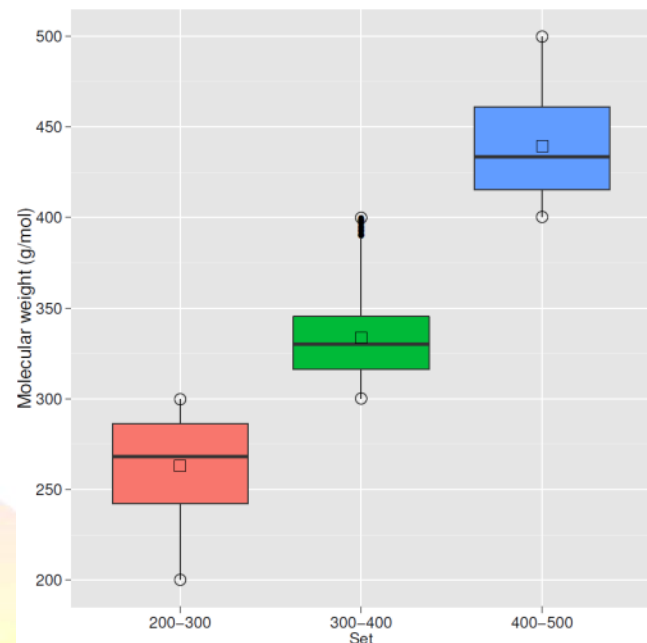
3000 ligands

3 molecular weight groups

[200,300], [300,400], [400,500]

1000 ligands each group

**8.69x ~
37.51x**



Program	200-300/mol		300-400/mol		400-500/mol	
	CPU	Elapsed	CPU	Elapsed	CPU	Elapsed
1HCL human cyclin-dependent kinase 2						
Vina	12.57	3.33	22.55	5.91	51.62	13.41
idock	0.63	0.16	0.92	0.24	1.38	0.36
Ratio	20.06	20.25	24.41	24.39	37.51	36.81
1J1B human tau protein kinase I						
Vina	9.07	2.47	14.69	3.92	32.28	8.49
idock	0.78	0.21	1.25	0.33	2.35	0.62
Ratio	11.55	11.92	11.73	11.87	13.73	13.73
1LI4 human S-adenosylhomocysteine hydrolase						
Vina	11.82	3.30	19.08	5.22	39.41	10.64
idock	0.89	0.23	1.55	0.40	3.15	0.82
Ratio	13.24	14.14	12.33	12.95	12.50	12.98
1V9U human rhinovirus 2 coat protein VP1						
Vina	9.80	2.95	15.55	4.62	29.75	8.49
idock	0.97	0.25	1.64	0.42	3.42	0.89
Ratio	10.11	11.74	9.49	10.91	8.69	9.56
2IQH influenza A virus nucleoprotein NP						
Vina	9.51	2.66	15.03	4.08	29.64	7.83
idock	0.92	0.24	1.59	0.41	3.41	0.88
Ratio	10.35	11.18	9.43	9.93	8.69	8.93
2XSK Escherichia coli curli protein CsgC - SeCys						
Vina	10.44	2.71	17.89	4.61	40.58	10.41
idock	0.71	0.19	1.16	0.30	2.16	0.56
Ratio	14.68	14.64	15.47	15.38	18.83	18.57
2ZD1 HIV-1 reverse transcriptase						
Vina	9.78	2.70	17.67	4.76	42.03	11.33
idock	0.97	0.25	1.52	0.39	2.60	0.69
Ratio	10.05	10.73	11.61	12.07	16.14	16.54
2ZNL influenza virus RNA polymerase subunit PA						
Vina	9.49	2.60	15.04	4.01	29.97	7.82
idock	0.89	0.23	1.56	0.40	3.41	0.87
Ratio	10.70	11.37	9.65	10.06	8.78	8.98
3BGS human purine nucleoside phosphorylase						
Vina	9.59	2.57	16.50	4.37	38.42	10.14
idock	0.95	0.25	1.55	0.40	2.81	0.74
Ratio	10.09	10.45	10.65	10.89	13.65	13.75
3H0W human S-adenosylmethionine decarboxylase						
Vina	9.85	2.64	17.67	4.70	41.69	11.04
idock	0.88	0.23	1.35	0.35	2.20	0.58
Ratio	11.17	11.50	13.07	13.28	18.99	19.11
3IAR human adenosine deaminase						
Vina	11.25	3.03	20.21	5.39	46.93	12.53
idock	0.80	0.21	1.21	0.32	2.01	0.53
Ratio	14.10	14.44	16.68	16.90	23.34	23.59
3KFN HIV protease						
Vina	10.53	2.80	18.37	4.83	42.43	11.03
idock	0.77	0.20	1.20	0.32	2.09	0.55
Ratio	13.69	13.85	15.29	15.32	20.32	20.12
Average across the above 12 receptors						
Vina	10.31	2.81	17.52	4.70	38.73	10.26
idock	0.85	0.22	1.38	0.36	2.58	0.67
Ratio	12.48	13.02	13.32	13.66	16.76	16.89

Availability

🌀 <https://github.com/HongjianLi/idock>

🌀 Free, C++, Apache License 2.0

🌀 32bit & 64bit Linux, Windows, Mac, FreeBSD, Solaris



github Explore Gist Blog Help HongjianLi

PUBLIC **HongjianLi / idock** Pull Request Unwatch Unstar 1 Fork 0

Code Network Pull Requests 0 Issues 0 Wiki Graphs Admin

idock is a multithreaded virtual screening tool for flexible ligand docking for computational drug discovery. — [Read more](#)
<http://istar.cse.cuhk.edu.hk/idock>

Clone in Windows ZIP **HTTP** SSH Git Read-Only <https://github.com/HongjianLi/idock.git> SSH HTTPS Git

branch: master **Files** Commits Branches 1 Tags 5 Downloads

idock / 341 commits

File	Time	Description
bin	2 months ago	Recompiled idock 1.6 for Windows on Windows 8 [HongjianLi]
examples	2 months ago	Added a new example 2VQZ [HongjianLi]
ligands	a month ago	Removed a large ligand from the ZINC folder [HongjianLi]
obj	10 months ago	Reverted obj/.gitignore [HongjianLi]
receptors	2 months ago	Removed non-polar hydrogens for 2VQZ receptor. Added MGT as the native [HongjianLi]

FULL REPRODUCIBILITY

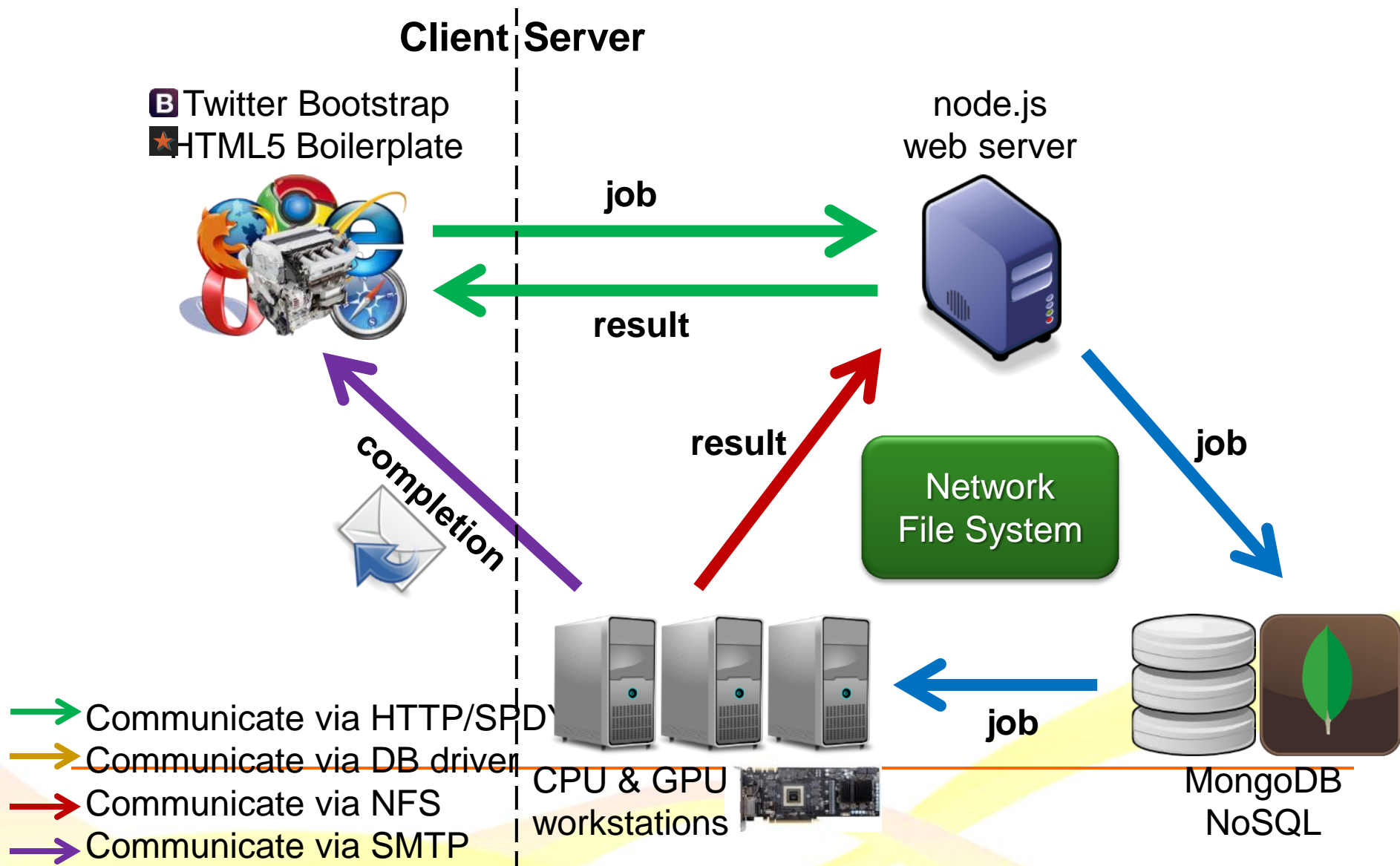


istar

Software as a Service



<http://istar.cse.cuhk.edu.hk>

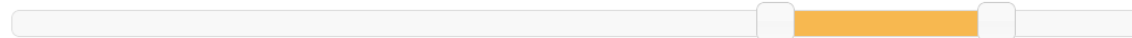


Ligand Filtering and Previewing

- Filter ligands with desired molecular properties
- Preview the number of ligands to dock

Number of ligands satisfying all the 9 filtering conditions: 188,820

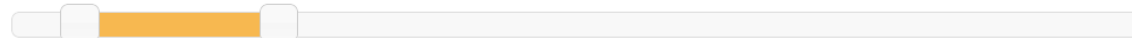
Molecular weight (g/mol): [400, 500]



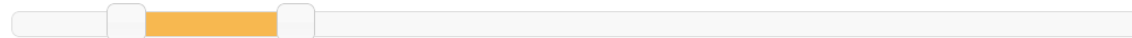
Partition coefficient xlogP: [0, 5]



Rotatable bonds: [2, 8]



Hydrogen bond donors: [2, 5]



Hydrogen bond acceptors: [2, 10]



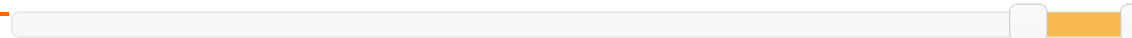
Net charge: [0, 0]



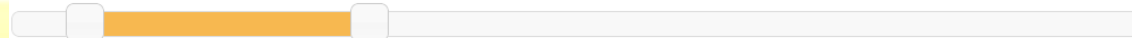
Apolar desolvation (kcal/mol): [0, 12]



Polar desolvation (kcal/mol): [-50, 0]




Polar surface area tPSA (\AA^2): [20, 100]



Real-Time Progress

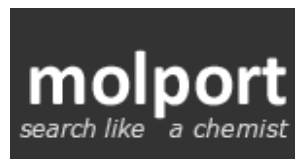
- Monitor job progress in real time
 - Progress reporting mechanism in daemon
 - Ajax timer and table

Your jobs

Ligands	Submitted on	Status	Progress	Result
8	2012/10/13 21:50:51	Done on 2012/10/13 22:13:29	100.00000%	
1,590,058	2012/10/20 19:57:54	Phase 1 in progress	0.07289%	

Supplier Output

Help purchase compounds from vendors 



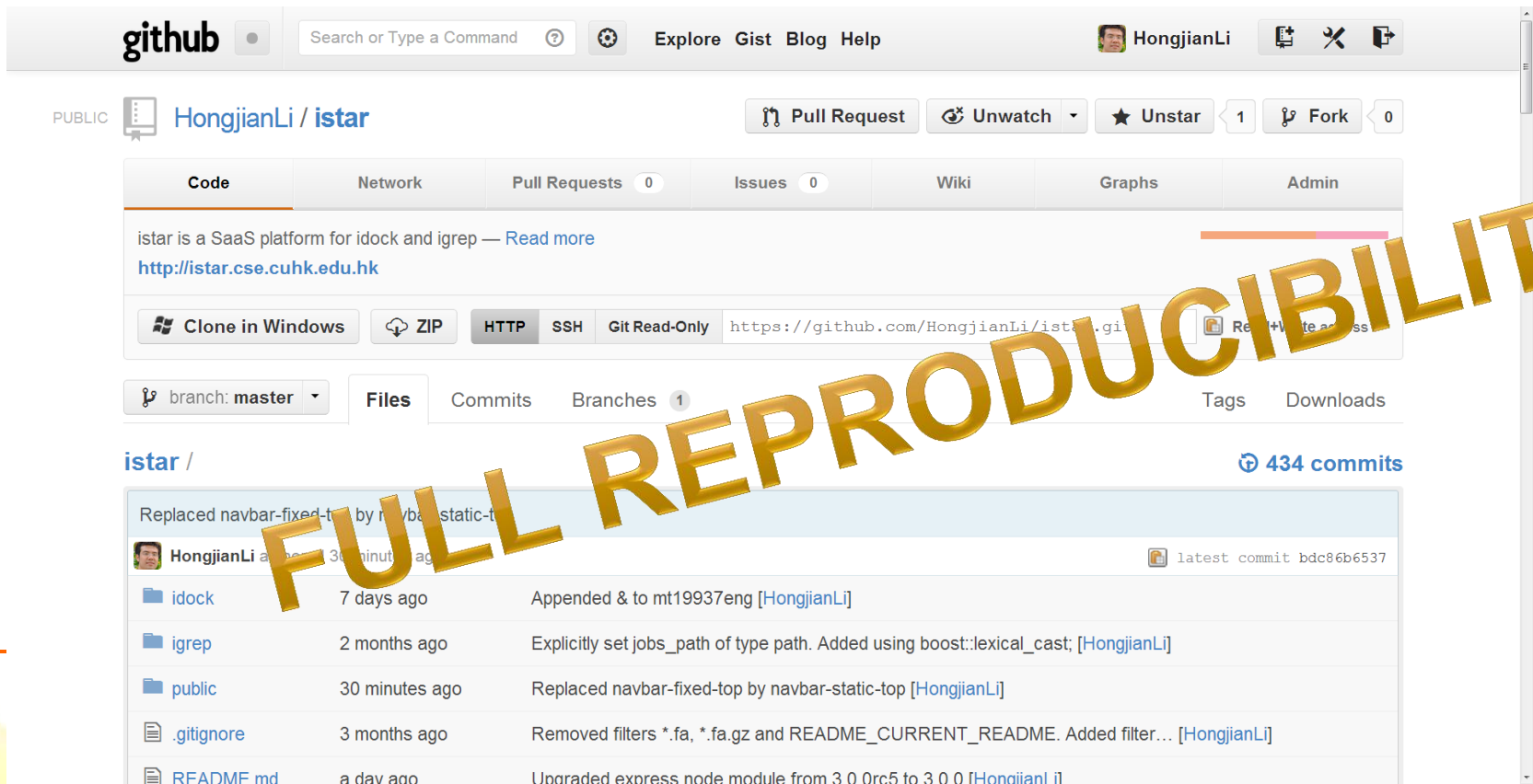
	AZ	BA
1	Substance information	Suppliers
2	http://zinc.docking.org/substance/25922195	1 uorsy
3	http://zinc.docking.org/substance/67742829	5 ambint chbr chemonaut emol molport
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		

Availability

🌀 <https://github.com/HongjianLi/istar>     

🌀 Free, Apache License 2.0, Javascript and C++

🌀 Chrome 19+, Firefox 12+, IE9+, Safari 5+, Opera 12+



github Search or Type a Command ? ? Explore Gist Blog Help HongjianLi

PUBLIC HongjianLi / **istar** Pull Request Unwatch Unstar 1 Fork 0

Code Network Pull Requests 0 Issues 0 Wiki Graphs Admin

istar is a SaaS platform for idock and igrep — [Read more](#)
<http://istar.cse.cuhk.edu.hk>

Clone in Windows ZIP HTTP SSH Git Read-Only <https://github.com/HongjianLi/istar.git> Refresh Write access

branch: master Files Commits Branches 1 Tags Downloads

istar / 434 commits

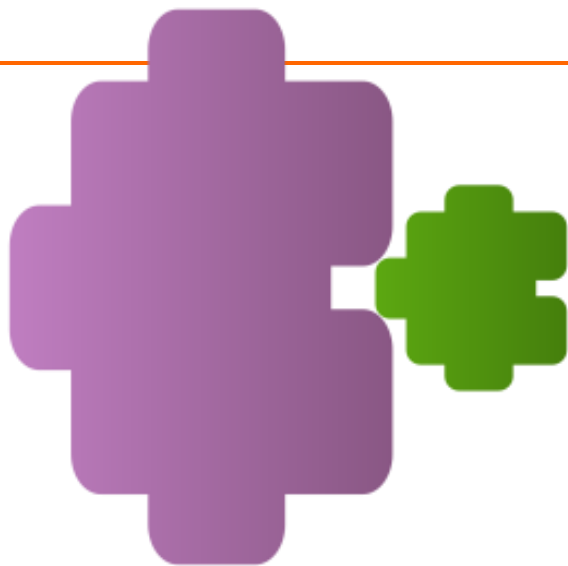
File	Commit	Message
idock	7 days ago	Appended & to mt19937eng [HongjianLi]
igrep	2 months ago	Explicitly set jobs_path of type path. Added using boost::lexical_cast; [HongjianLi]
public	30 minutes ago	Replaced navbar-fixed-top by navbar-static-top [HongjianLi]
.gitignore	3 months ago	Removed filters *.fa, *.fa.gz and README_CURRENT_README. Added filter... [HongjianLi]
README.md	a day ago	Upgraded express node module from 3.0.0rc5 to 3.0.0 [HongjianLi]

FULL REPRODUCIBILITY



idock 2.0

GPU Acceleration



NVIDIA GK104 Block Diagram

☞ GTX 680 US\$593

☞ 3.09 TFLOPS SP

☞ 128 GFLOPS DP

☞ 2GB GDDR5

☞ PCIE 3.0 192GB/s

☞ TDP 195W

☞ 4 GPCs

☞ 4 raster engines

☞ 8 SMX units

☞ 1536 CUDA cores



AMD Tahiti Block Diagram

€7970 US\$516

€3.79 TFLOPS SP

€947 GFLOPS DP

€3GB GDDR5

€264GB/s

€TDP 250W

€32 GCN cores

€2048 stream

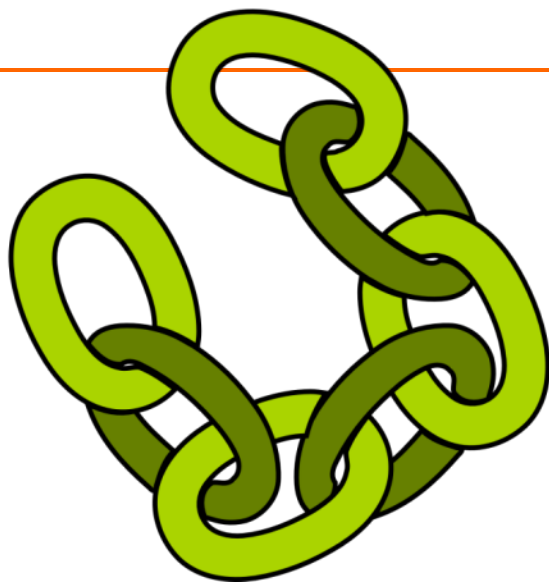
processors





idock 3.0

De Novo Ligand Design

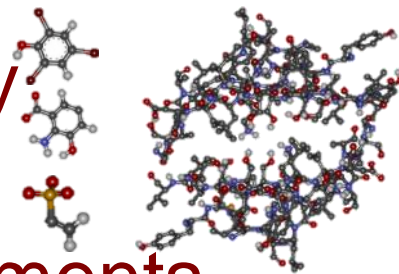


Motivation

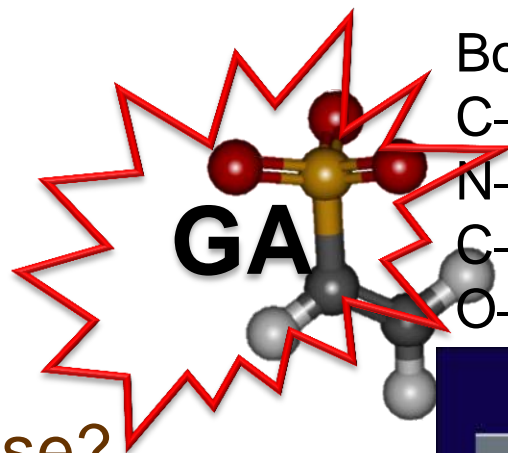
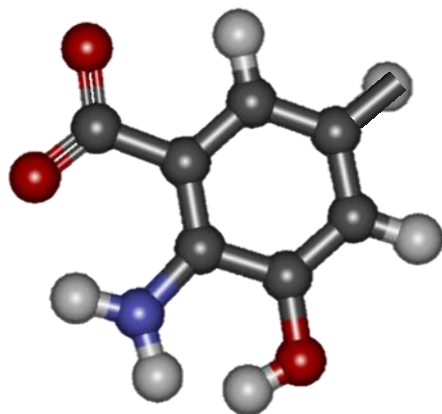
Virtual screening \rightarrow *de novo* strategy

$\approx 10^{60} - 10^{100}$ drug-like molecules

Grow an initial scaffold by adding fragments



Design ligands that have higher binding affinities



Bond length

C–C: 1.530 Å

N–N: 1.425 Å

C–N: 1.469 Å

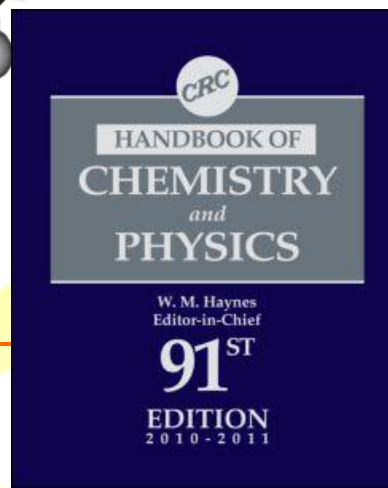
O–O: 1.469 Å

Which fragment to choose?

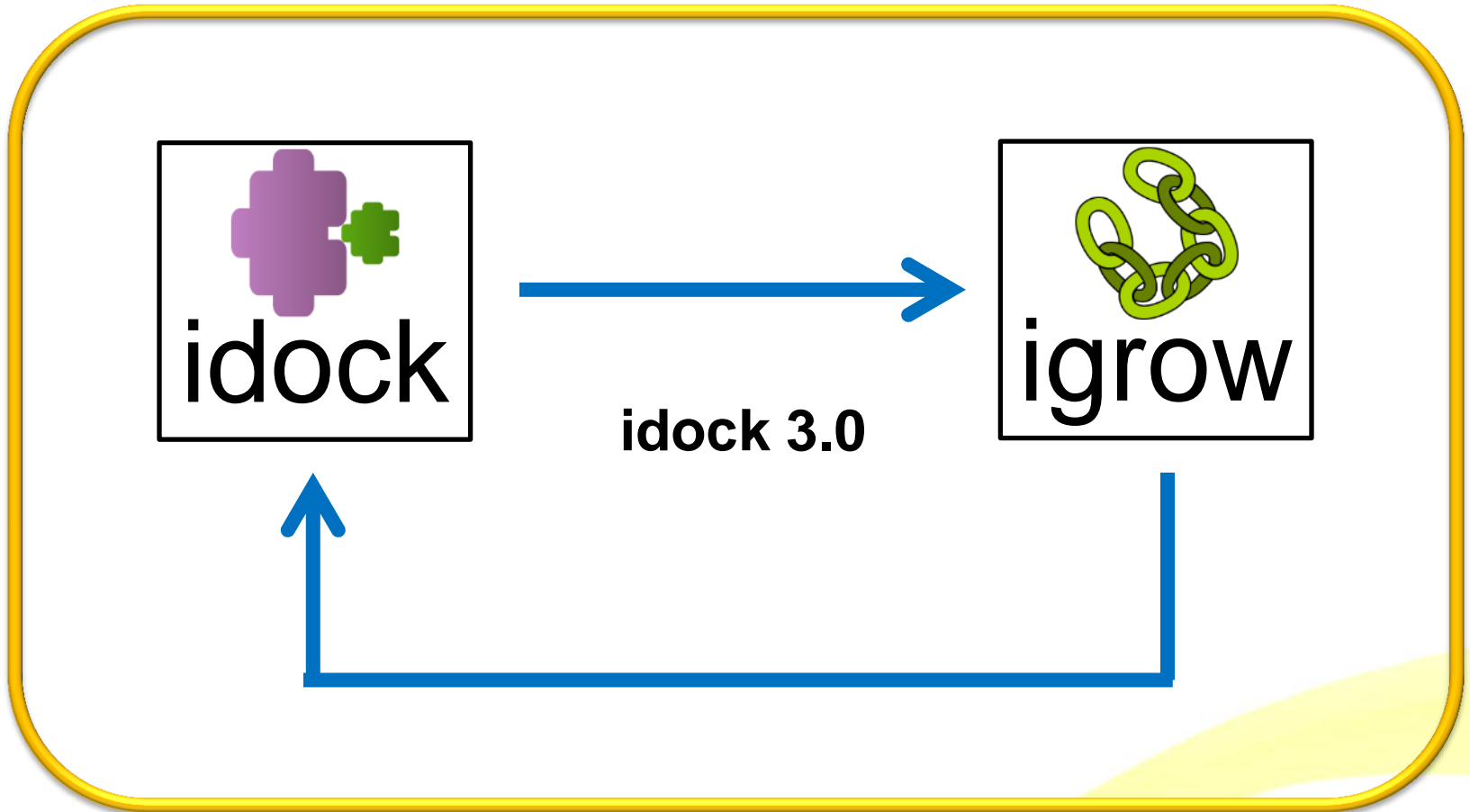
Which linker atom to choose?

How to join the fragment in 3D?

Combinatorial optimization problem

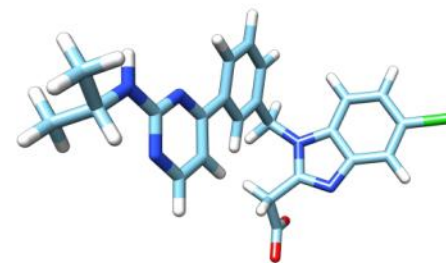
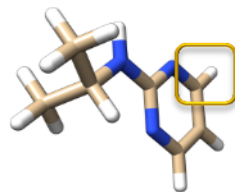
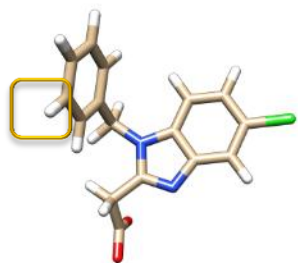


Genetic Operator: Selection



Genetic Operator: Addition

☞ Merge a ligand and a fragment



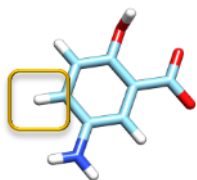
● Elitist 1

● Fragment

● Elitist 2

-4.457 kcal/mol

-7.818 kcal/mol



● Elitist 2

● Fragment

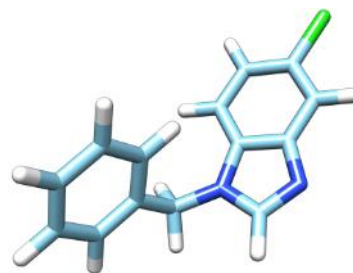
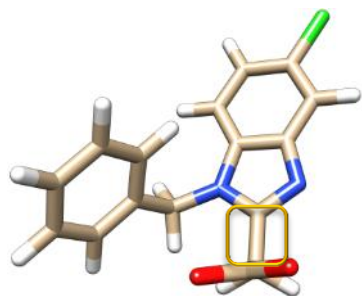
● Elitist 3

-7.818 kcal/mol

-9.043 kcal/mol

Genetic Operator: Subtraction

Drop part of a ligand

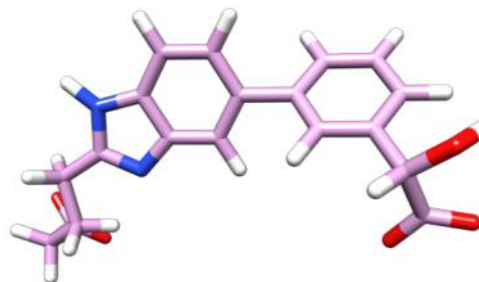
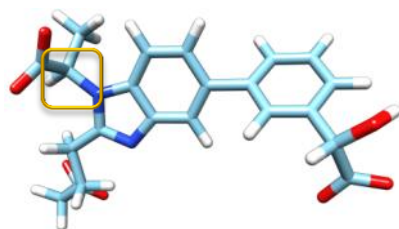


● Elitist 1

-4.457 kcal/mol

● Elitist 2

-7.818 kcal/mol



● Elitist 3

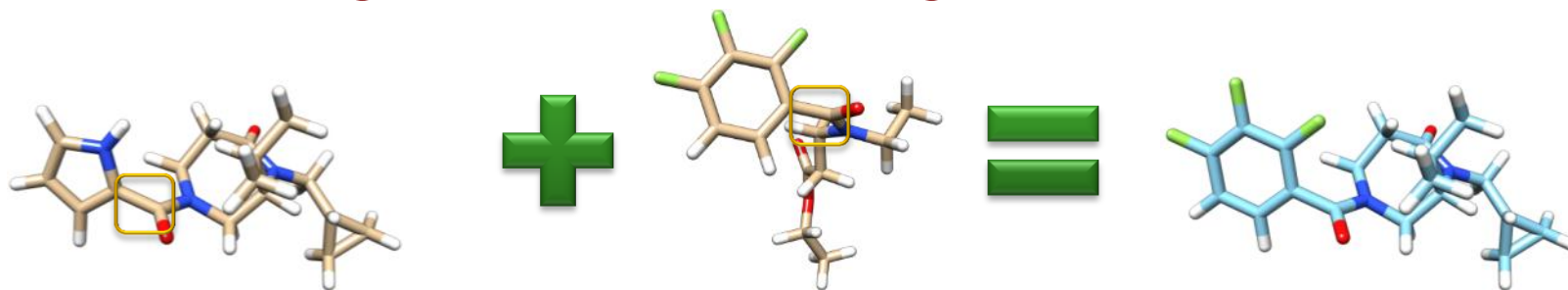
-7.818 kcal/mol

● Elitist 4

-9.043 kcal/mol

Genetic Operator: Crossover

Exchange parts of two ligands

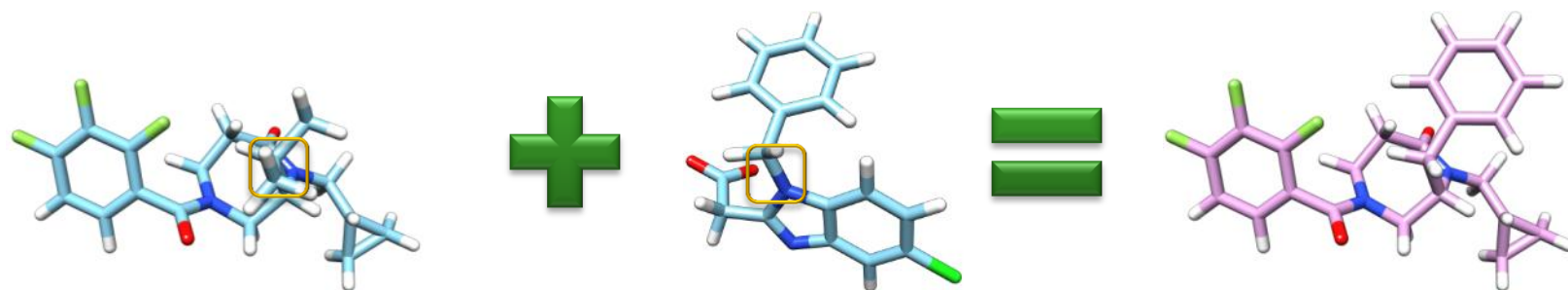


● Elitist 1

● Elitist 2

● Elitist 3

-3.072 kcal/mol -5.027 kcal/mol -7.337 kcal/mol



● Elitist 3

● Elitist 4

● Elitist 5

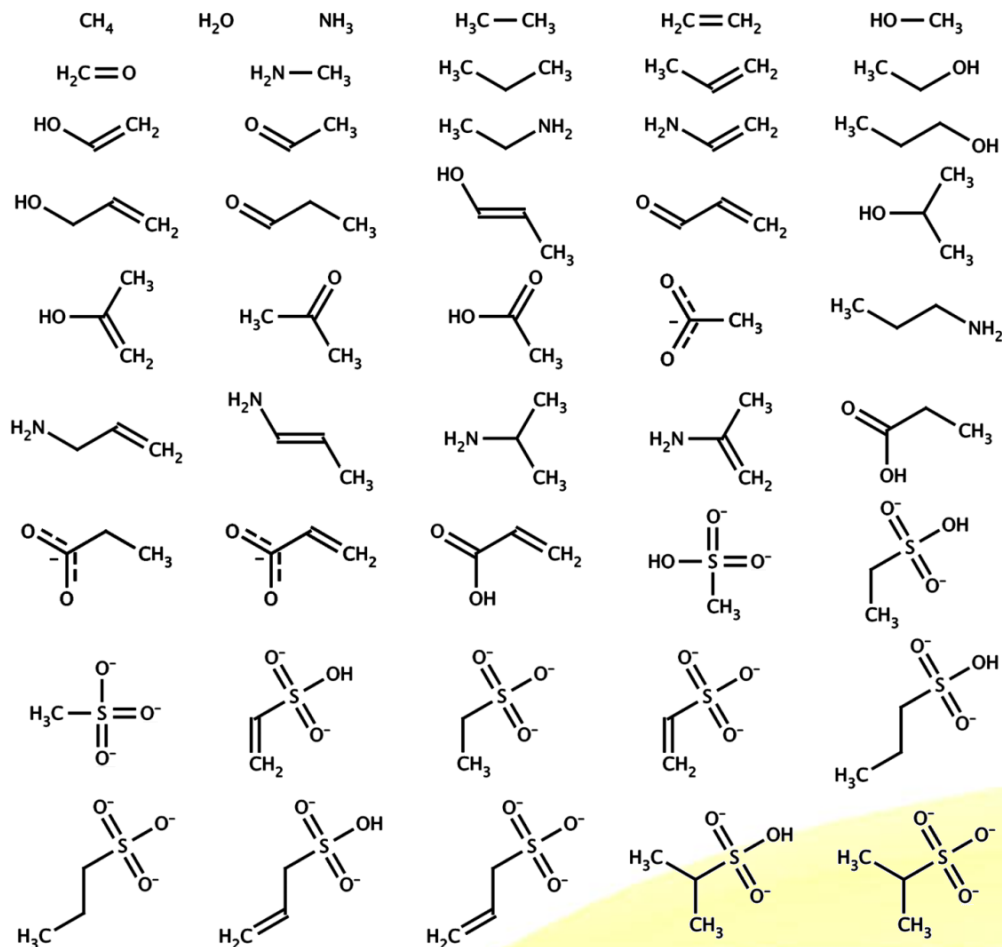
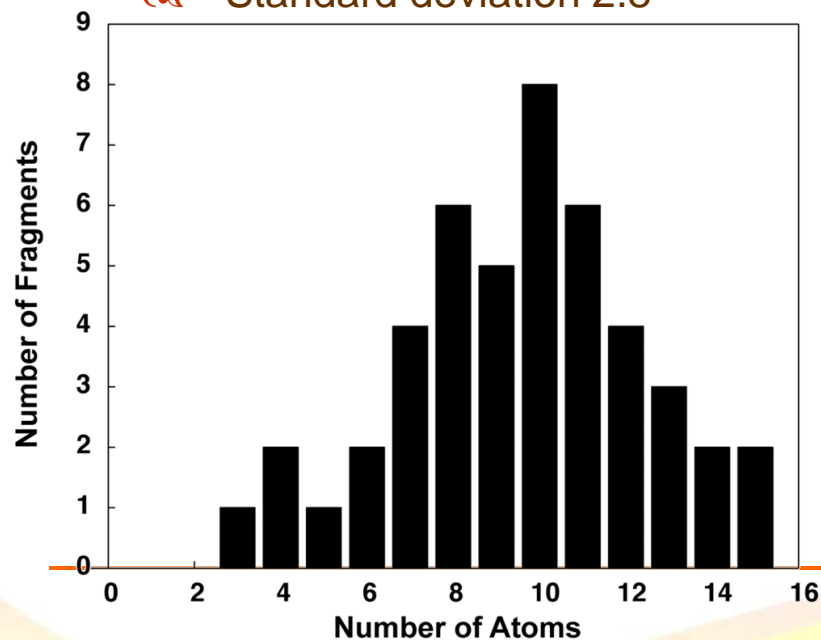
-7.337 kcal/mol -6.126 kcal/mol -8.200 kcal/mol

Fragment Library

☞ An Example of

Small-fragment library

- ☞ Provided by AutoGrow
- ☞ 46 fragments
- ☞ 3 to 15 atoms
- ☞ Average 9.6 atoms
- ☞ Standard deviation 2.8

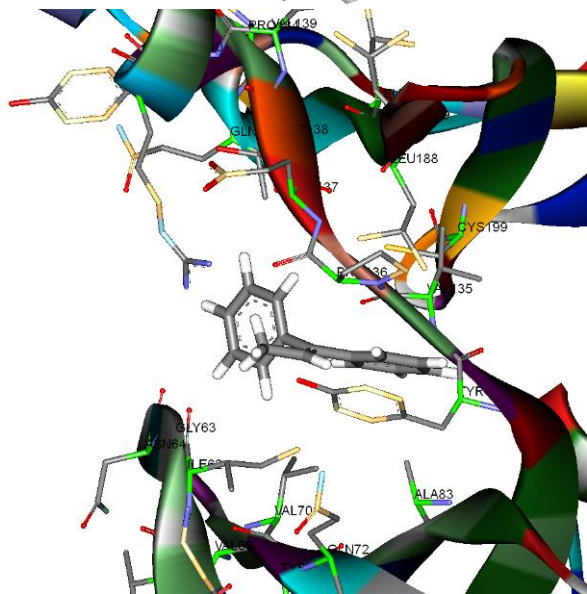
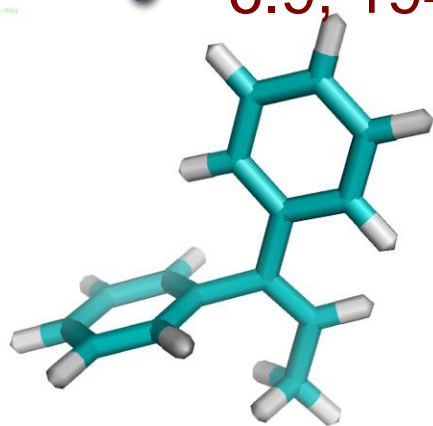


Initial Results: GSK3 β -ZINC01019824

GSK-3 β inhibitor reduces Alzheimer's pathology and rescues neuronal loss

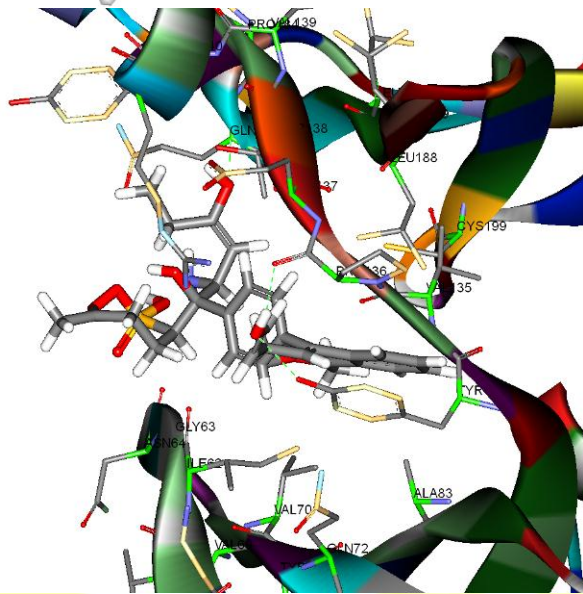
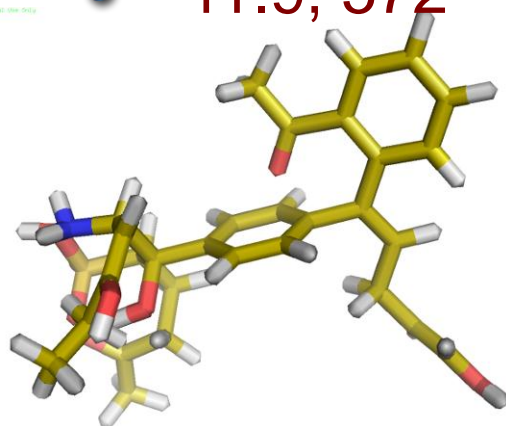
Initial ligand

-6.9, 194



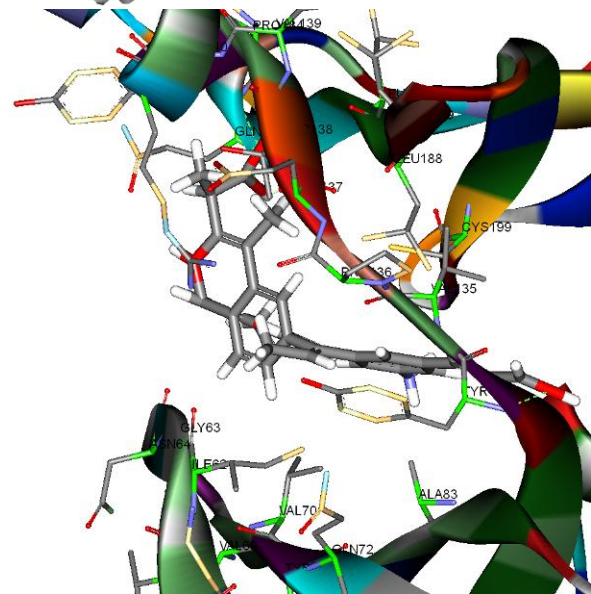
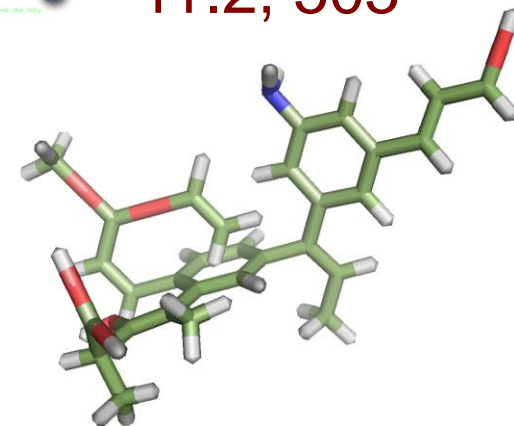
AutoGrow

-11.9, 572



iGrow

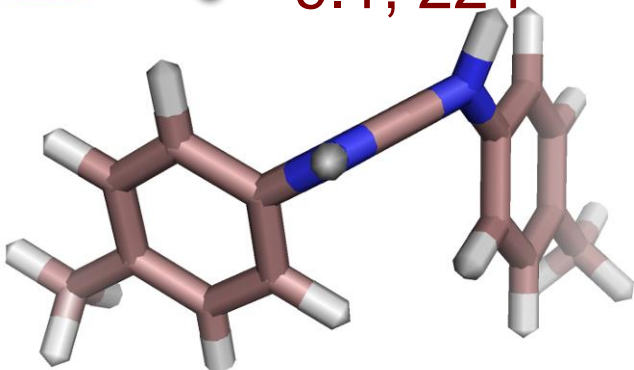
-11.2, 505



Results: HIV RT-ZINC08442219

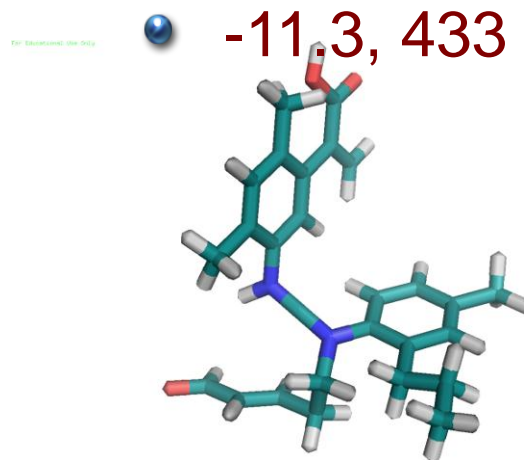
● Initial ligand

● -9.1, 224



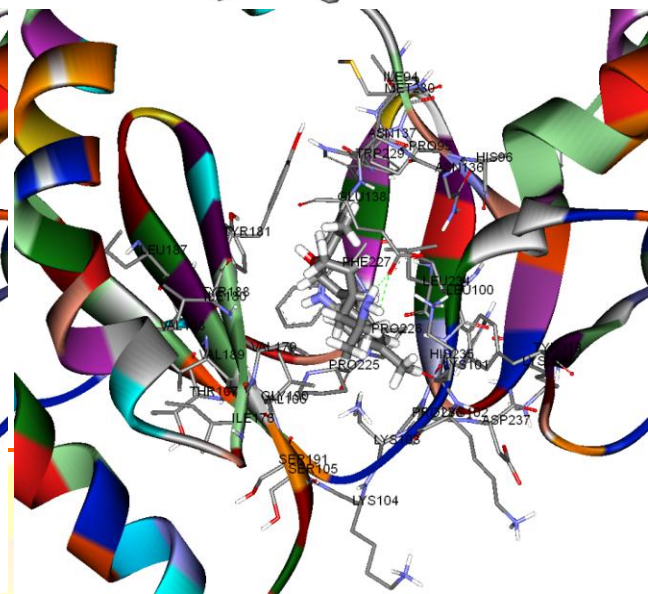
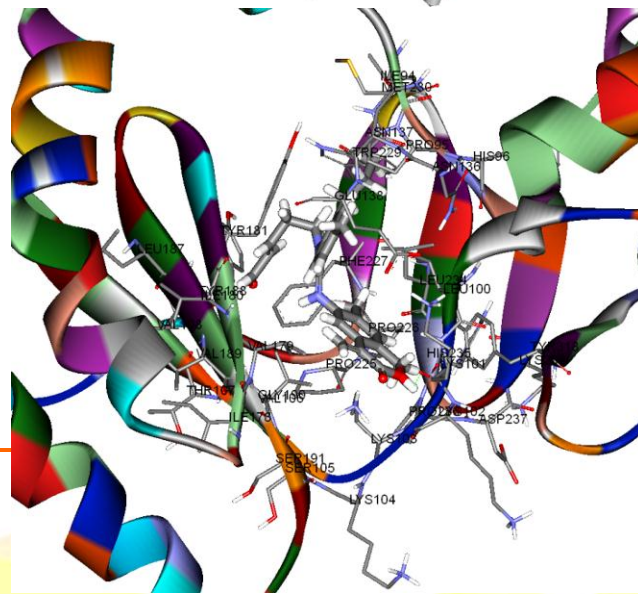
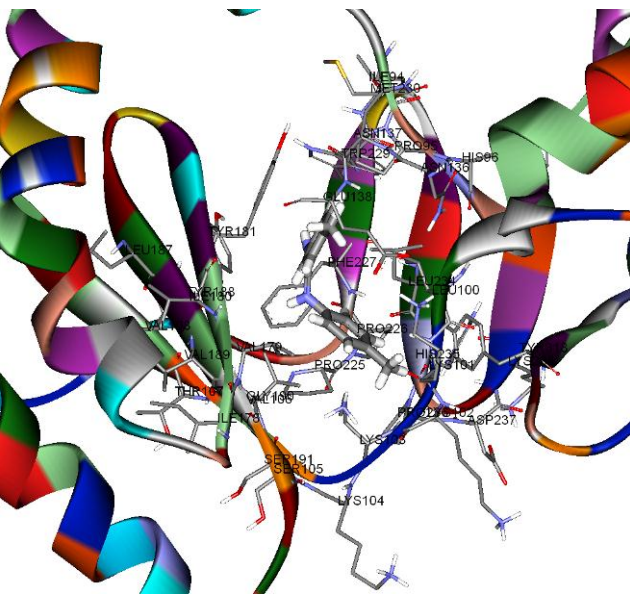
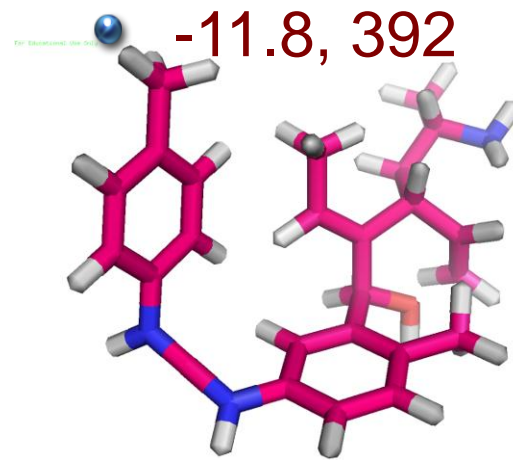
● AutoGrow

● -11.3, 433



● iGrow

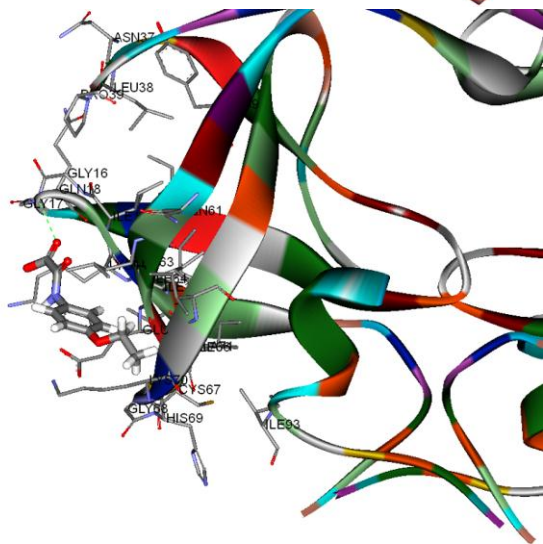
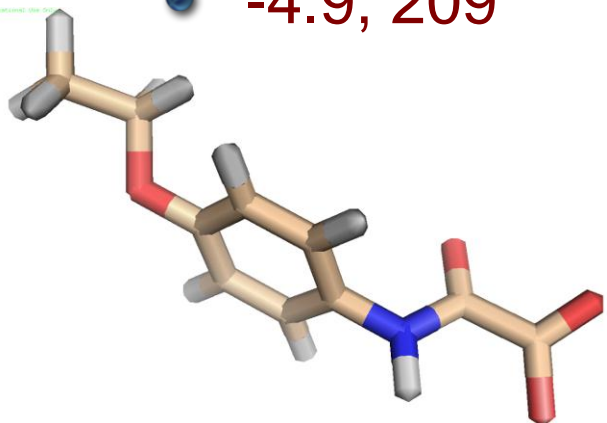
● -11.8, 392



Results: HIV PR-ZINC20030231

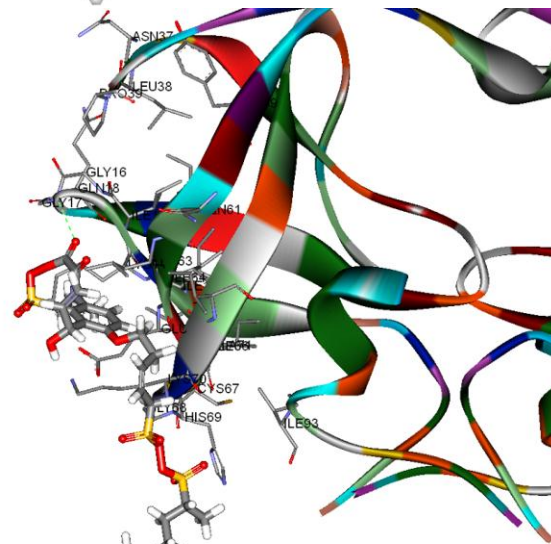
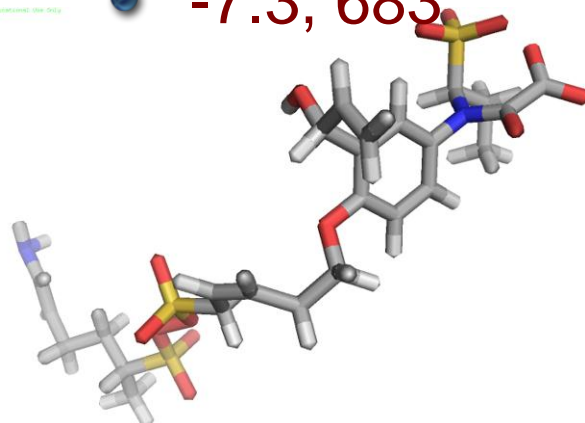
● Initial ligand

● -4.9, 209



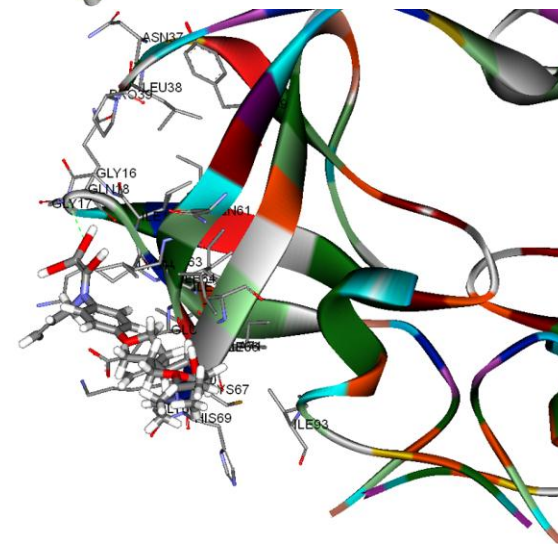
● AutoGrow

● -7.3, 683



● iGrow

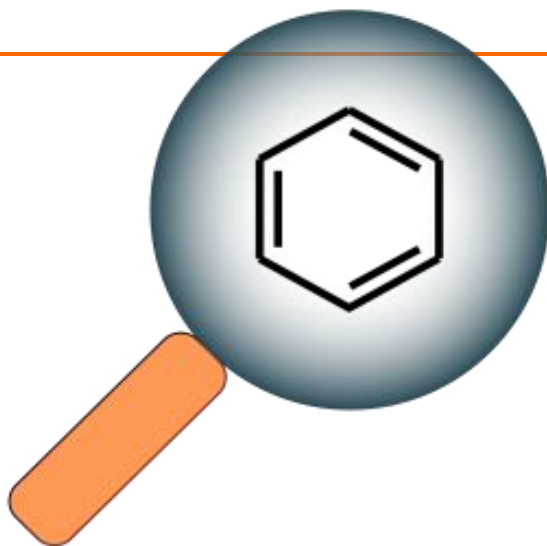
● -7.5, 489





iview

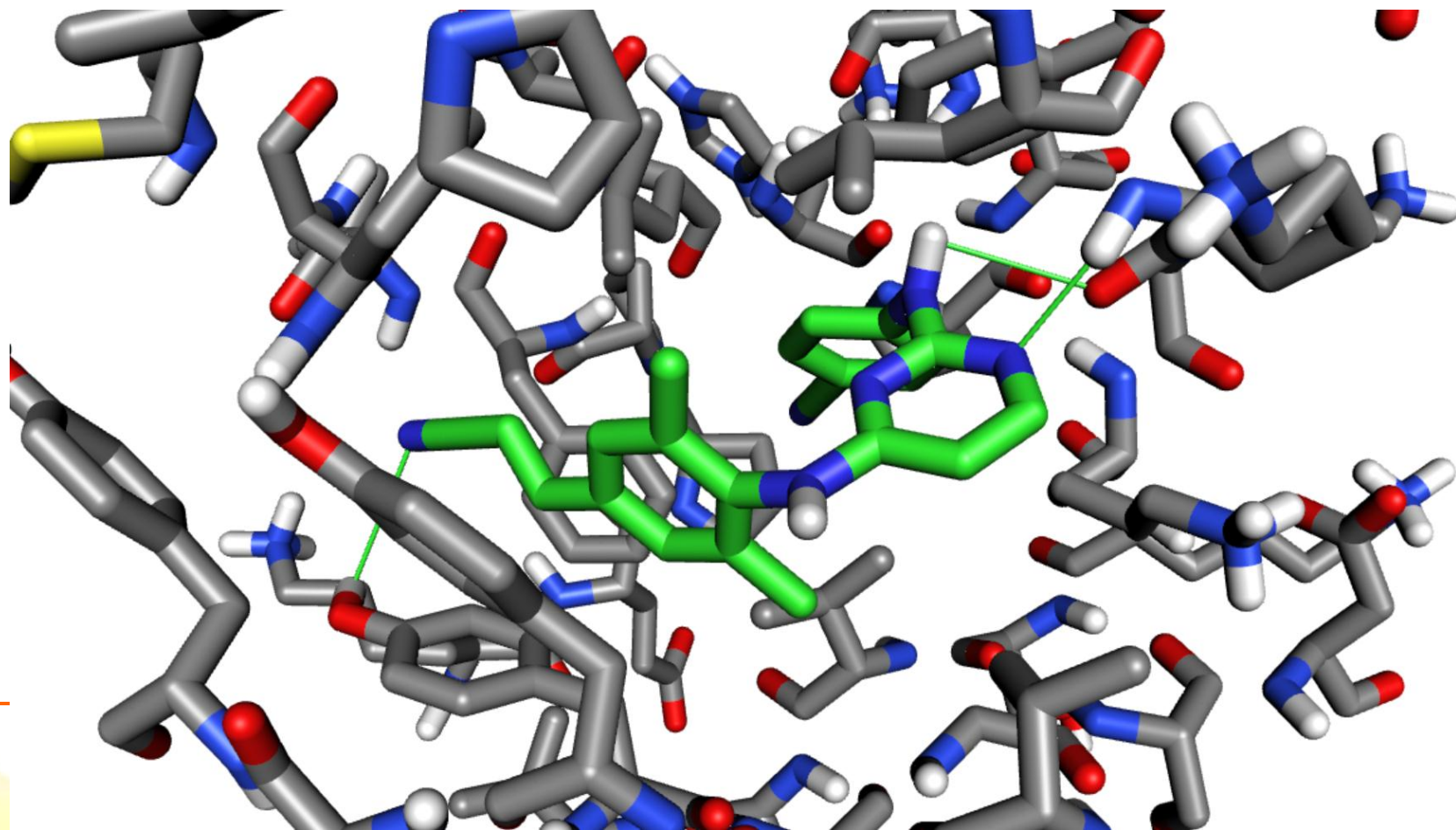
HTML5 Visualizer



Interactive HTML5 Visualizer

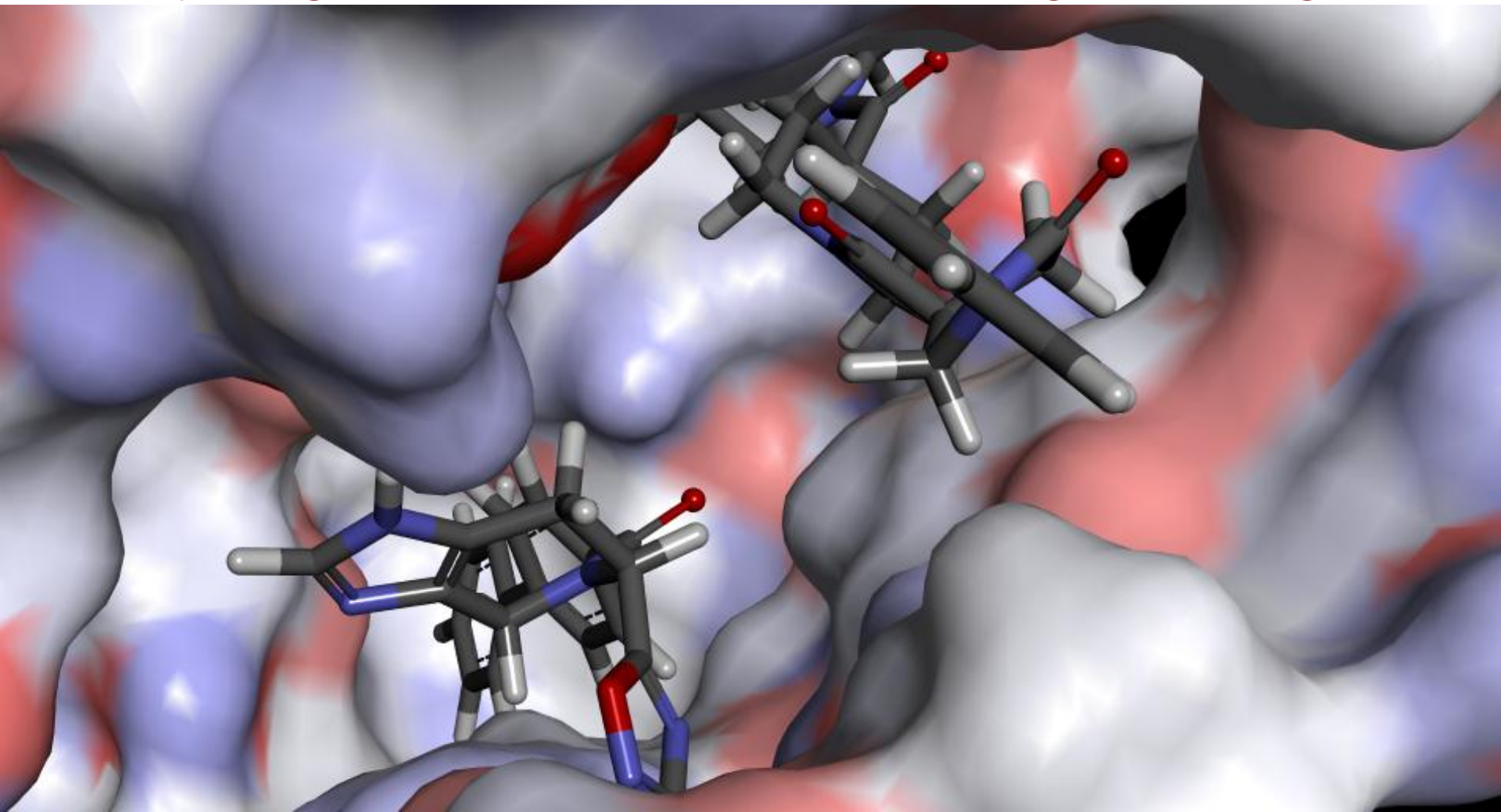
Based on canvas and WebGL

First HTML5 visualizer of protein-ligand complex



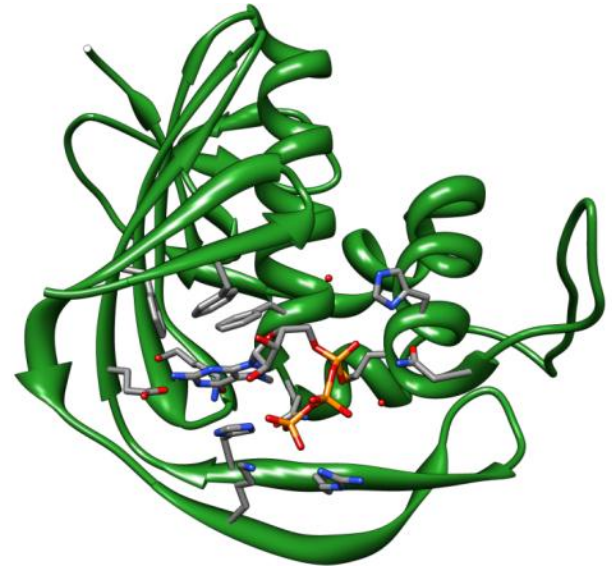
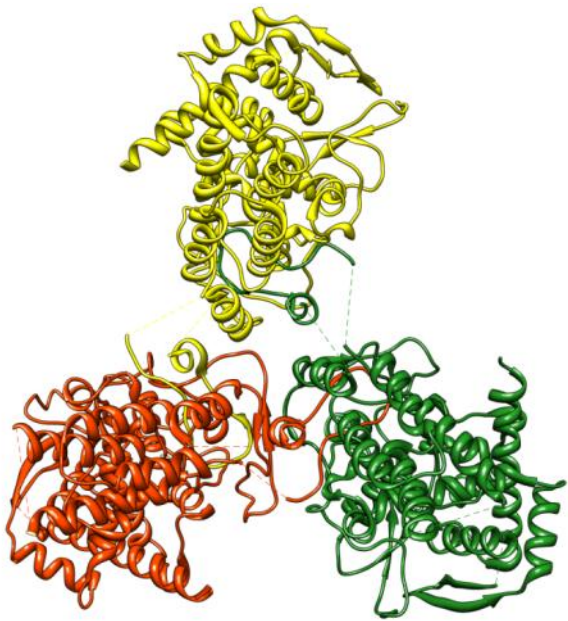
Dual Ligand Docking

☞ Synergistic effect, suitable for large binding





Case Study of Influenza A



Background

WHO fact sheets

250K–500K deaths, 3M–5M severe illness annually

Drug resistance

Proteins	Functions	Binding sites	Inhibitors
HA	Virus attachment to sialic acid receptors on host cell surface; fusion of virus and cell membranes	Sialic acid binding site; TBHQ binding site	Neu5Ac; TBHQ
NA	Cleavage of sialic acid receptors to release progeny viruses from host cells	Active site	Zanamivir; oseltamivir
M2	Acidification and uncoating of endosome-entrapped virus; virus assembly and budding	Inside pore near Ser31	Amantadine; rimantadine
NP	Capsidation of viral RNA and binding of three polymerase subunits to form ribonucleoprotein particles	Tail-loop binding site; RNA binding site	– ^a
Polymerase	Viral RNA transcription and replication	PA: endonuclease active site; PB1 binding site PB1: polymerase active site PB2: cap binding site; importing binding site	–
M1	Structural component of virion; nuclear export of ribonucleoprotein particles	NEP binding site	–
NEP	Nuclear export of ribonucleoprotein particles from host-cell nucleus	Crm1 binding site; M1 binding site	–
NS1	Protection against host-cell antiviral responses	Double-stranded RNA binding site; CPSF30 binding site	–

Our Progress

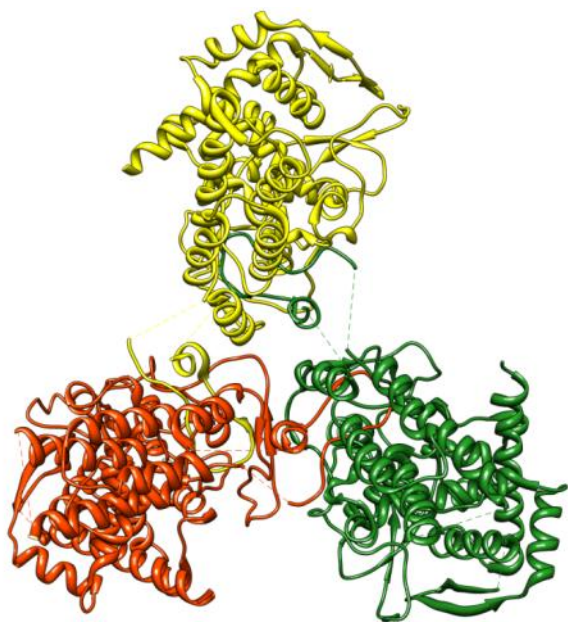
☞ Nucleoprotein

☞ 2IQH

☞ idock 1.4

☞ 4 Mac@CSE

☞ 7M ligands



☞ Polymerase PA

☞ 2ZNL

☞ idock 1.5

☞ 1 Mac@CSE

☞ 73K ligands



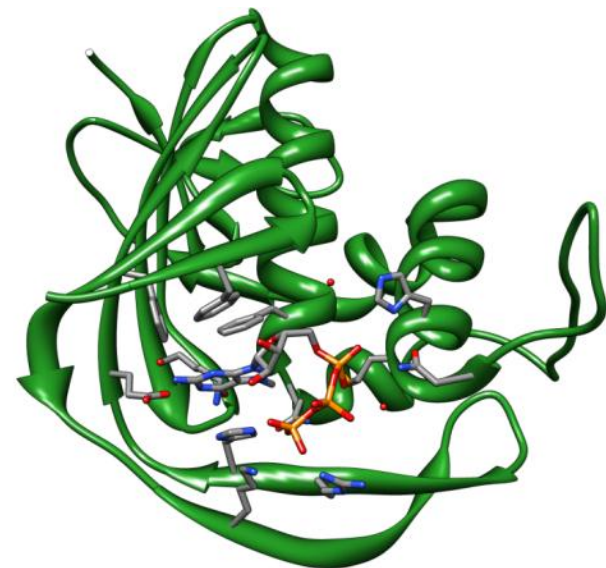
☞ PolymerasePB2

☞ 2VQZ

☞ idock 1.6

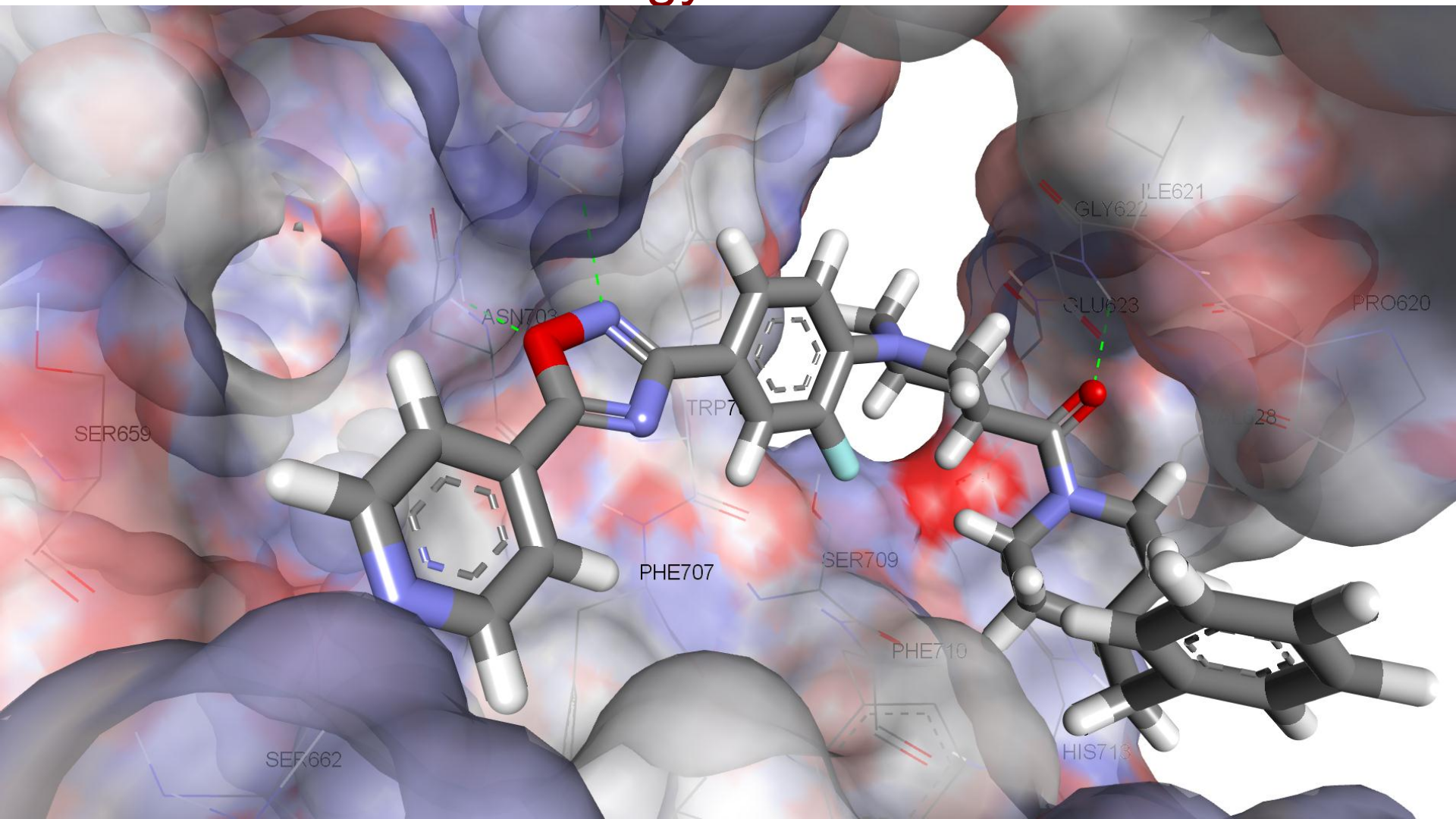
☞ 2 Linux@ITSC

☞ 2M ligands



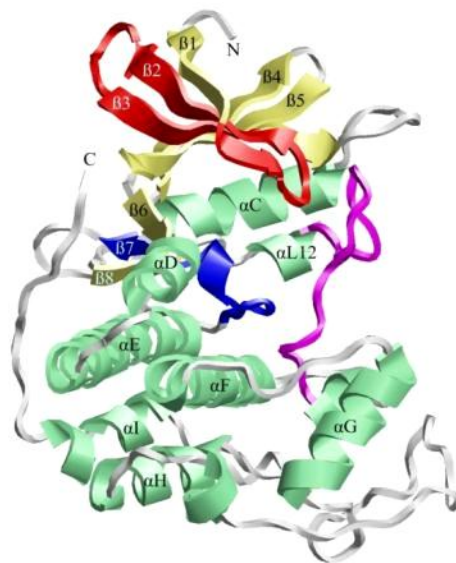
Polymerase PA w/ ZINC40879809

∞ Predicted free energy -11.465 kcal/mol





Case Study of CCRK-Related Cancers



CCRK (Cell Cycle-Related Kinase)

CCRK aliases: p42, PNQLARE, CDK20

4 transcript variants by alternative splicing

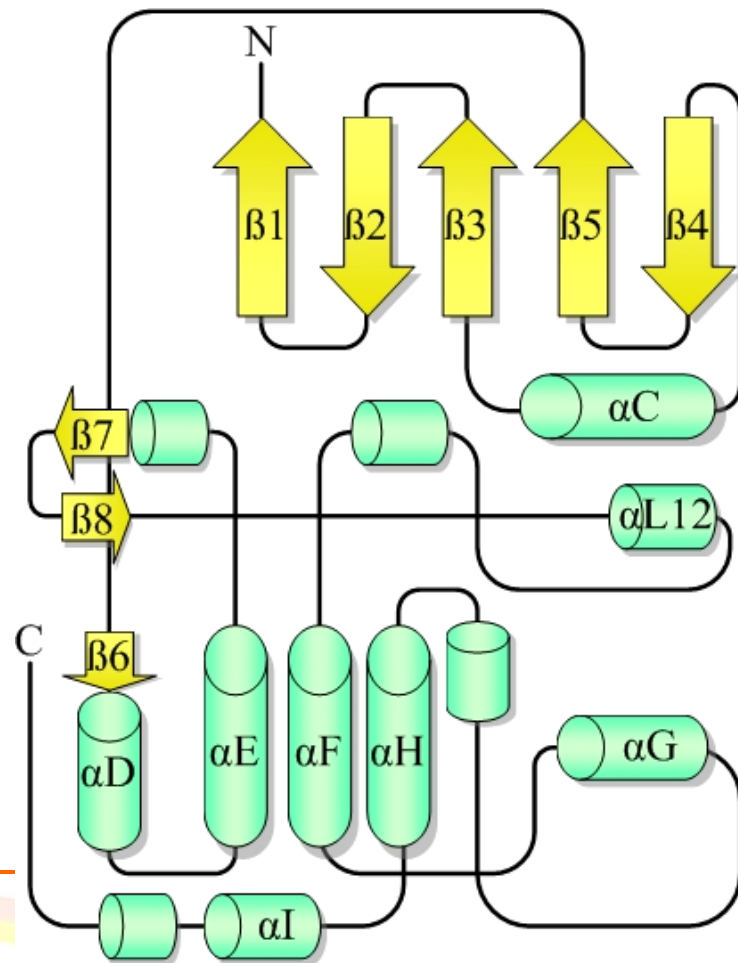
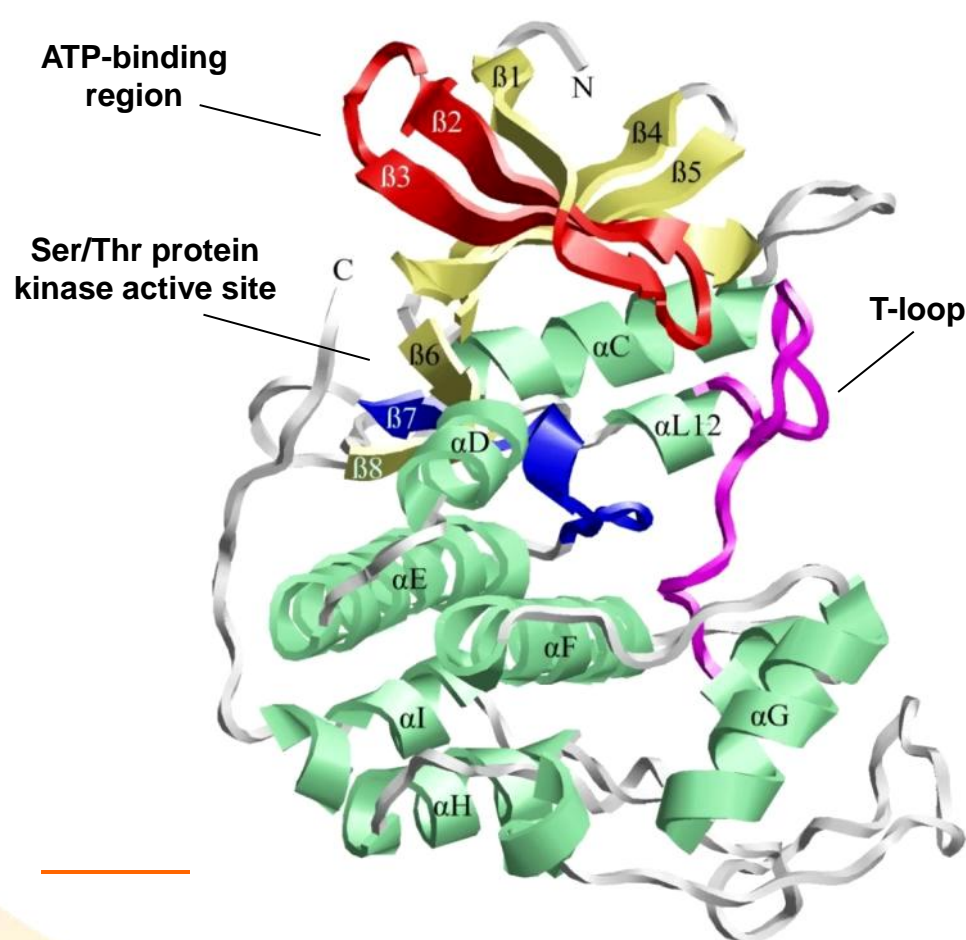


Widely expressed in various cancers

Glioblastoma, cervical adenocarcinoma, colorectal carcinoma, osteogenic sarcoma, breast adenocarcinoma, ovarian carcinoma, lung fibroblast, myoblast, and lymphocyte

CCRK Homology Model from 1HCL

Done with SWISS-MODEL

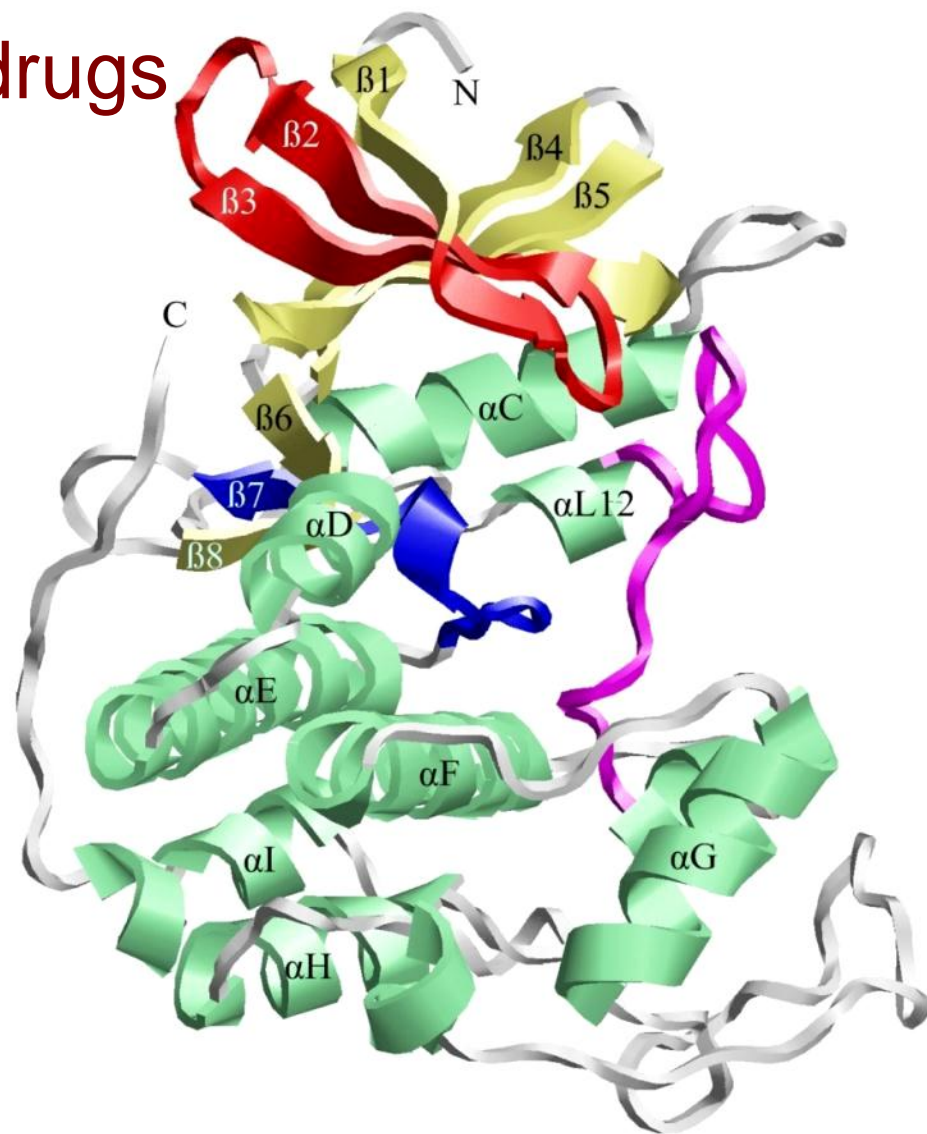


Our Progress

☞ Repurpose approved drugs

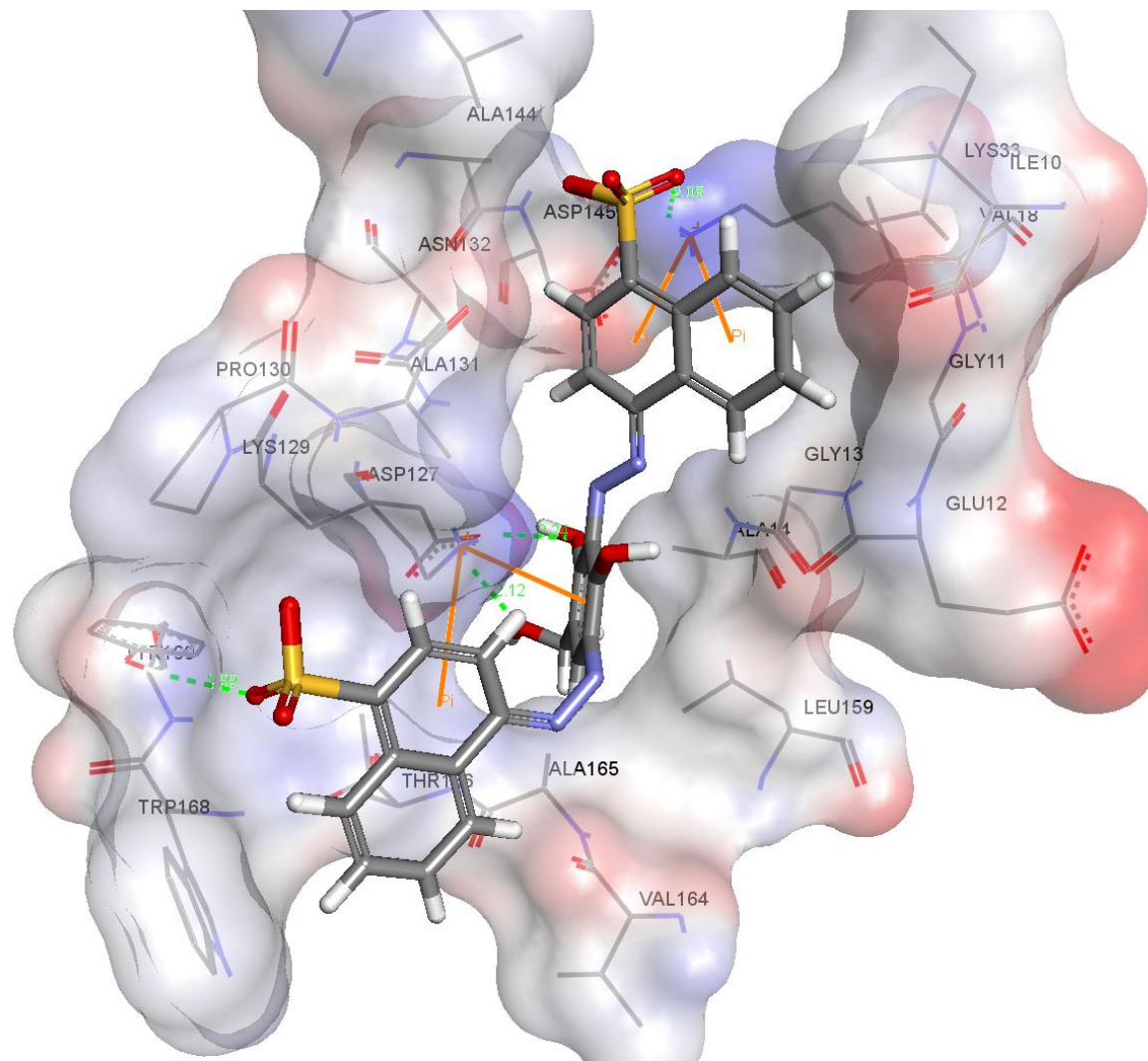
☞ 1,715 via DrugBank

☞ 3,176 via DSSTOX



CCRK in complex w/ ZINC03830332

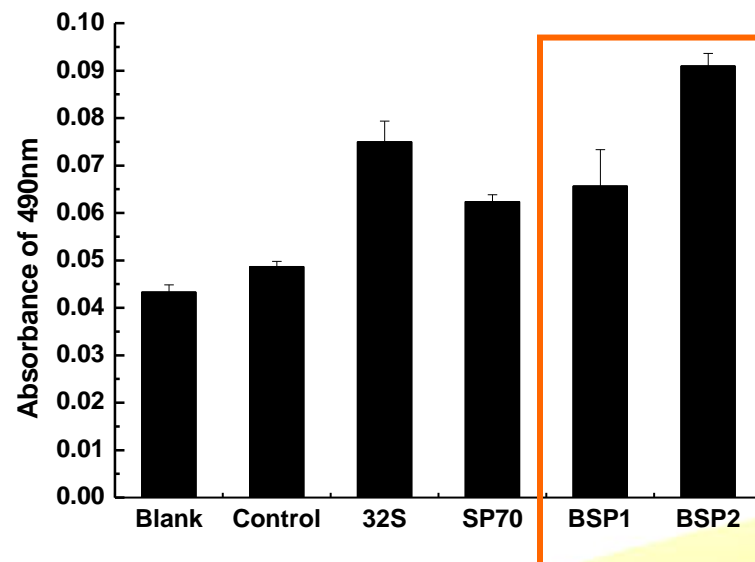
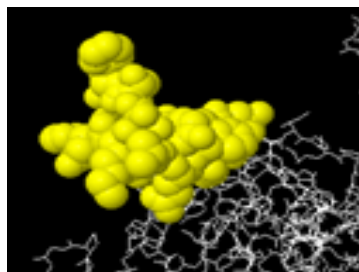
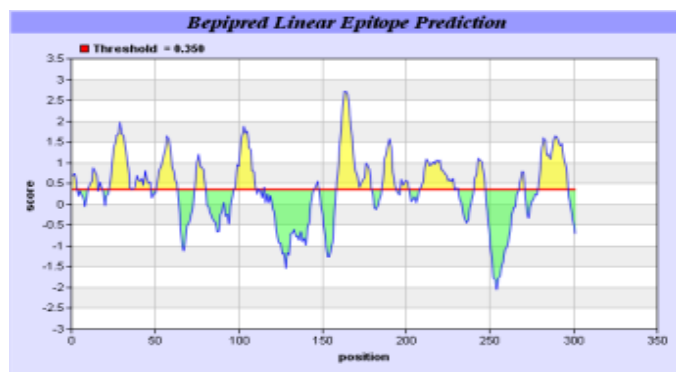
☞ Predicted free energy -10.306 kcal/mol



Other Drug Discovery Results

Current collaboration with biomedical experts:

--combined prediction helps identify a novel B-cell epitope with the best wet-lab immune responses, a potential vaccine for EV71 (hand foot and mouth disease)



Immune response in mouse, higher the better
 BSP1-Computational control (conformational only)
 BSP2-Shortlisted epitope (best combined result)
 32S, SP70: known and documented epitopes

IV. Discussion and Conclusion

☞ Summary

☞ Discussion

Summary

☞ In this talk

- ☞ A brief introduction to Bioinformatics research problems
- ☞ Discovering approximate protein-DNA interaction sequence patterns for better understanding gene regulation (the essential control mechanisms of life)
- ☞ Computer-aided drug discovery via protein-ligand docking and de novo ligand design. Case studies on influenza and cancers.
- ☞ Encouraging results have been achieved and promising direction has been pointed out

Discussion

- ❧ Bioinformatics becomes more and more important in life sciences and biomedical applications
- ❧ Most computational fields (ranging from string algorithms to graphics) have applications in Bioinformatics
- ❧ Still long way to go (strong potentials to explore)
 - ❧ Massive data are available but annotations are still limited
 - ❧ Lack of full knowledge in many biological mechanisms
 - ❧ Biological systems are very complicated and stochastic

Selected Publications (2008-now)

- T.M. Chan, K.S. Leung, K.H. Lee, M.H. Wong, C.K. Lau, Stephen K.W. Tsui, Subtypes of Associated Protein-DNA (Transcription Factor-Transcription Factor Binding Site) Patterns, *Nucleic Acids Research*, 2012, 40 (19), pp. 9392-9403 (IF:8.026)
- Po-Yuen Wong, Tak-Ming Chan, Man-Hon Wong and Kwong-Sak Leung, Predicting Approximate Protein-DNA Binding Cores Using Association Rule Mining, In Proceedings of *IEEE ICDE 2012*, pp. 965-976 (Acceptance Rate: 17.7%).
- T.M. Chan, K.S. Leung, K.H. Lee, "Memetic Algorithms for de novo Motif Discovery". *IEEE Transactions on Evolutionary Computation*, 2012, 16(5), pp. 730-748.
- T.M. Chan, K.C. Wong, K.H. Lee, M.H. Wong, C.K. Lau, Stephen K.W. Tsui, K.S. Leung, Discovering approximate-associated sequence patterns for protein-DNA interactions. *Bioinformatics*, 2011, 27(4), pp. 471-478. (IF:5.468)
- (S.K. LOU), J.W. LI, H. QIN, Aldrin K.Y. YIM, L.Y. Lo, Bing Ni, K.S. Leung, Stephen K.W. TSUI, and T.F. CHAN, "Detection of splicing events and multiread locations from RNA-seq data based on a geometric-tail (GT) distribution of intron length", *BMC Bioinformatics*, 2011.07.27Vol12 suppl.5 S2.
- Leung, KS, (Wong, KC), (Chan, TM), Wong, MH, Lee, KH, Lau, CK, and Tsui, Stephen, "Discovering Protein-DNA Binding Sequence Patterns Using Association Rule Mining," *Nucleic Acids Research*. 2010, 38(19), pp. 6324-6337.
- (S.K. Lou)†, (B. Ni)†, L.Y. Lo, Stephen K.W. Tsui, T.F. Chan and K.S. Leung, "ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping", *Bioinformatics*, Oxford Journal, 2010.02.01 †co-1st authors
- (T.M. Chan), (G. Li), K.S. Leung and K.H.Lee, Discovering multiple realistic TFBS motifs based on a generalized model, *BMC Bioinformatics*, 2009, 10:321
- (G. Li), (T.M. Chan), K.S. Leung and K.H.Lee, A Cluster Refinement Algorithm for Motif Discovery, *IEEE/ACM Transaction on Computational Biology and Bioinformatics*. pp.654-668., 2010.10.01
- KS Leung, KH Lee, (JF Wang), (Eddie YT Ng), Henry LY Chan, Stephen KW Tsui, Tony SK Mok, C.H. Tse, Joseph JY Sung, "Data Mining on DNA Sequences of Hepatitis B Virus". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. vol.8 no.2, pp.428-40. 2011.03.
- (Chan, T.M.), Leung, K.S., and Lee, K.H., "TFBS identification based on genetic algorithm with combined representations and adaptive post-processing." *Bioinformatics*, Vol.24, No.3, pp341-349, Oxford Journals, Feb 2008
- Hongjian Li, Kwong-Sak Leung, and Man-Hon Wong. idock: A Multithreaded Virtual Screening Tool for Flexible Ligand Docking. 2012 IEEE Symposium on CIBCB, pp.77-84, 2012.
- C. M. Tse, H. J. Li, K. S. Leung, K. H. Lee, and M. H. Wong. Iterative Drug Design in Virtual Reality. 15th International Conference on Information Visualisation (IV), pp.226-231, 13-15 2011.

10 Related Patents including:

- SUNG Joseph Jao Yiu; CHAN Lik Yuen Henry; TSUI Kwok Wing; LEUNG Kwong Sak; et al. "Genomic Markers of Hepatitis B Virus in Hepatocellular Carcinoma". United States Patent no. US7439020B2. U.S.A, 2008.10.21.
- SUNG Jao Yiu, Joseph; CHAN Lik Yuen, Henry; TSUI Kwok Wing, Stephen; LEUNG Kwong Sak; et al. "Genomic Markers of Hepatitis B Virus Associated with Hepatocellular Carcinoma." United States Patent no. US7871780. U.S.A, 2011.01.18.

The End

☞ Thank you!

☞ Q&A

Appendix

Introduction: Bridging

II: Results and Analysis: Statistical Significance

II: Results and Analysis

Statistical Significance ($W=5$)

- Simulated on over 100,000 rules for each setting
- The majority (**64%-79%**) for $R_{TF-TFBS}$ are statistically significant
- For $E=0$, although the $0.05 < p(R_{TF} \geq 1) < 0.07$, the majority (**74%-82%**) achieve the best possible p-values

TY R_*	$W = 5, E = 0$						$W = 5, E = 1$					
	0.0		0.1		0.3		0.0		0.1		0.3	
	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS	TF	TF-TFBS
P-value < 0.05	0 (127*)	110	0 (165*)	147	0 (636*)	567	223	226	278	272	1974	2023
Rule No.	172	172	211	211	774	774	346	346	396	396	2559	2559
Significant Ratio	0 (0.74*)	0.64	0 (0.78*)	0.70	0 (0.82*)	0.73	0.64	0.65	0.70	0.69	0.77	0.79