Robust Regression Analysis of Protein-Ligand Free Energy Models: Virtual Screening and Target Identification via Grid and Cloud Computing

Jung-Hsin Lin
Division of Mechanics, Research Center for Applied Sciences and Institute of Biomedical Sciences, Academia Sinica
School of Pharmacy, National Taiwan University

http://rx.mc.ntu.edu.tw/~jlin/

Seminar in 2010 South East Asia International Program, December 9, 2010

Taiwan Pharmaceutical Databank (TPD)

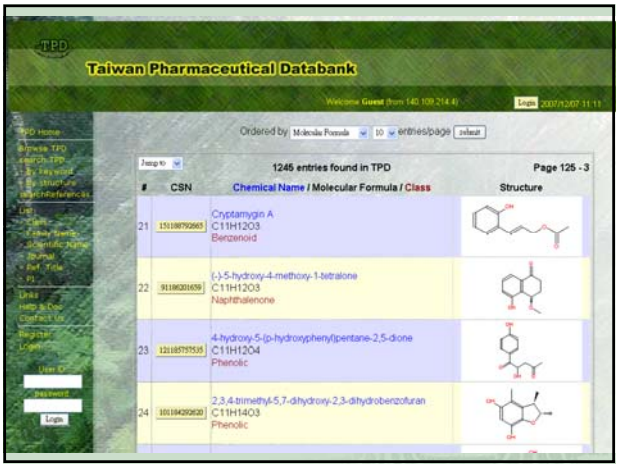
http://tpd.mc.ntu.edu.tw

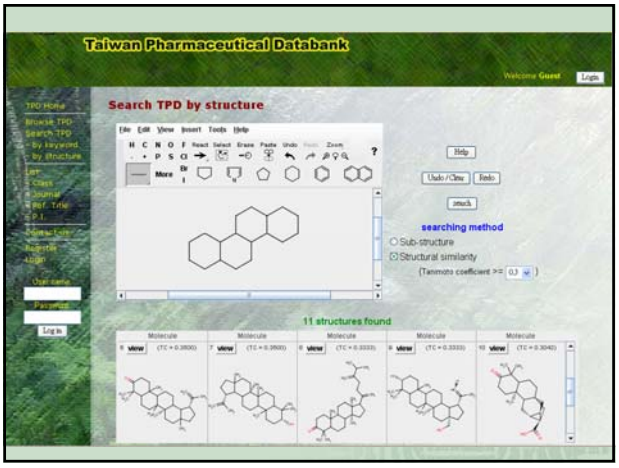
- Web-based Database
  LAMP: Linux OS, Apache web server, MySQL relational database, PHP
  Dynamic HTML, JAVA scripts
  JAVA applets are used for structure drawing and presentation
  JME and Marvin from ChemAxon
- User interface in Chinese and/or in English
- Currently TPD contains entries for
  - 3600+ compounds which are classified into 130 classes
  - 200+ natural sources which are from 84 families
  - 540+ references which are published in 57 journals

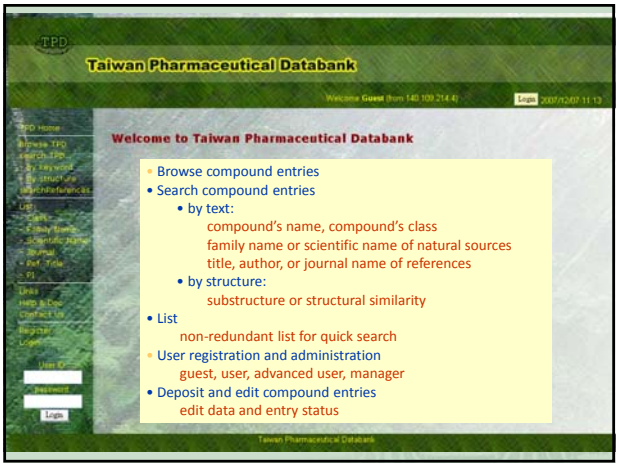
Taiwan Pharmaceutical Databank

Comparison with Beilstein database

- Chinese information also included
- Compounds studied by Taiwanese scholar
- Classification for compounds included
- Natural sources specified and the collected time and location documented.
- Capability of displaying spectrum images
- Annotated PDF files available for TPD manager and/or user







Taiwan Pharmaceutical Databank

The information for each compound entry may include:

Structure and Nomenclature
- structure (still image and interactive presentation), compound's class
- chemical name, IUPAC name, synonym, CAS Registry Number

• Physical and Chemical Properties
- molecular formula, molecular weight, melting point, solubility, store condition

• Biological Activity and Toxicity
- text and/or image

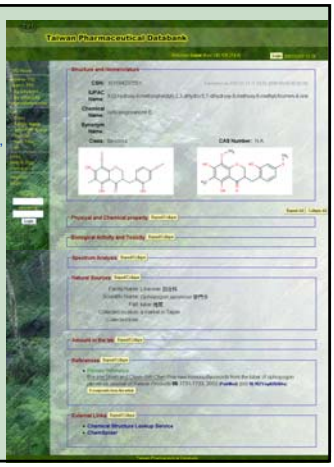
• Spectrum Analysis
- text and/or image

• Natural Sources
- family name, scientific name, part used, collected time and location

• Amount in the lab

• References
- reference information and links to PubMed and/or DOI if available

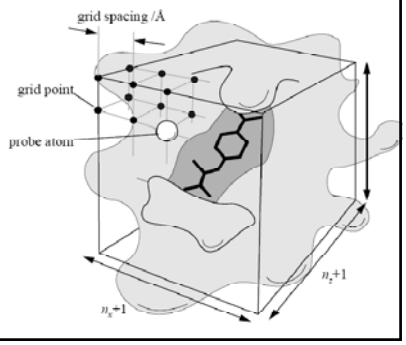
• External links
- Chemical Structure Lookup Service, ChemSpider



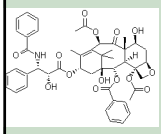
Finding Good Drug Targets in Druggable Genomes

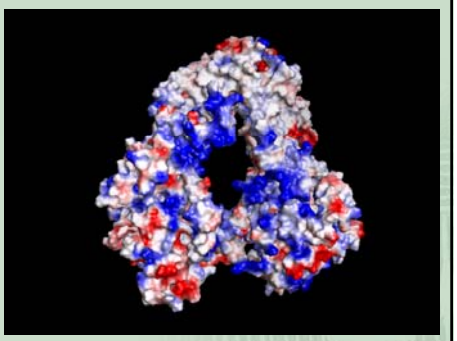
- Not all proteins are druggable: not essential in signaling pathways, not easily accessible to drugs, no suitable binding pocket, etc.
- Too hydrophobic pockets will lead to too hydrophobic ligands, causing solubility or even permeability problems.
- No protein cavity or crevices available for binding ligands tightly .

The Flexible Docking Problem



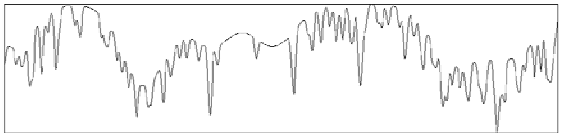
Prediction of binding pose could be challenging: the case of P-glycoprotein with Paclitaxel (Taxol)





Characteristics of Biological Complex Problems

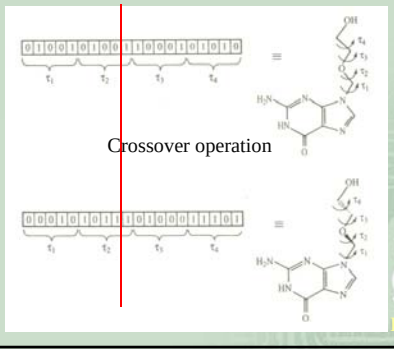
- The potential energy function is extremely rugged.
- The potential energy surface is usually highly asymmetric.
- The true global minimum is often surrounded by many deceptive local minima.
- The biological complex problems are mostly in the space of high dimensionality.



Genetic Algorithm

1. **[Start]** Generate random population of *n* chromosomes (suitable solutions for the problem)
2. **[Fitness]** Evaluate the fitness *f(x)* of each chromosome *x* in the population
3. **[New population]** Create a new population by repeating following steps until the new population is complete
   a. **[Selection]** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
   b. **[Crossover]** With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
   c. **[Mutation]** With a mutation probability mutate new offspring at each locus (position in chromosome).
   d. **[Accepting]** Place new offspring in the new population
4. **[Replace]** Use new generated population for a further run of the algorithm
5. **[Test]** If the end condition is satisfied, **stop**, and return the best solution in current population
6. **[Loop]** Go to step **2**

Chromosomes for Flexible Docking



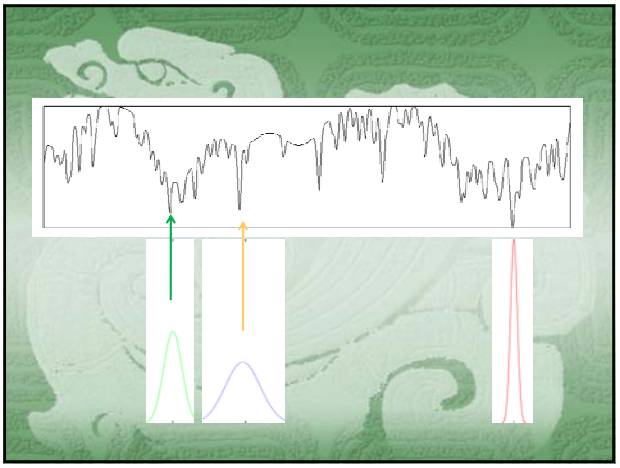
Leach, 2001.

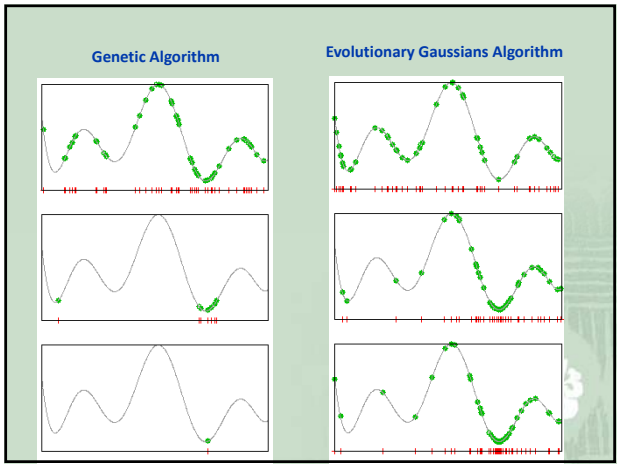
The Evolutionary Gaussians Algorithm

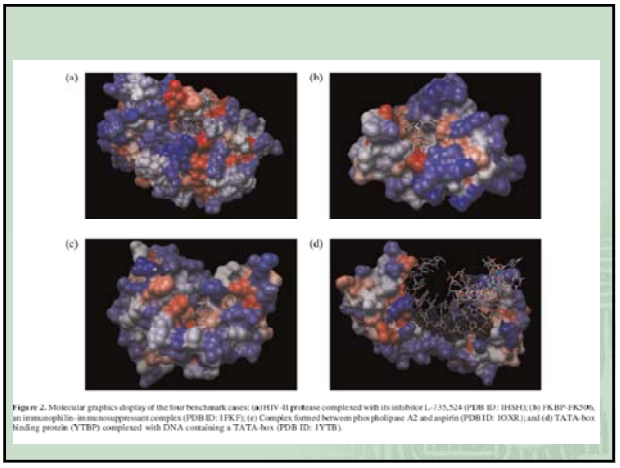
Nucleic Acids Research **33:** W233-W238 (2005)

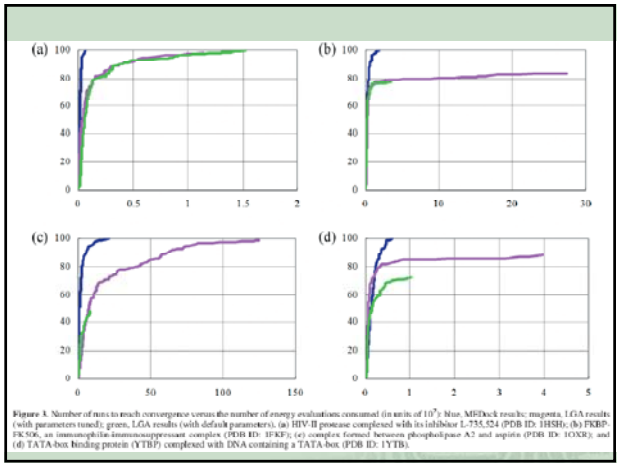
- *n* individuals, denoted by $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, are generated. Each $\mathbf{s}_i$ is a vector corresponding to a point in the domain of the objective function f. In order to achieve a scale-free representation, each component of $\mathbf{s}_i$ is linearly mapped to the numerical range of [0,1].
- The individuals in each generation of population are then sorted in the ascending order based on the values of the energy function on evaluated on these individuals. Let $\mathbf{t}_1, \mathbf{t}_2, \ldots \mathbf{t}_n$ denote the ordered individuals and we have $f(\mathbf{t}_1) < f(\mathbf{t}_2) < f(\mathbf{t}_n)$.
- *n* Gaussian distributions, denoted by $G_1, G_2, \ldots G_n$, are generated before the new generation of population is created. The center of each Gaussian distribution is selected randomly and independently from $\mathbf{t}_1, \mathbf{t}_2, \ldots \mathbf{t}_n$, where the probability is not uniform but instead follows a discrete diminishing distribution, $n : n\text{-}1 : \ldots : 1$.

$$G_i(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}\cdot\sigma_i}\right)\exp\left(-\frac{(\mathbf{x}-\mathbf{t}_k)^2}{2\sigma_i^2}\right) \qquad \sigma_i^2 = \alpha + \frac{(\beta-\alpha)(k-1)}{n-1}$$









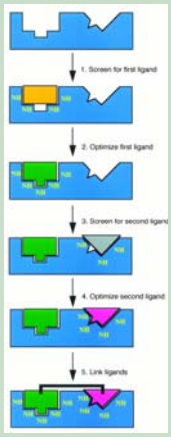
Other global optimization algorithms for molecular docking

- Simulated Annealing
- Biased Probability Monte Carlo (ICM)
- Differential Evolution (GEMDOCK, MolDock)
- Particle Swarm Optimization (Tribe-PSO, SODOCK)

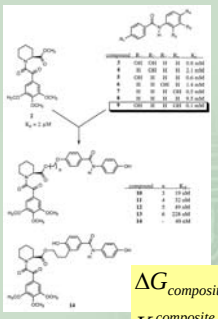
The Relaxed Complex Scheme

- Accommodating receptor flexibility by molecular dynamics
- Assigning compound molecular properties using rigorous quantum chemical approaches
- Rapid docking using Lamarckian Genetic Algorithm (or now the Evolutionary Gaussian Algorithm)
- Ranking compounds by *binding free energy spectra*, instead of single binding free energy.
- Multivalent drug design in a building-block fashion
- Computational analogue of "SAR by NMR"

Lin *et al. J. Am. Chem. Soc.*, **124**, 5632 (2002)
Lin *et al. Biopolymers*, **68**, 47 (2003)



Multivalent Drug Design in the "SAR by NMR" Method

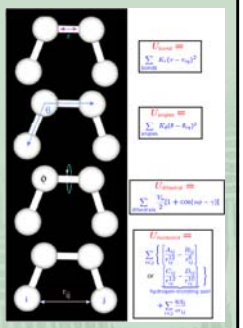


$$\Delta G_{composite} \approx \Delta G_1 + \Delta G_2$$
$$K_D^{composite} \approx K_D^1 \times K_D^2$$

Molecular dynamics simulations

$$m_i \dot{v}_i = -\nabla_i U + m_i \xi \left(\frac{T_0}{T} - 1 \right) v_i$$

$$v_i' = \lambda v_i$$
$$\lambda = \left[1 + \frac{\Delta t}{\tau_T} \left(\frac{T_0}{T} - 1 \right) \right]^{1/2}$$

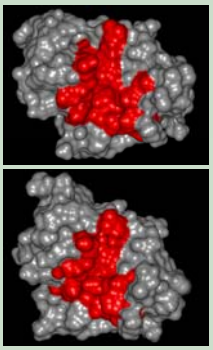


http://www.rcsb.org/pdb/

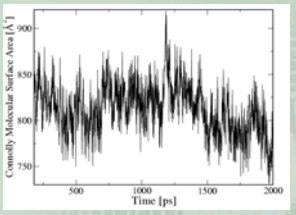
http://agave.wustl.edu/pdb2pqr/server.html

http://molprobity.biochem.duke.edu/

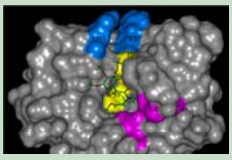
Dynamical Changes of the Solvent Accessible Surface Area of the FKBP Active Site



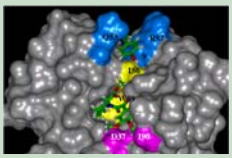
- Global breathing motion
- ~200Å^2 variation (comparable with small compounds)

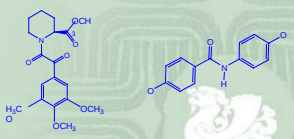


The Relaxed Complex Method can Reproduce the SAR-by-NMR Results

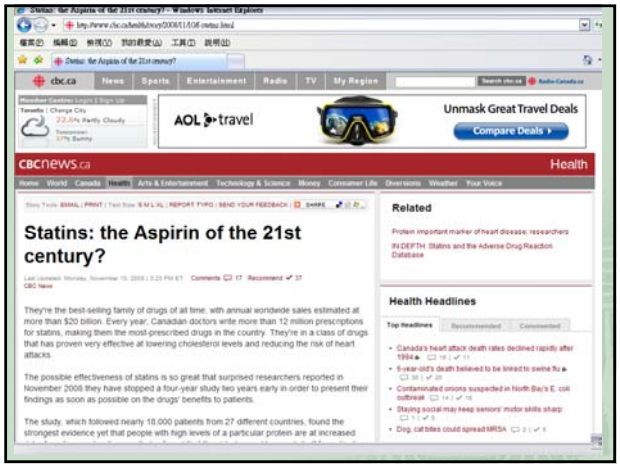


- Using a building-block approach to design a more specific and more potent drug
- A crucial test for the race between costly experimental approaches and computational methods.





2 9



STATINS AND CANCER PREVENTION

Marie-France Demierre†, Peter D. R. Higgins‡, Stephen B. Gruber§, Ernest Hawk‖ and Scott M. Lippman¶*

Abstract | Randomized controlled trials for preventing cardiovascular disease indicated that statins had provocative and unexpected benefits for reducing colorectal cancer and melanoma. These findings have led to the intensive study of statins in cancer prevention, including recent, large population-based studies showing statin-associated reductions in overall, colorectal and prostate cancer. Understanding the complex cellular effects (for example, on angiogenesis and inflammation) and the underlying molecular mechanisms of statins (for example, 3-hydroxy-3-methylglutaryl coenzyme-A (HMG-CoA) reductase-dependent processes that involve geranylgeranylation of Rho proteins, and HMG-CoA-independent processes that involve lymphocyte-function-associated antigen 1) will advance the development of molecularly targeted agents for preventing cancer. This understanding might also help the development of drugs for other ageing-related diseases with interrelated molecular pathways.

Demierre *et al.*, Nature Rev. Cancer **5:** 930-942 (2005)

The Risk of Cancer in Users of Statins

Matthijs R. Graaf, Annette B. Beiderbeck, Antoine C.G. Egberts, Dick J. Richel, and Henk-Jan Guchelaar

A B S T R A C T

Purpose
Several preclinical studies suggested a role for 3-hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors (statins) in the treatment of cancer. The objective of this study was to compare the risk of incident cancer between users of statins and users of other cardiovascular medication.

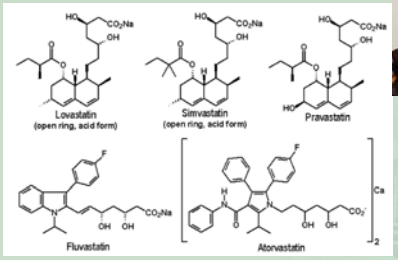
Methods
Data were used from the PHARMO database, containing drug dispensing records from community pharmacies and linked hospital discharge records for residents of eight Dutch cities. The study base included all patients with one or more prescriptions for cardiovascular drugs in the period between January 1, 1985 and December 31, 1998. Cases were identified as patients in the study base with a diagnosis of incident cancer and matched with four to six controls on sex, year of birth, geographic region, duration of follow-up, and index date. The analysis was adjusted for diabetes mellitus; prior hospitalizations; comorbidity; and use of diuretics, angiotensin-converting enzyme inhibitors, calcium-channel blockers, nonsteroidal anti-inflammatory drugs, sex hormones, and other lipid-lowering drug therapies.

Results
In the study base, 3,129 patients were identified and matched to 16,976 controls. Statin use was associated with a risk reduction of cancer of 20% (adjusted odds ratio [OR], 0.80; 95% CI, 0.66 to 0.96). Our data suggest that statins are protective when used longer than 4 years (adjusted OR, 0.64; 95% CI, 0.44 to 0.93) or when more than 1,350 defined daily doses are taken (adjusted OR, 0.60; 95% CI, 0.40 to 0.91).

Conclusion
This observational study suggests that statins may have a protective effect against cancer.

Demierre *et al.*, J. Clin. Onco. **22:** 2388-2394 (2004)

Can statins inhibit HDAC?

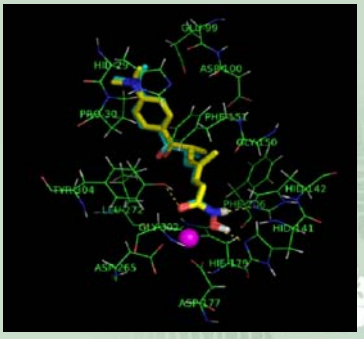


Professor Ching-Chow Chen, Department of Pharmacology, College of Medicine, National Taiwan University

Reproducing the X-ray crystallographic results

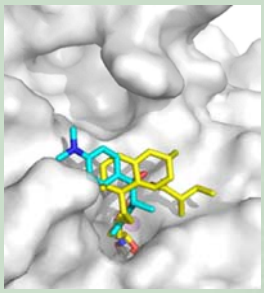
Structure: HDLP

CPK: Experimental pose
Yellow: Docked pose



Lin *et al. Cancer Res.* 68: 2375-2383 (2008)

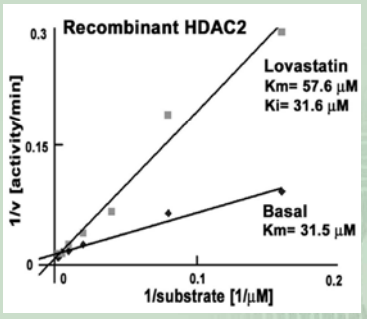
Lovastatin and TSA both can bind competitively at the catalytic site



Cyan: TSA
Yellow: Lovastatin

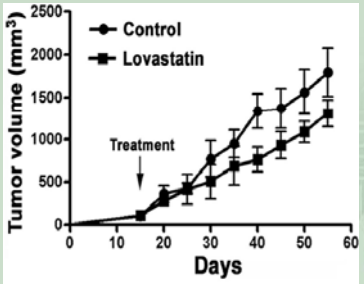
Lin *et al. Cancer Res.* **68:** 2375-2383

Lineweaver-Burk plot of enzymatic assay confirms nearly competitive inhibition of lovastatin



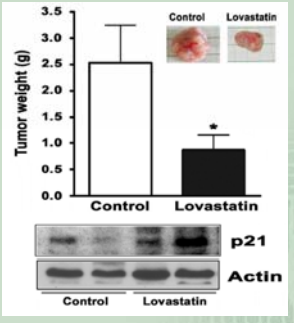
Lin *et al. Cancer Res.* **68:** 2375-2383

Statins reduced tumor growth rate in the xenograft nude mice



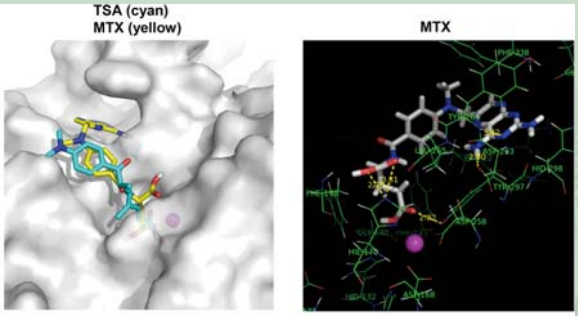
Lin *et al. Cancer Res.* **68:** 2375-2383

Lovastatin reduced the tumor size and weight after 45 days



Lin *et al. Cancer Res.* **68:** 2375-2383

The antifolate drug methotrexate is also an HDAC inhibitor



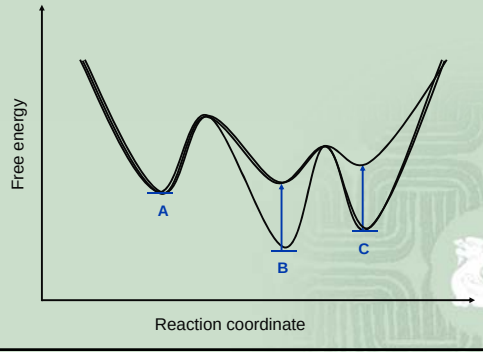
Biochem. Biophys. Res. Comm. **391** 1396-1399 (2010)

SLITHER: a web server for prediction of multiple binding site prediction and substrate translocation pathways

http://slither.rcas.sinica.edu.tw; http://bioinfo.mc.ntu.edu.tw/slither/

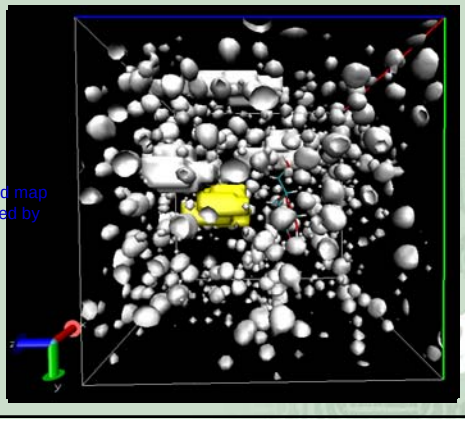
Nucleic Acids Research **37:** W559-W564 (2009)

Schematic illustration of SLITHER's methodology



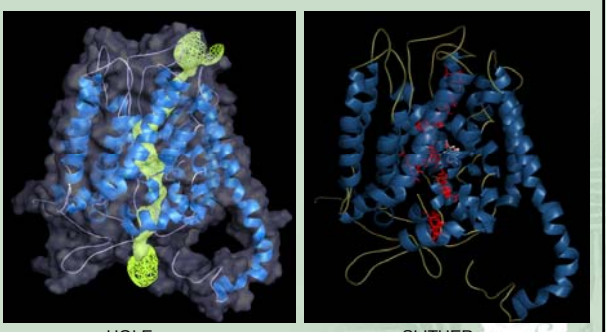
Prediction of the transport pathway using SLITHER

AutoGrid map visualized by VMD



Prediction of the transport pathway using SLITHER

Prediction of the glucose transport pathway of D-glucose in the GLUT1 using SLITHER



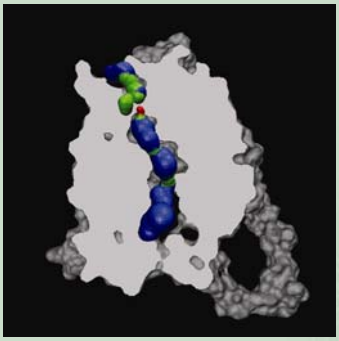
HOLE

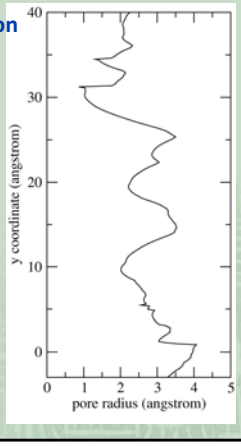
SLITHER

Biophys.J. **65**:2455-2460 (1993)

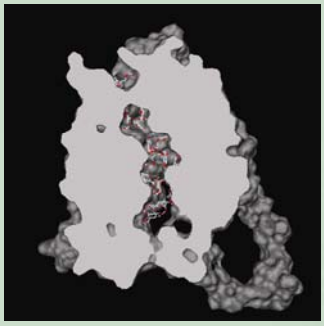
Nucleic Acids Research **37:** W559-W564 (2009)

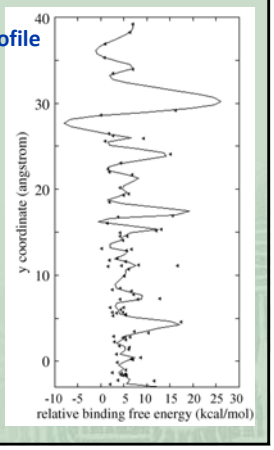
HOLE: Geometrical Characterization of Channel Structures





SLITHER: Binding Free Energy Profile of the Substrate in the Channel

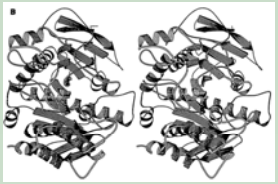




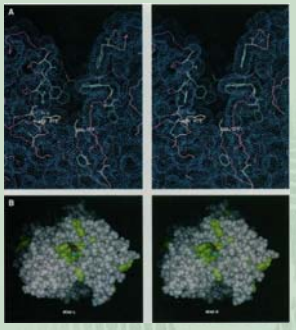
SLITHER: Extension of Docking

- The adjacent binding poses of two identical compounds of two different compounds can be used for fragment-based drug design.
- SLITHER performs more comprehensive global search, and therefore can be considered as extended version of docking.
- In principle, the docked free energies should follow the ascending order of the SLITHER iterations. This provides a basic assessment for the docking protocols and parameters.

Acetylcholinesterase has 20 Å- long gorge

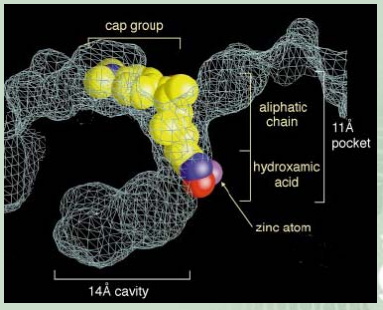


The active site gorge. The most remarkable feature of the structure is a deep and narrow gorge, ~20 Å long, that penetrates halfway into the enzyme and widens out close to its base (Fig. 7A). We have named this cavity the "active site gorge" because it contains the AChE catalytic triad. The O_γ atom of Ser^{200}, which can be seen from the surface of the enzyme (Fig. 7B), is ~4 Å above the base of the gorge. Fourteen aromatic residues line a substantial portion of the surface of the gorge (~40 percent) (Figs. 2, 4, 7, and 8). These residues and their flanking sequences, which are highly conserved in AChE's from different species (Fig. 4), come primarily from loops



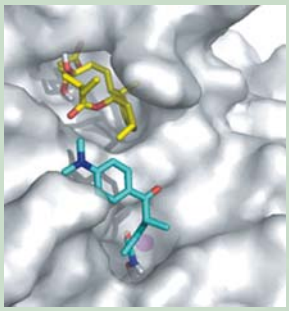
Science **253:** 872-879 (1991)

HDAC-like proteins have long internal cavity



Nature **401:** 188-193 (1999)

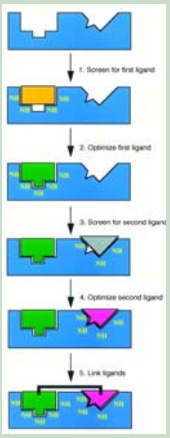
Another binding mode of lovastatin?

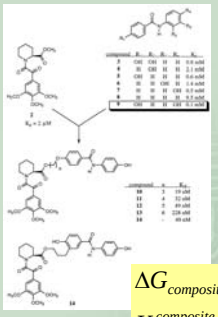


Cyan: TSA
Yellow: Lovastatin

Lin, *et al. Cancer Res.* **68:** 2375-2383 (2008)

Multivalent Drug Design in the "SAR by NMR"



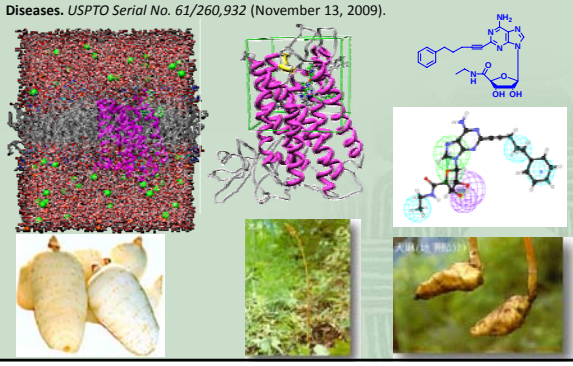


$$\Delta G_{composite} \approx \Delta G_1 + \Delta G_2$$

$$K_D^{composite} \approx K_D^1 \times K_D^2$$

Designing new therapeutic agents for treating Huntington's disease

Dual Functional Adenosine Analogues and use Therefore in Treating Neurodegenerative Diseases. *USPTO Serial No. 61/260,932* (November 13, 2009).



Keys for successfully applying the relaxed complex scheme

- The sampled conformations should cover the most relevant phase space.
- The scoring function should include important physical energetic factors.
- The ligand efficiency should be taken into account in the fragment-based drug design.

Lin, *Curr. Top. Med. Chem.* (2010)

Least square (LS) regression

Gauss, 1800

$$y_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip} + e_i$$

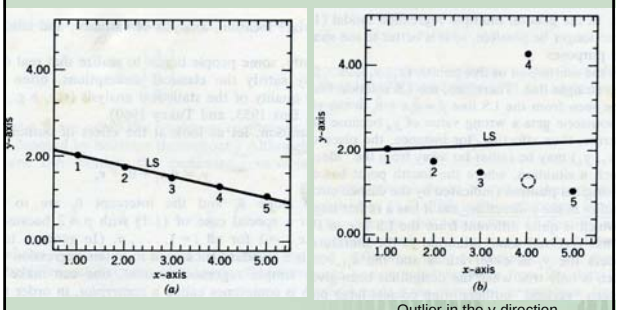
$$i = 1, \ldots, n$$

Sample size

$$r_i = y_i - \hat{y}_i \qquad \mathbf{w} = (w_1, w_2, \cdots, w_p)$$

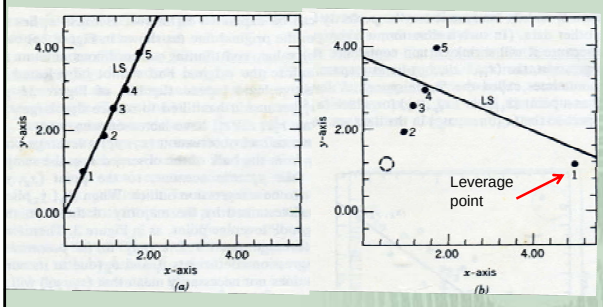
$$\underset{\mathbf{w}}{Minimize} \sum_{i=1}^{n} r_i^2$$

The problem of least square (LS) regression



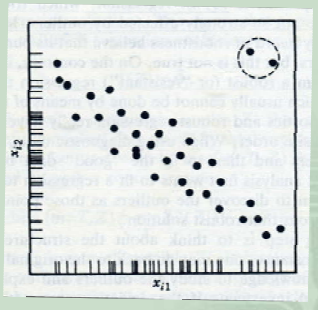
Outlier in the y-direction

The problem of least square (LS) regression



Outlier in the x-direction

Leverage points in two dimension



Leverage points are not easy to detect by checking the ranges of variables

Regression diagnostics versus robust regression

- Regression outliers pose a serious threat to standard least square analysis.
- Regression diagnostics: Use some quantity to pinpoint the influential points, remove the outliers, and then LS.
- Robust regression: Devise estimators not so strongly affected by outliers. Fit to the majority of data.

Robust regression

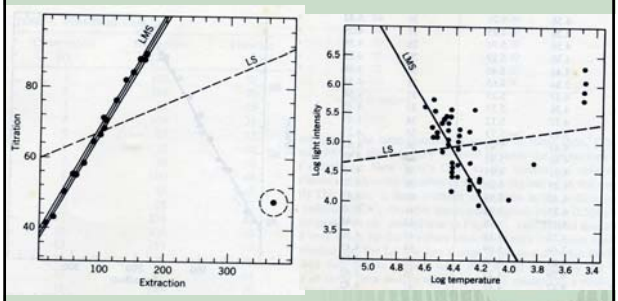


Table 1
X-Score Regression Coefficients Calculated with the 1555 Dataset

	C_0	W_{vdW}	W_{HB}	$W_{HP}/W_{HM}/W_{HS}$	W_{RT}	[a]RMSE	[b]$RMSE_{cv}$	r^2
HPSCORE	4.1018	0.0038	-0.0191	0.0087	-0.0628	2.198	2.199	0.3397
HMSCORE	4.1411	0.0040	-0.0175	0.2480	-0.0840	2.192	2.194	0.3433
HSSCORE	4.0627	0.0041	-0.0298	0.0032	-0.0812	2.207	2.208	0.3339
	C_0	W_{vdW}	W_{HB}	$W_{HP}/W_{HM}/W_{HS}$	W_{RT}	[a]RMSE	[b]$RMSE_{cv}$	r^2
HPSCORE	4.3313	0.0030	-0.0454	0.0086		2.217	2.218	0.3277
HMSCORE	4.4125	0.0032	-0.0653	0.1979		2.225	2.226	0.3231
HSSCORE	4.3363	0.0035	-0.0776	0.0024		2.239	2.238	0.3339

[a]All RMSE and $RMSE_{cv}$ values are in kcal/mol.
[b] $RMSE_{cv}$ is the root-mean-square error of 5-fold cross-validation.
All *F*-statistics of the models in Table 1: *p*-value < 2.2×10^{-16}

Table 2
X-Score Regression Coefficients Calculated with the 192 Dataset

	C_0	W_{vdW}	W_{HB}	$W_{HP}/W_{HM}/W_{HS}$	W_{RT}	[a]RMSE	[b]$RMSE_{cv}$	r^2
HPSCORE	2.8108	0.0070	0.0215	0.0039	-0.1249	2.433	2.467	0.3900
HMSCORE	2.8407	0.0067	0.0410	0.1835	-0.1445	2.418	2.454	0.3980
HSSCORE	2.7661	0.0073	0.0144	0.0015	-0.1384	2.436	2.481	0.3887
	C_0	W_{vdW}	W_{HB}	$W_{HP}/W_{HM}/W_{HS}$	W_{RT}	[a]RMSE	[b]$RMSE_{cv}$	r^2
HPSCORE	3.3048	0.0051	-0.0147	0.0049		2.500	2.572	0.3532
HMSCORE	3.3263	0.0054	-0.0301	0.1046		2.504	2.589	0.3507
HSSCORE	3.2330	0.0062	-0.0567	-0.00004		2.512	2.607	0.3466

[a]All RMSE and $RMSE_{cv}$ values are in kcal/mol.
[b] $RMSE_{cv}$ is the root-mean-square error of 5-fold cross-validation.
All *F*-statistics of the models in Table 2: *p*-value < 2.2×10^{-16}

Table 3
AutoDock4 Scoring Function Regression Coefficients Calculated with the 1555 Dataset

	W_{desolv}	W_{elec}	W_{vbond}	W_{vdw}	W_{vdw}		[a]RMSE	[b]$RMSE_{cv}$	
with N_{tor}	-0.0053	0.0170	-0.0051	0.2176	0.1525		2.917	2.919	
without N_{tor}	0.0136	0.0160	-0.0101		0.1310		3.053	3.055	
	W_{desolv}	W_{elec}	W_{vbond}	W_{vdw}	W_{vdw}	C_0	[a]RMSE	[b]$RMSE_{cv}$	r^2
with N_{tor}	0.0666	0.0099	0.0001	0.0654	0.0835	-5.8750	2.167	2.168	0.3624
without N_{tor}	0.0746	0.0093	-0.0011		0.0748	-6.1151	2.182	2.182	0.3534

[a]All RMSE and $RMSE_{cv}$ values are in kcal/mol.
[b] $RMSE_{cv}$ is the root-mean-square error of 5-fold cross-validation.
All *F*-statistics of the models in Table3: *p*-value < 2.2×10^{-16}

AutoDock4 Scoring Function Regression Coefficients Calculated with the 192 Dataset

	W_{desolv}	W_{elec}	W_{vbond}	W_{tors}	W_{vdw}		[a]RMSE	[b]RMSE$_{cv}$	
with N_{tor}	-0.0327	0.0041	-0.0089	0.2525	0.1546		2.704	2.671	
without N_{tor}	-0.0095	0.0177	-0.0237		0.1324		2.883	2.852	
	W_{desolv}	W_{elec}	W_{vbond}	W_{tors}	W_{vdw}	C_0	[a]RMSE	[b]RMSE$_{cv}$	r^2
with N_{tor}	0.0313	0.0023	0.0097	0.1498	0.1068	-4.0884	2.418	2.444	0.4044
without N_{tor}	0.0533	0.0092	0.0047		0.0878	-4.7069	2.479	2.505	0.3706

[a]All RMSE and RMSE$_{cv}$ values are in kcal/mol.
[b] RMSE$_{cv}$ is root-mean-square error of 5-fold cross-validation.
All *F*-statistics of the models in Table 4: *p*-value < 2.2×10^{-16}

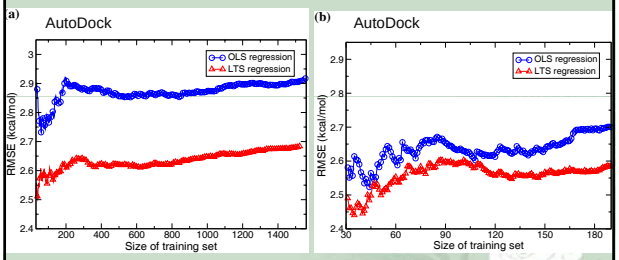
INTERNATIONAL JOURNAL OF SYSTEMS AND SYNTHETIC BIOLOGY
© International Science Press
1(2) December 2010, pp. 339-354

Robust Regression Analysis of Protein-ligand Binding Free Energy Models: Toward the Identification of Druggable Genomes

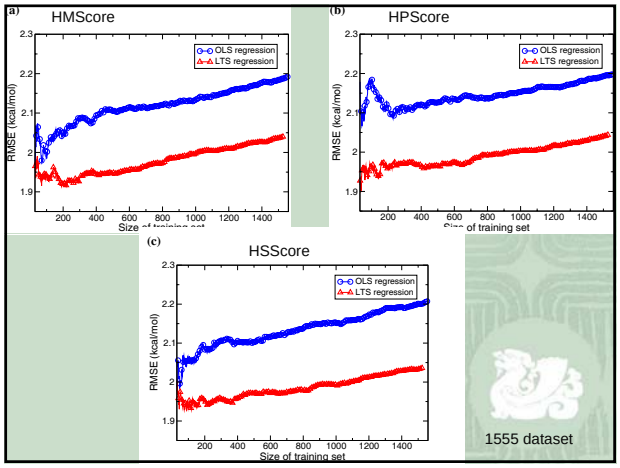
Jui-Chih Wang[1] and Jung-Hsin Lin[2]

[1]Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan;
[2]School of Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan
[2]Division of Mechanics, Research Center for Applied Science, Academia Sinica, Taipei, Taiwan;
[2]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan;

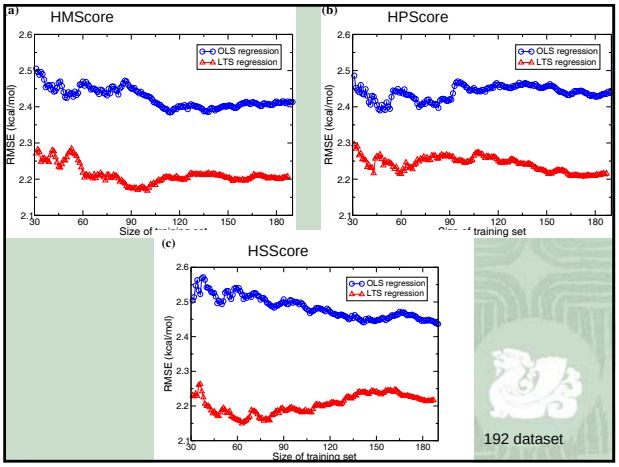
Evolutionary regression procedure



1555 dataset 192 dataset



1555 dataset



192 dataset

Conclusion

- Robust regression analysis indeed has beneficial effect on the construction of free energy models for protein-ligand interactions.
- With an evolutionary regression procedure, we showed that the root-mean-square of error will be substantially reduced when the outliers detected by the robust regression analysis were removed.
- The structures of protein-ligand complexes determined by X-ray crystallography or NMR need to be refined to obtain high quality energy terms for regression analysis.

Acknowledgments

Group members
(present and past) :

Gwan, Jean-Fang
Chu, Pei-Ying
Chen, Yo-Hsuan
Wang, Jui-Chih
Chang, Che-Chia
Chern, Ting-Rong
Chen, Jing-Yeh
Cheng, An-Liang
Lee, Po-Hsien
Liu, Min-Wei
An-Liang Cheng
Tsai, Cheng-Che
Shu-Hao Yeh
Kuei-Ling Kuo
Yo-Hsaun Tu

Funding and support :
National Science Council,
Research Center for Applied Sciences, Academia Sinica
National Center for High Performance Computing