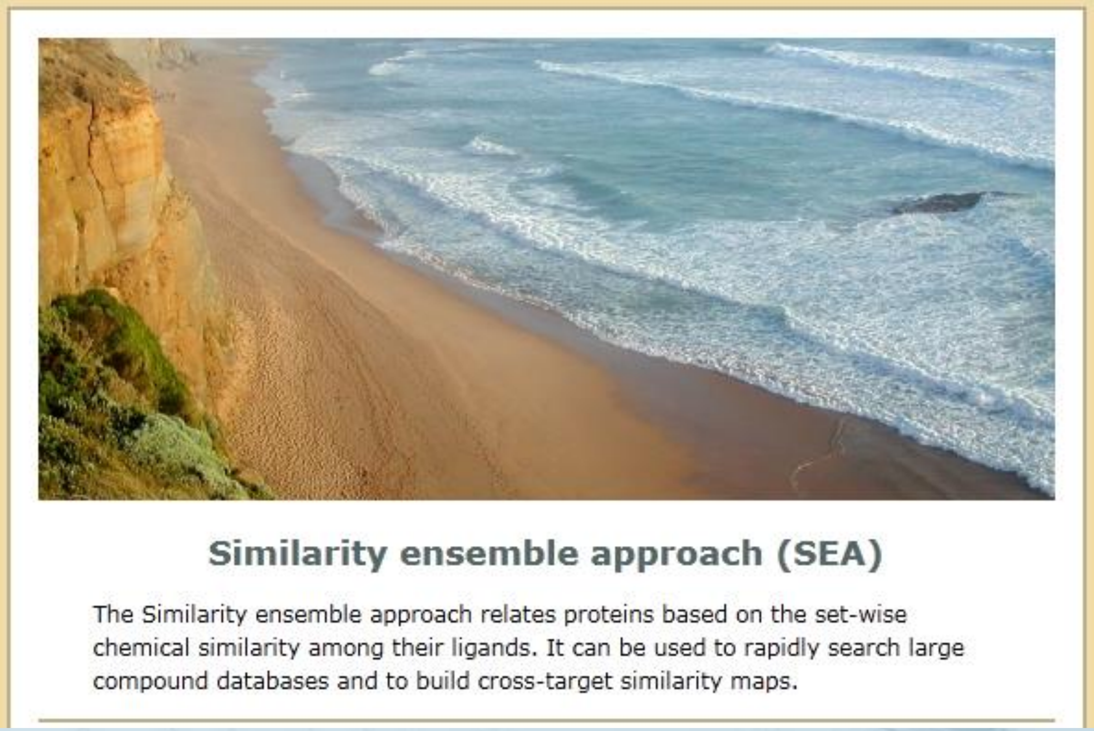
Journal Club

Reporter: Zhixiong Zhao
2012.10.8

Nature **2012**,361-368

Large-scale prediction and testing of drug activity on side-effect targets

Brian K. Shoichet ‘s lab, Novartis Institute



Similarity ensemble approach (SEA)

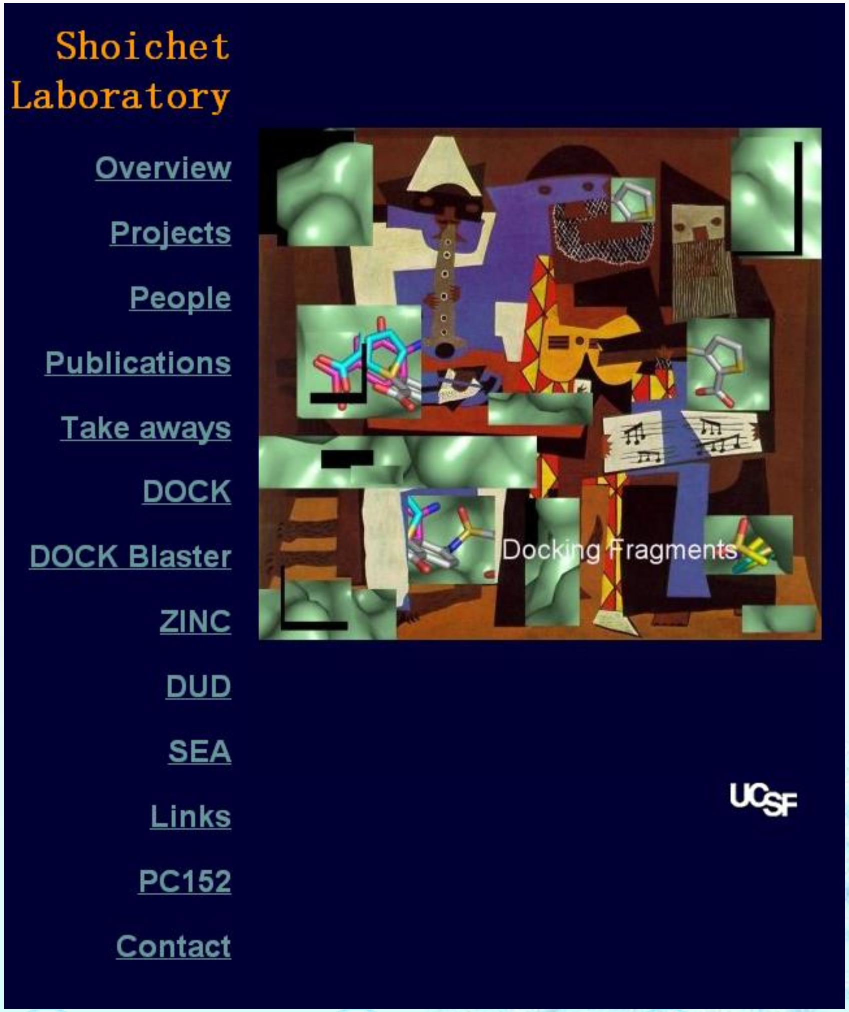
The Similarity ensemble approach relates proteins based on the set-wise chemical similarity among their ligands. It can be used to rapidly search large compound databases and to build cross-target similarity maps.

Author Introduction

The corresponding author Brian K. Shoichet is a professor in UCSF. He received his Ph.D for work with Tack Kuntz in UCSF. His team is one of the team to maintain and develop the famous DOCK program. And the ZINC database is also developed in Shoichet's lab.

Shoichet's lab focus on the research of docking, protein-ligand interaction. "Two crazy goals for the next five years are predicting one ligand for every accessible protein target, and for every drug predicting one new target to which it binds. "

Here, I will introduce their work on SEA.



Shoichet Laboratory

Overview

Projects

People

Publications

Take aways

DOCK

DOCK Blaster

ZINC

DUD

SEA

Links

PC152

Contact

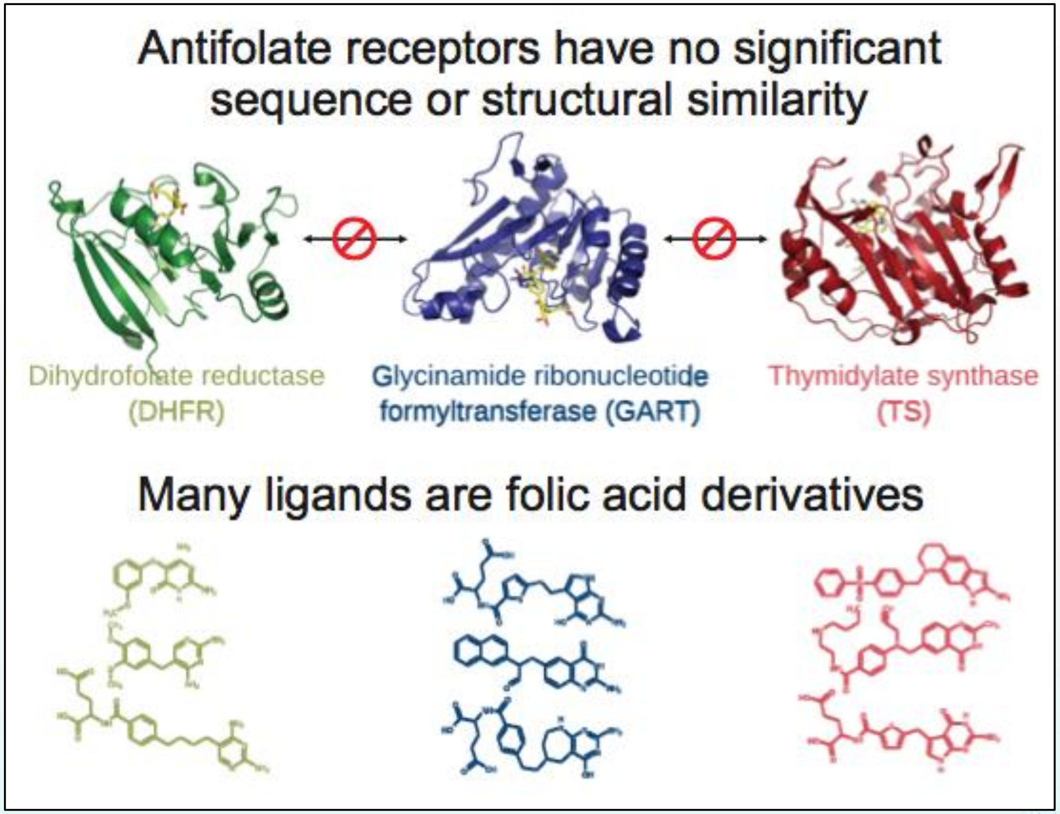
Three main articles on this topic:

- First is focus on the basic method called **SEA**, which use the chemical similarity for grouping the proteins from their ligands. (*Nature Biotechnology **2007**,197*)

- Second is applying SEA for off-target prediction. (*Nature **2009**,175*)

- This paper focus on the application of SEA for large-scale testing the activities of off-targets and relating the adverse drug reaction(ADR) to the off-target. (*Nature **2012**,361*)

One Drug, Multi Targets



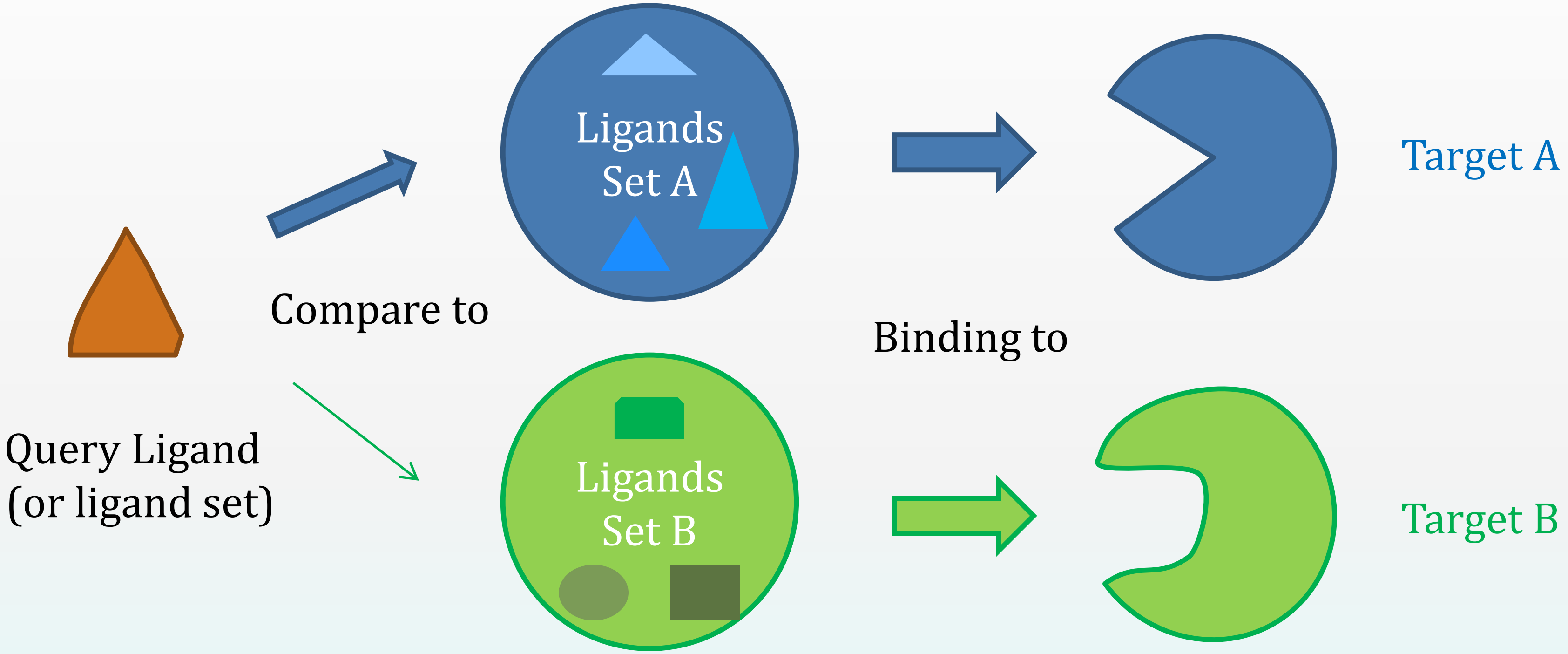
Antifolate receptors have no significant sequence or structural similarity

Dihydrofolate reductase (DHFR)
Glycinamide ribonucleotide formyltransferase (GART)
Thymidylate synthase (TS)

Many ligands are folic acid derivatives

Adverse drug reactions(ADR), often caused by off-target interaction, limits the use of effective drug and may cause disastrous events. Prediction of unknown off-target drug interaction might prevent the disastrous and allow safer molecules to be prioritized for pre-clinical development. However, sometimes the off-target protein have no significant sequence or structural similarity to the expected target. Chemoinformatics methods on the ligands may offer more opportunity.

SEA (Similarity Ensemble Approach)



Query Ligand
(or ligand set)

Compare to

Ligands
Set A

Ligands
Set B

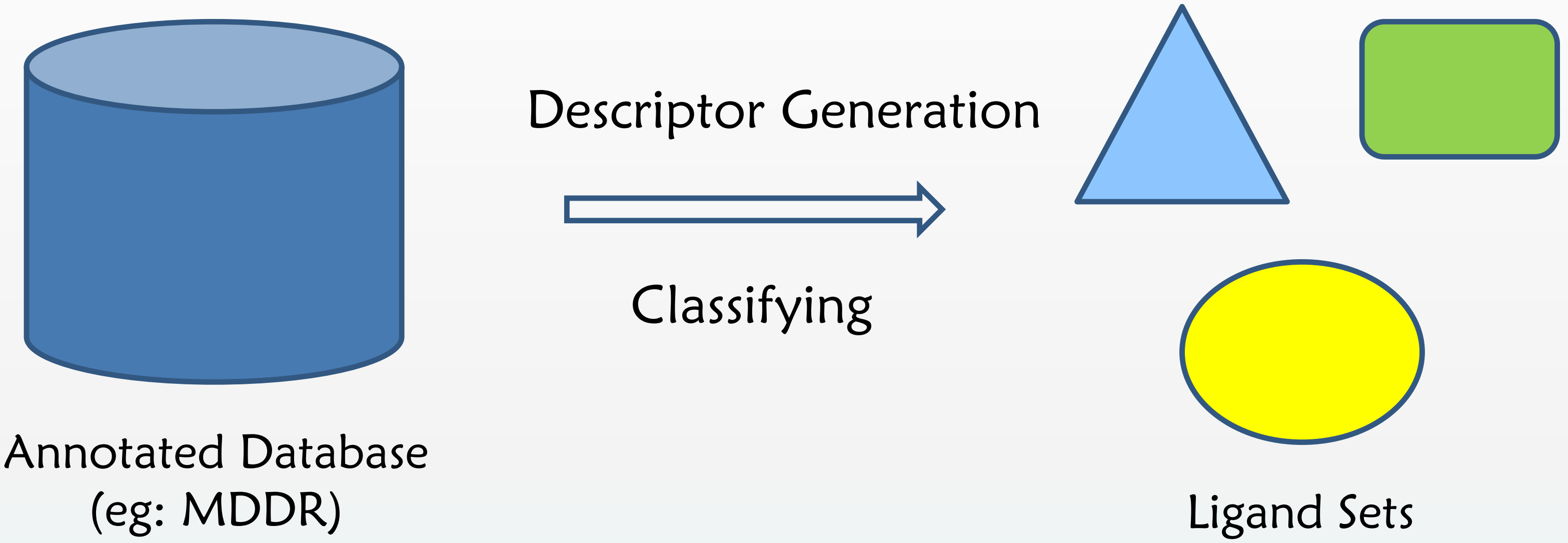
Binding to

Target A

Target B

The key idea of SEA is using the known binding ligands as a set for representing the target protein, then the query ligand or ligand set could be compared to the target ligand set by chemical similarity to identify the potential target. Its unique feature is utilizing a statistical method similar to the BLAST.

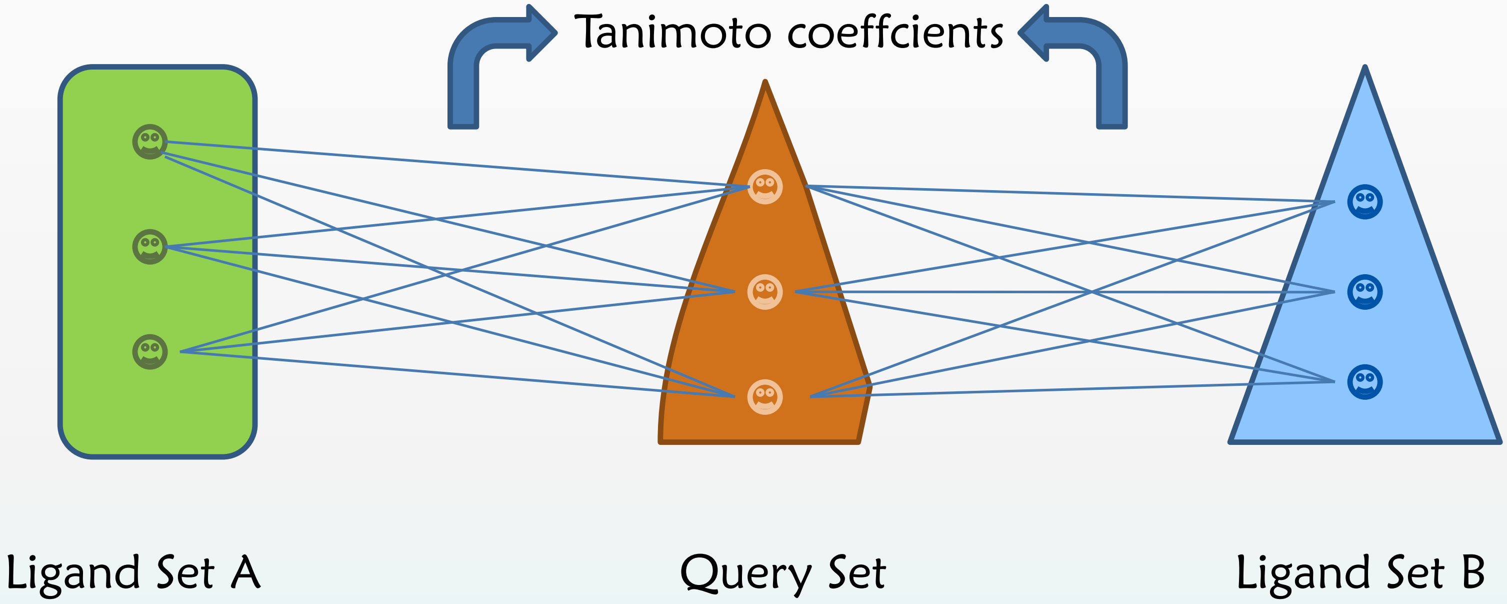
Preparation of ligand set



65,241 ligands in MDL Drug Data Report (MDDR) were used. Their annotation on the modulated targets were extracted and used for the classification. Only sets containing five or more ligands were used. Finally, they were classified to 246-receptor subsets.

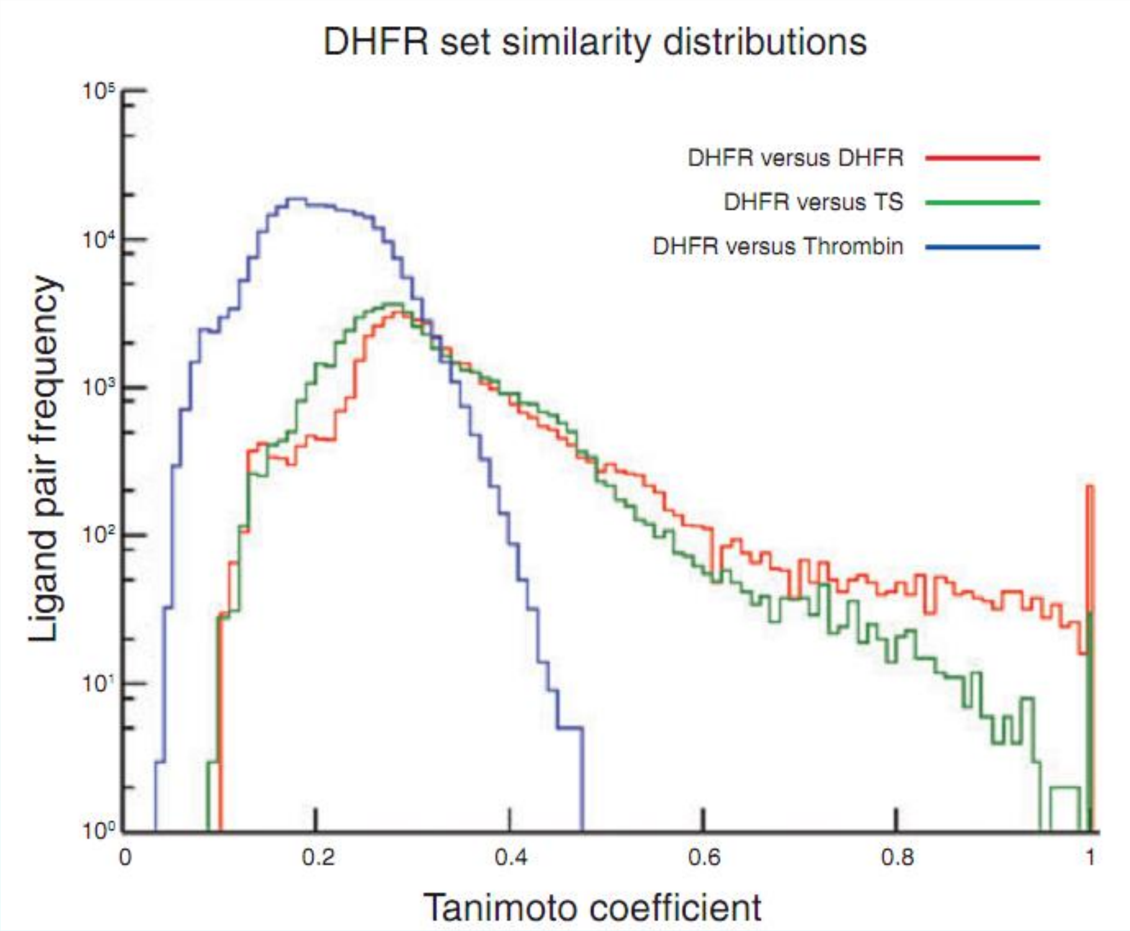
*The classification depends on the kind of database and the chose annotation.

Set Comparison



Then each ligand in the query set was compared to the each ligand in all target ligand sets, and the Tanimoto coeffcients (Tc) for chemical similarity were calculated via 2-dimensional fingerprints by Daylight or ECFP_4. (Tc=0~1)

Similarity distribution



DHFR: antifolate enzyme,
216 ligands
TS: thymidylate synthase,
related protein, 253 ligands
Thrombin: unrelated protein,
1226 ligands

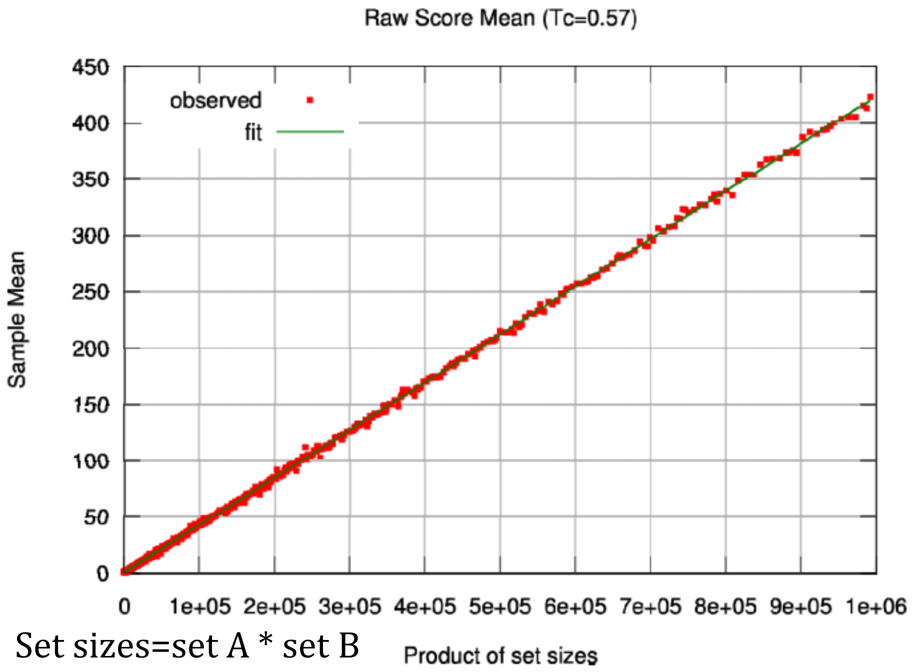
When comparing the 216 ligands of the antifolate enzyme DHFR to themselves, 80.4% of the pairs had a Tc in the 0.1 to 0.4 range, with only 5.2% having more substantial scores in the 0.6–1.0 range

When the Tc is low, the ligands compared were considered insubstantial similarity. Thus, the ligand pair with Tc>=0.57 were used and raw similarity score between sets could be sum up the Tc of the similar ligand pairs. Most pairs of ligand sets have no raw similarity score(raw score=0, 70.8%).

Necessity for “random” similarity

- Though the raw similarity score could help to identify the relationship between query set and target, the size and chemical composition bias could not be ignored. Thus, it’s improper to use the raw score directly to evaluate the similarity between different sets.
- To compare the significance of the set similarity raw scores across sets of different sizes, they developed a statistical model of the similarity they would expect at random for sets drawn from the same large but finite database of ligands.

Fit for expect raw score



Set sizes=set A * set B

Raw score mean *vs* Product of set size
From random set from the same DB
(Background sets)

$$y_{\mu} = mx^{n} + p$$

⇓

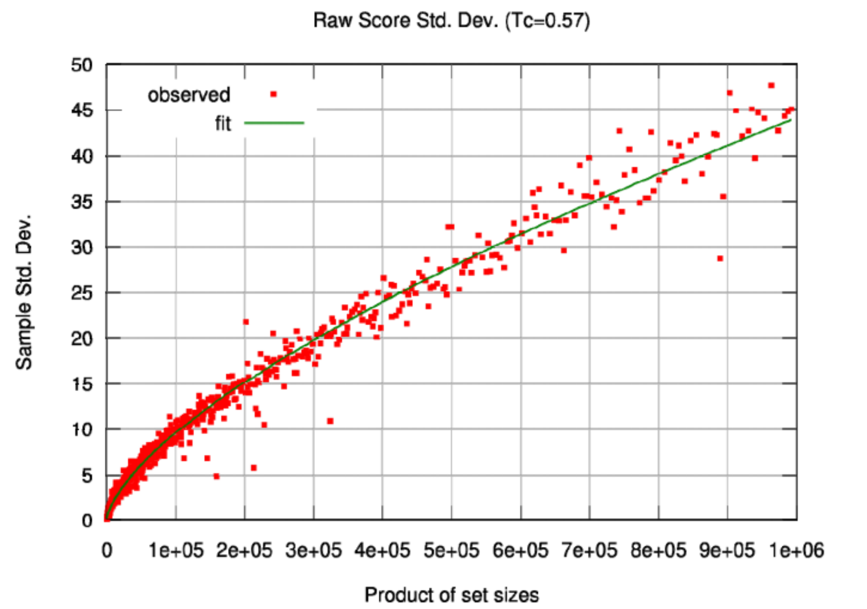
$$\mu(x) \approx (4.24\times10^{-4})x$$

μ(x):expected raw score mean

•Different size ligand sets was extracted from the same database randomly. The similarities via Tc were calculated and different threshold Tc (0~1,0.01 step)was used to calculate the raw score. Nonlinear fitter was used for the expected raw score mean and standard deviation via least-square.

•It's easy to understand the linear relation between the raw score and the product of set size.

Fit for standard deviation



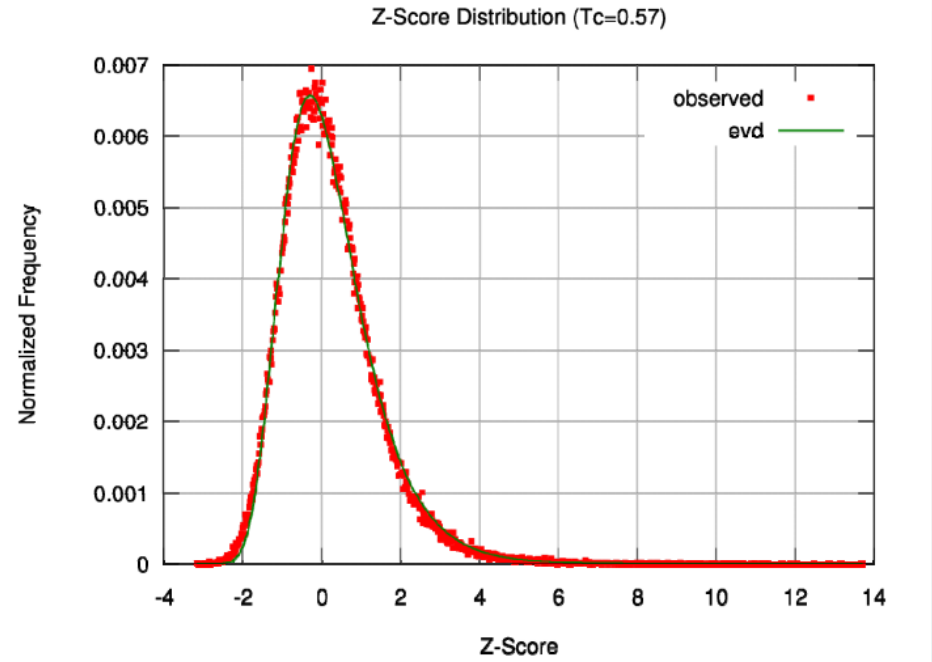
S.D. *vs* Product of set size

$$y_\sigma = qx^r + s$$

$$\sigma(x) \approx (4.49\times10^{-3})\,x^{0.665}$$

Bin the data by the x-axis values(>=5 points) and then calculate the standard deviation of each bin with Laplacian correction. Finially, fit the S.D. vs product of set size.

Z-score and E-value



$$z = (rs(S_1,S_2) - \mu(n(S_1,S_2))) / \sigma(n(S_1,S_2))$$

$rs(S_1,S_2)$ =raw score of set S1 vs S2

$$P(Z > z) = 1 - \exp(-e^{-z\pi/\mathrm{sqrt}(6) - \Gamma'(1)})$$

Then the set comparison Z-scores were calculated as a function of the set raw scores, expected raw scores and s.d. The histogram of Z-scores (right) of the random sets conformed to an extreme value distribution, which underlied BLAST comparisons of protein and DNA sequences. Finally, the probability of the score being achieved by random chance alone, given the Z-score, was converted to an expectation value (E-value).
The threshold Tc was obtained from the best chi-square to fit the extreme value distribution, here threshold Tc=0.57

Method testing

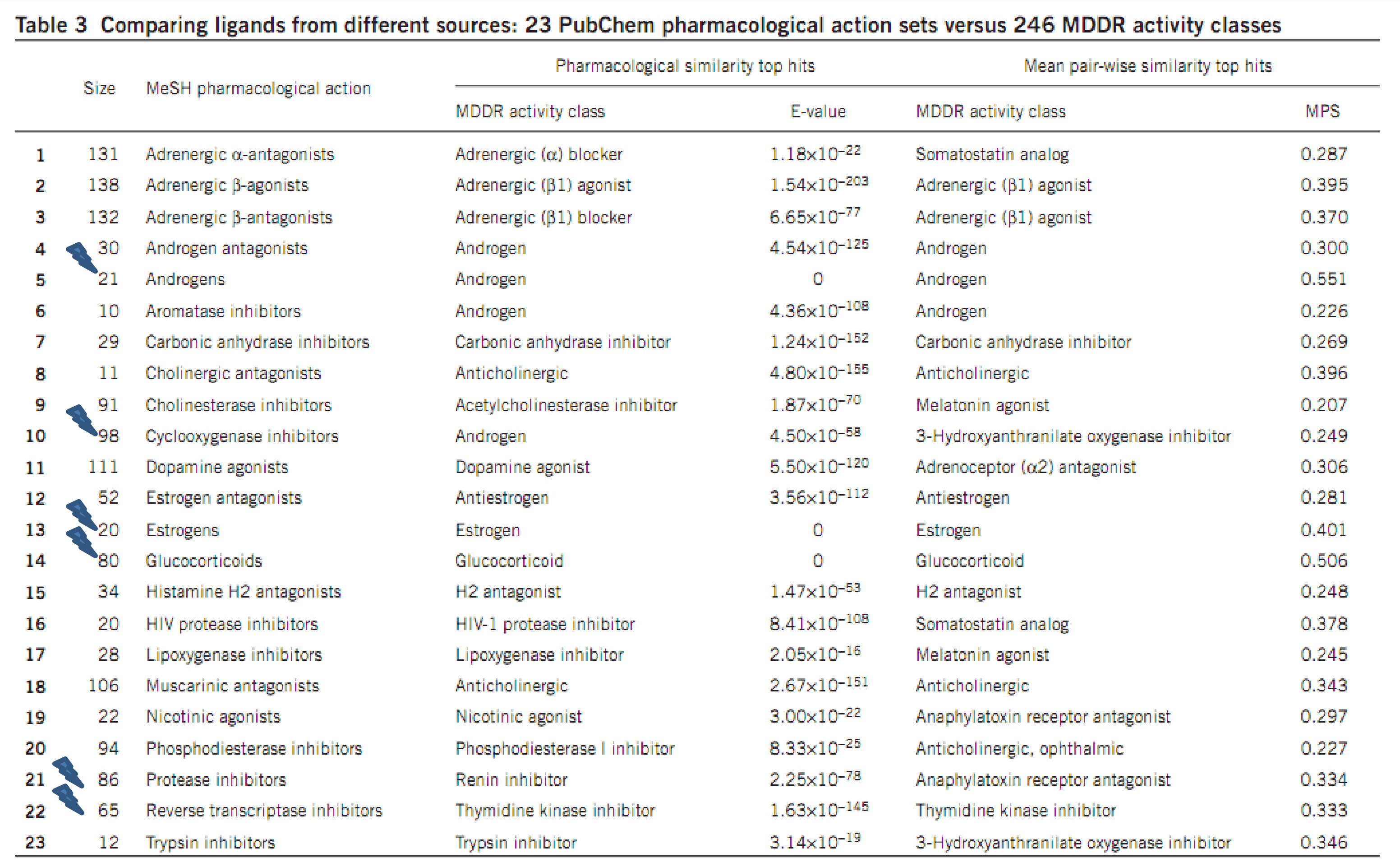


Table 3 Comparing ligands from different sources: 23 PubChem pharmacological action sets versus 246 MDDR activity classes

	Size	MeSH pharmacological action	Pharmacological similarity top hits		Mean pair-wise similarity top hits	
			MDDR activity class	E-value	MDDR activity class	MPS
1	131	Adrenergic α-antagonists	Adrenergic (α) blocker	1.18×10^{-22}	Somatostatin analog	0.287
2	138	Adrenergic β-agonists	Adrenergic (β1) agonist	1.54×10^{-203}	Adrenergic (β1) agonist	0.395
3	132	Adrenergic β-antagonists	Adrenergic (β1) blocker	6.65×10^{-77}	Adrenergic (β1) agonist	0.370
4	30	Androgen antagonists	Androgen	4.54×10^{-125}	Androgen	0.300
5	21	Androgens	Androgen	0	Androgen	0.551
6	10	Aromatase inhibitors	Androgen	4.36×10^{-108}	Androgen	0.226
7	29	Carbonic anhydrase inhibitors	Carbonic anhydrase inhibitor	1.24×10^{-152}	Carbonic anhydrase inhibitor	0.269
8	11	Cholinergic antagonists	Anticholinergic	4.80×10^{-155}	Anticholinergic	0.396
9	91	Cholinesterase inhibitors	Acetylcholinesterase inhibitor	1.87×10^{-70}	Melatonin agonist	0.207
10	98	Cyclooxygenase inhibitors	Androgen	4.50×10^{-58}	3-Hydroxyanthranilate oxygenase inhibitor	0.249
11	111	Dopamine agonists	Dopamine agonist	5.50×10^{-120}	Adrenoceptor (α2) antagonist	0.306
12	52	Estrogen antagonists	Antiestrogen	3.56×10^{-112}	Antiestrogen	0.281
13	20	Estrogens	Estrogen	0	Estrogen	0.401
14	80	Glucocorticoids	Glucocorticoid	0	Glucocorticoid	0.506
15	34	Histamine H2 antagonists	H2 antagonist	1.47×10^{-53}	H2 antagonist	0.248
16	20	HIV protease inhibitors	HIV-1 protease inhibitor	8.41×10^{-108}	Somatostatin analog	0.378
17	28	Lipoxygenase inhibitors	Lipoxygenase inhibitor	2.05×10^{-16}	Melatonin agonist	0.245
18	106	Muscarinic antagonists	Anticholinergic	2.67×10^{-151}	Anticholinergic	0.343
19	22	Nicotinic agonists	Nicotinic agonist	3.00×10^{-22}	Anaphylatoxin receptor antagonist	0.297
20	94	Phosphodiesterase inhibitors	Phosphodiesterase I inhibitor	8.33×10^{-25}	Anticholinergic, ophthalmic	0.227
21	86	Protease inhibitors	Renin inhibitor	2.25×10^{-78}	Anaphylatoxin receptor antagonist	0.334
22	65	Reverse transcriptase inhibitors	Thymidine kinase inhibitor	1.63×10^{-145}	Thymidine kinase inhibitor	0.333
23	12	Trypsin inhibitors	Trypsin inhibitor	3.14×10^{-19}	3-Hydroxyanthranilate oxygenase inhibitor	0.346

23 ligand sets from PubChem (1421 compounds, not in MDDR) were tested by SEA, and 17 found a matching activity class as the top-ranked hit (average rank 1.4). But in mean pair-wise similarity (MPS) method, only 9 found a matching top-ranked hit (average rank 8.2). This result show that SEA was effective for ligand classification.

Off-target prediction

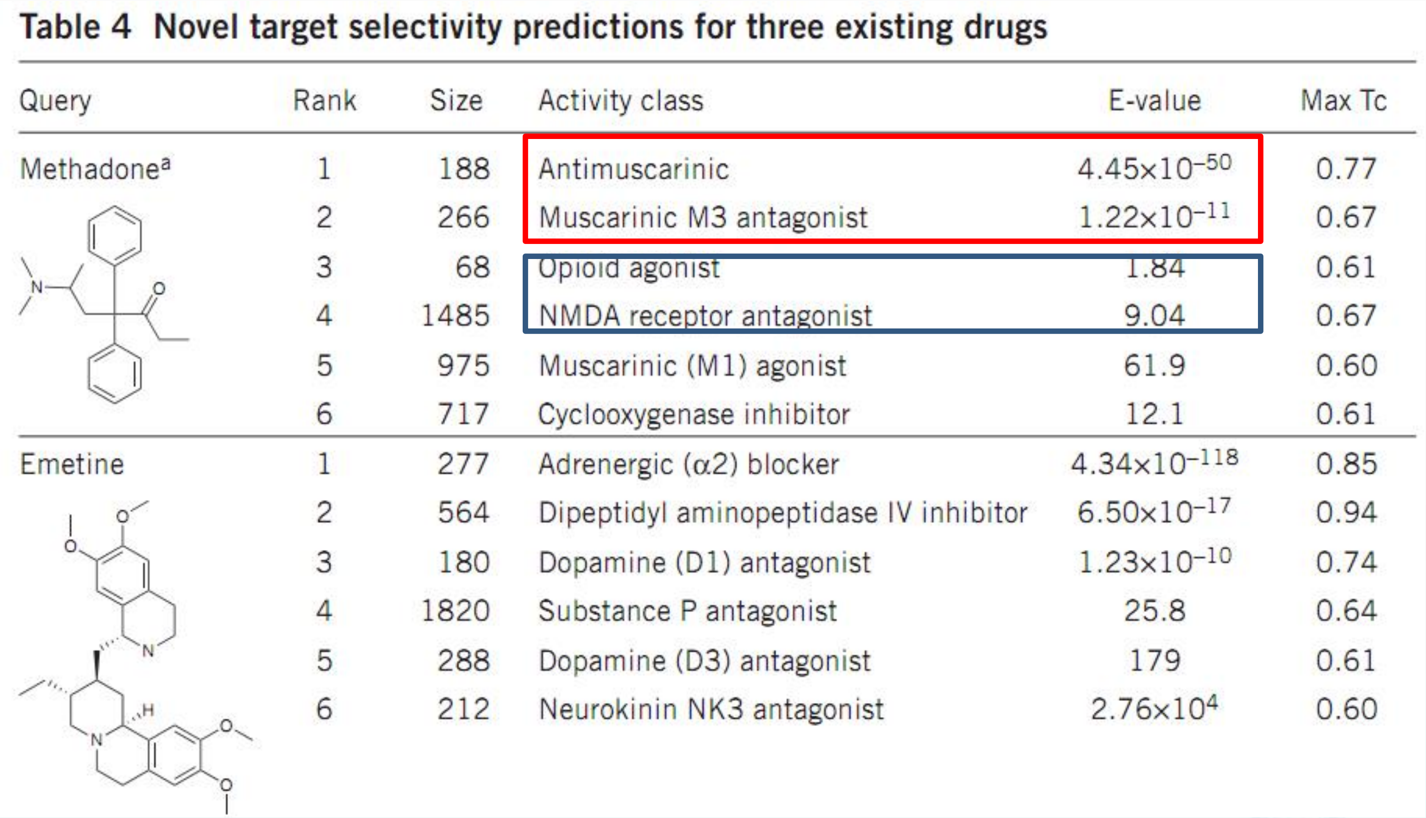


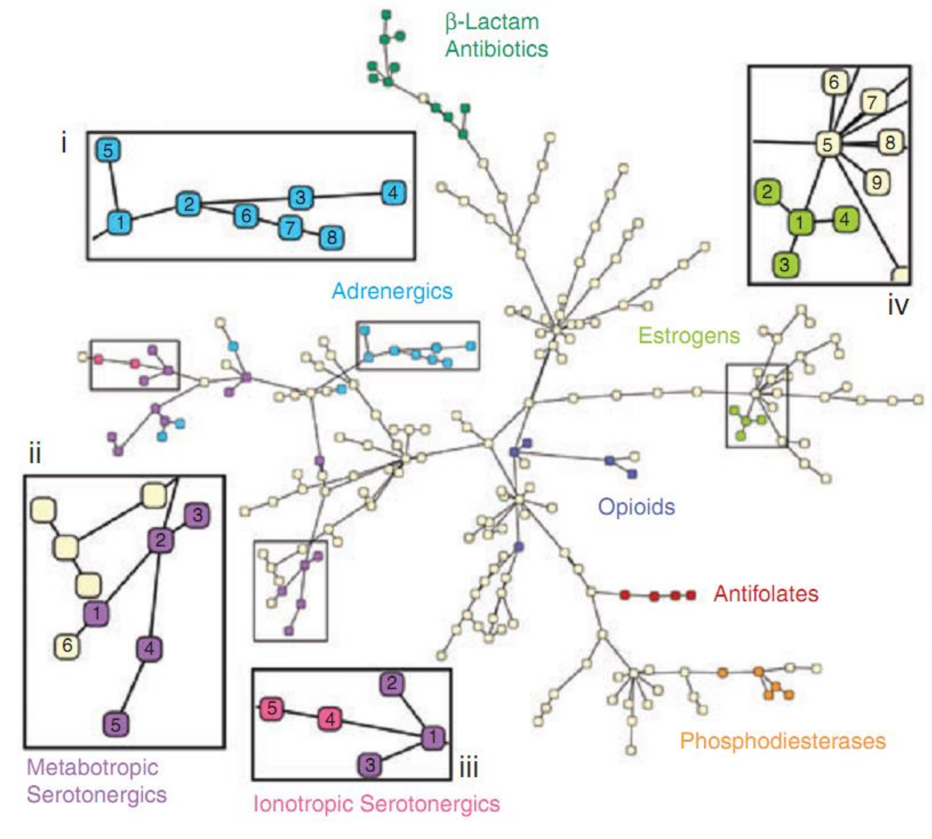
Table 4 Novel target selectivity predictions for three existing drugs

Query	Rank	Size	Activity class	E-value	Max Tc
Methadone[a]	1	188	Antimuscarinic	4.45×10^{-50}	0.77
	2	266	Muscarinic M3 antagonist	1.22×10^{-11}	0.67
	3	68	Opioid agonist	1.84	0.61
	4	1485	NMDA receptor antagonist	9.04	0.67
	5	975	Muscarinic (M1) agonist	61.9	0.60
	6	717	Cyclooxygenase inhibitor	12.1	0.61
Emetine	1	277	Adrenergic (α2) blocker	4.34×10^{-118}	0.85
	2	564	Dipeptidyl aminopeptidase IV inhibitor	6.50×10^{-17}	0.94
	3	180	Dopamine (D1) antagonist	1.23×10^{-10}	0.74
	4	1820	Substance P antagonist	25.8	0.64
	5	288	Dopamine (D3) antagonist	179	0.61
	6	212	Neurokinin NK3 antagonist	2.76×10^{4}	0.60

Methadone with its analogs as a set, targeting NMDA and μ-opioid, has been predicted to muscarinic M3 antagonist by SEA. Further test show Ki=1.0 uM. Similar result is for emetine.

PS: More information could refer to another paper focusing on off-target prediction.(Nature **2009**,175)

Finding relationship between proteins



Similarity maps for 246 enzymes and receptors based on the E-values

After comparison of 246 ligand sets with each other, the E-value between them could be obtained and could be used to plot the relationship tree between the 246 targets via minimal spanning tree.

Clusters of biologically related targets may be observed as an emergent property, as no explicit biological information, only ligand information, is used to calculate the cross-target similarity.

New large-scale prediction

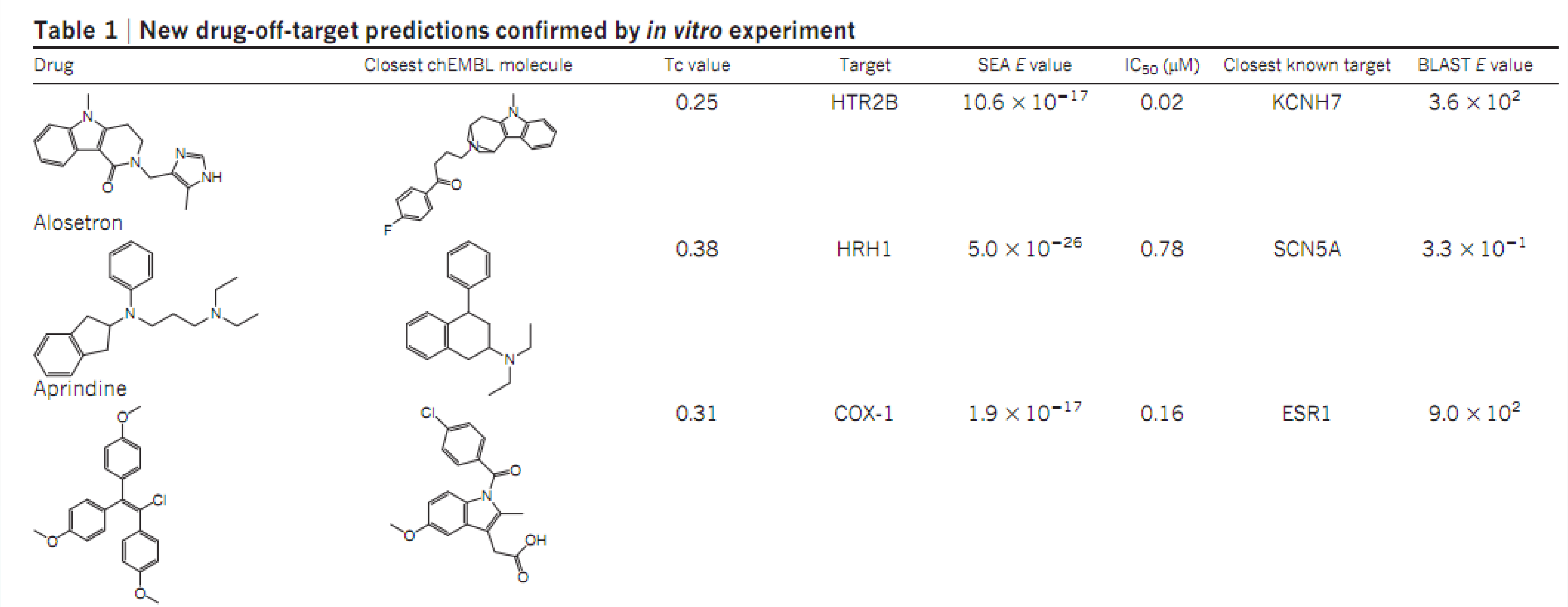


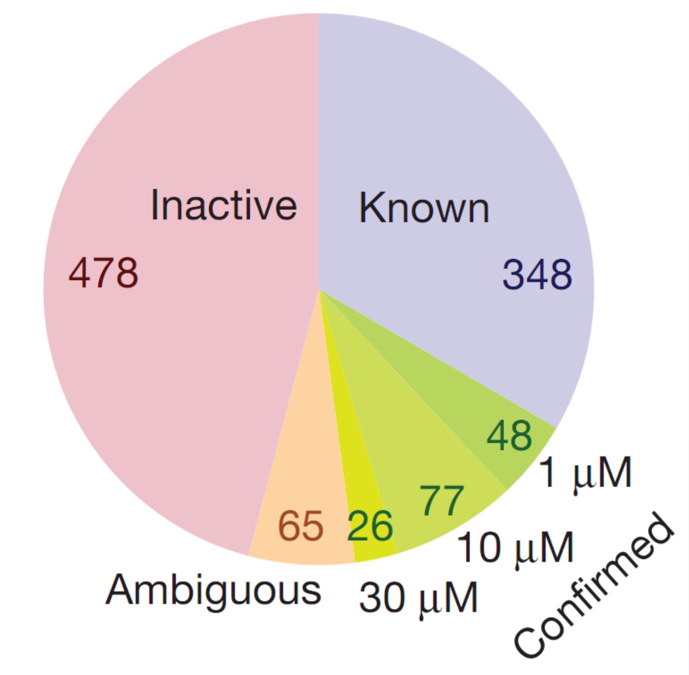
Table 1 | New drug-off-target predictions confirmed by *in vitro* experiment

Drug	Closest chEMBL molecule	Tc value	Target	SEA *E* value	IC_{50} (μM)	Closest known target	BLAST *E* value
Alosetron		0.25	HTR2B	10.6×10^{-17}	0.02	KCNH7	3.6×10^{2}
Aprindine		0.38	HRH1	5.0×10^{-26}	0.78	SCN5A	3.3×10^{-1}
		0.31	COX-1	1.9×10^{-17}	0.16	ESR1	9.0×10^{2}

656 marketed drugs were screened by SEA for their likelihood to bind to 73 targets potentially related to ADR. The ligand sets for these targets were from ChEMBL. Totally, 151 new drug-target association were confirmed by in vitro test(IC50<=30uM).

Nature **2012**,361-368

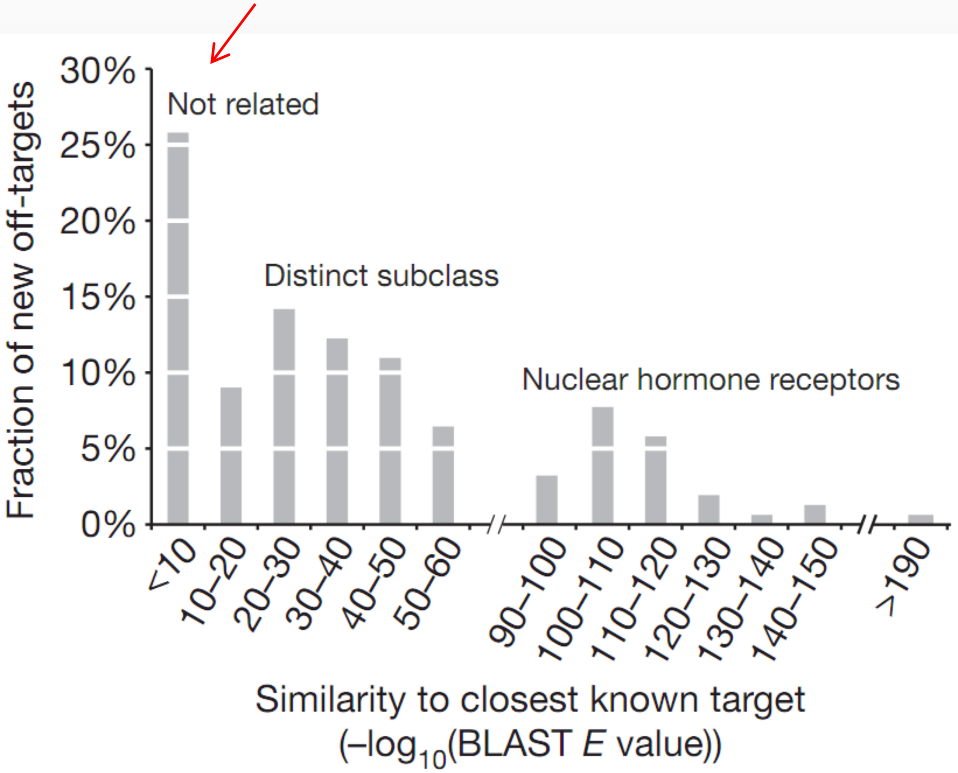
Validation for prediction result



Only 1,644 of the more than 47,000 possible drug–target pairs had significant E values. Of these, 403 were already known in ChEMBL and so were trivially confirmed. In the remaining 1,241 predictions, 348 (28%) were unknown to ChEMBL, but could be found in proprietary ligand–target databases that were unavailable to SEA. The others were tested for their bioacitivities. 151 have IC50 <=30uM.

In summary, of the 1,042 predictions that were tested (694 by assay, 348 by databases), 48% were confirmed either in proprietary databases or in Novartis assays in full concentration responses, and just under 46% were disproved.

Compare to sequence similarity

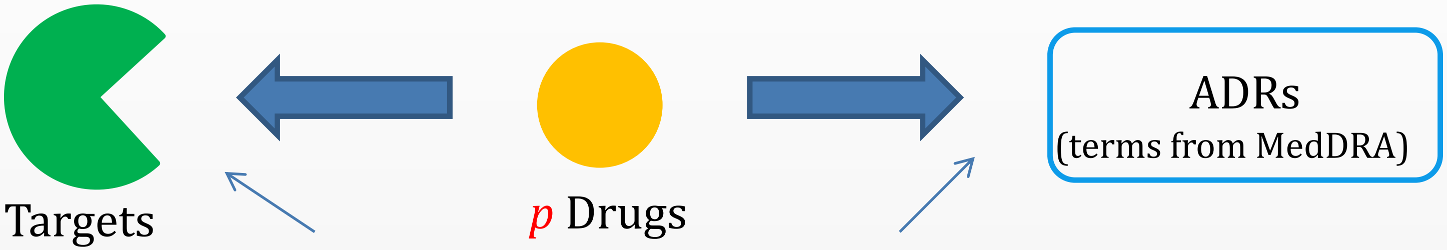


- The BLAST sequence similarity of predicted targets to any known target of a drug was calculated to identify the similar known targets.

- It could be found that 39 in 151(26%) have BLAST E-value>10e-5, suggesting the previously known targets shared no sequence similarity with the new off-targets.

This result indicated that many off-targets are hard to be predicted by the methods based on sequence similarity.

Association of target with ADR



ADRs
(terms from MedDRA)

Targets

p Drugs

All ADR-drug related (total T)

punish

All target-drug related (total A)

To evaluate the relation between targets and ADRs, the author found all the Targets-drugs and Drugs-ADRs relations and enumerated(枚举) all the possible Targets-ADRs relations. Then, an enrichment score (EF) based on guilt-by-association metric(“连坐”) was calculated as follow:

$$\mathrm{EF} = p/(A \times T/P)$$

p is the co-occurrence of target X and ADR Y on a drug, A is the number of times ADR Y was linked to any drug–target pair, T is the number of times target X was linked with any drug–ADR pair, and P is the total number of target–ADR pairs (681797 total).

The normal p-value and P-value for chi-squared test were calculated to control the result. Finally 3257 associations with p-value<0.05 were retained.

Drug-Target-ADR

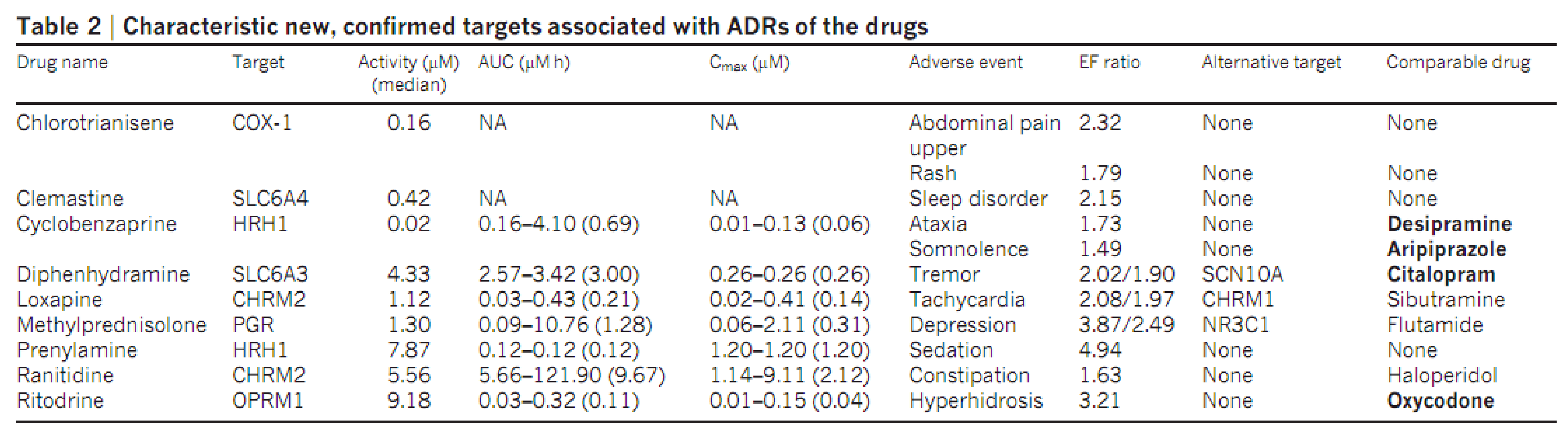


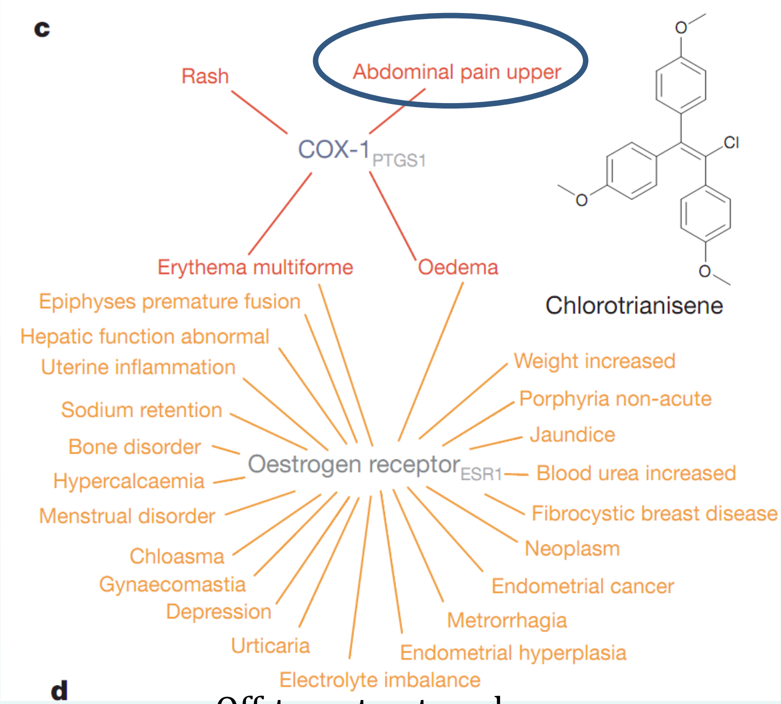
Table 2 | Characteristic new, confirmed targets associated with ADRs of the drugs

Drug name	Target	Activity (μM) (median)	AUC (μM h)	C_{max} (μM)	Adverse event	EF ratio	Alternative target	Comparable drug
Chlorotrianisene	COX-1	0.16	NA	NA	Abdominal pain upper	2.32	None	None
					Rash	1.79	None	None
Clemastine	SLC6A4	0.42	NA	NA	Sleep disorder	2.15	None	None
Cyclobenzaprine	HRH1	0.02	0.16–4.10 (0.69)	0.01–0.13 (0.06)	Ataxia	1.73	None	**Desipramine**
					Somnolence	1.49	None	**Aripiprazole**
Diphenhydramine	SLC6A3	4.33	2.57–3.42 (3.00)	0.26–0.26 (0.26)	Tremor	2.02/1.90	SCN10A	**Citalopram**
Loxapine	CHRM2	1.12	0.03–0.43 (0.21)	0.02–0.41 (0.14)	Tachycardia	2.08/1.97	CHRM1	Sibutramine
Methylprednisolone	PGR	1.30	0.09–10.76 (1.28)	0.06–2.11 (0.31)	Depression	3.87/2.49	NR3C1	Flutamide
Prenylamine	HRH1	7.87	0.12–0.12 (0.12)	1.20–1.20 (1.20)	Sedation	4.94	None	None
Ranitidine	CHRM2	5.56	5.66–121.90 (9.67)	1.14–9.11 (2.12)	Constipation	1.63	None	Haloperidol
Ritodrine	OPRM1	9.18	0.03–0.32 (0.11)	0.01–0.15 (0.04)	Hyperhidrosis	3.21	None	**Oxycodone**

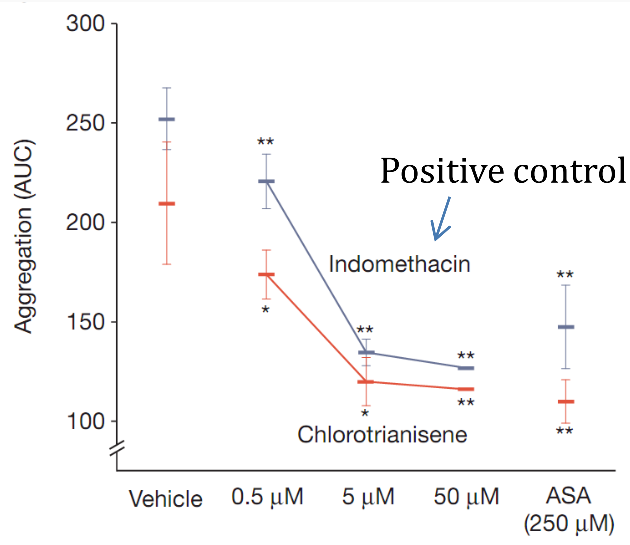
Of all 151 new confirmed drug-target association(IC50<=30uM), 82 were significantly associated with one or more ADR, resulting in a total of 247 drug–target–ADR links. In 116 cases, the EF of the new drug–target–ADR link was stronger than that for any previously known target (*Alternative target* in the table).

The author found that many comparable drugs(in table) with similar activity, pharmacodynamics and pharmacokinetics share the same ADR and could be further used to confirm the drug-target-ADR relationship.

Drug-Target-ADR Net



Off-target network



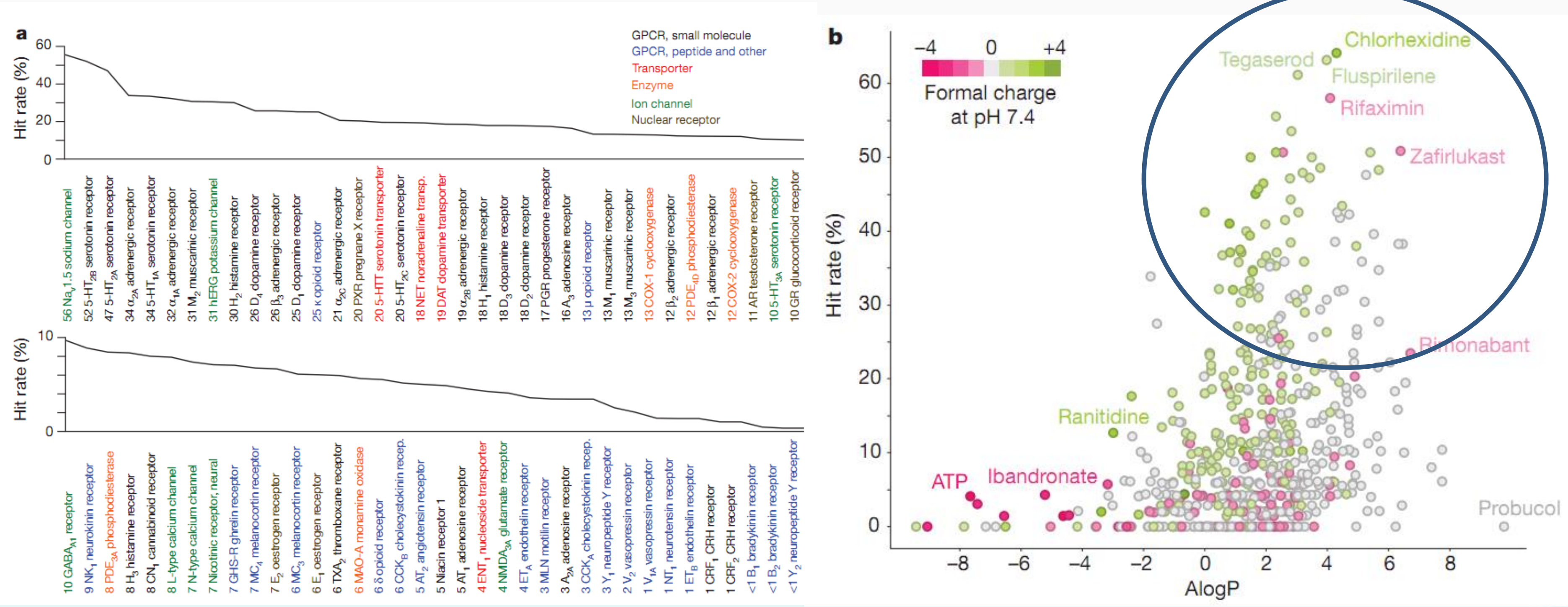
Positive control

ex vivo test for platelet aggregation

The visual network could help to find the shared ADR between different targets and even new ADR of the drug. Example as chlorotrianisene(氯烯雌醚,非甾体激素), which could target COX-1 causing abdominal pain upper(epigastralgia,胃脘痛), similar to NSAIDs(非甾体抗炎药). Generally, synthetic oestrogen(雌激素) may promote platelet aggregation(血小板聚集); however, NSAIDs inhibit platelet aggregation via inhibition of COX-1. The *ex vivo* test confirmed the speculation: chlorotrianisene could promote platelet aggregation as NSAIDs, not as synthetic oestrogen.

Target and drug promiscuity

Hit rate=tested hit drugs(targets)/total tested drugs(targets)



- •The hit rate for the targets and drugs may reflect their promiscuity(杂泛性). The larger hit rate for target means that the target may be affected by more drugs.
- •The right picture show the target promiscuity, showing that GPCR for small molecules were most promiscuity and reversely for peptides.
- •The left picture show the drug promiscuity, showing that the positive hydrophobic drugs may be more promiscuity, thus they may have more ADRs.

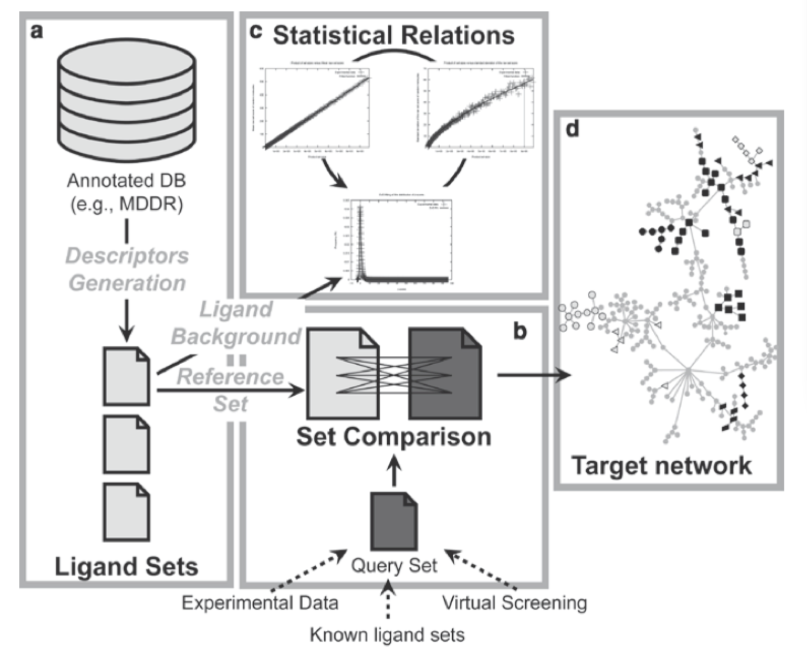
Summary

- SEA is a method for predicting the relation between query ligands with target based on chemical similarity (query ligands *vs* target ligands) and statistical method.

- Using SEA to study the off-target and ADR could help to improve the understanding of the nature of the drug and to optimize the drug design procedure. Many off-target activities might be found before pre-clinical.

Thank you!

How does SEA work ?

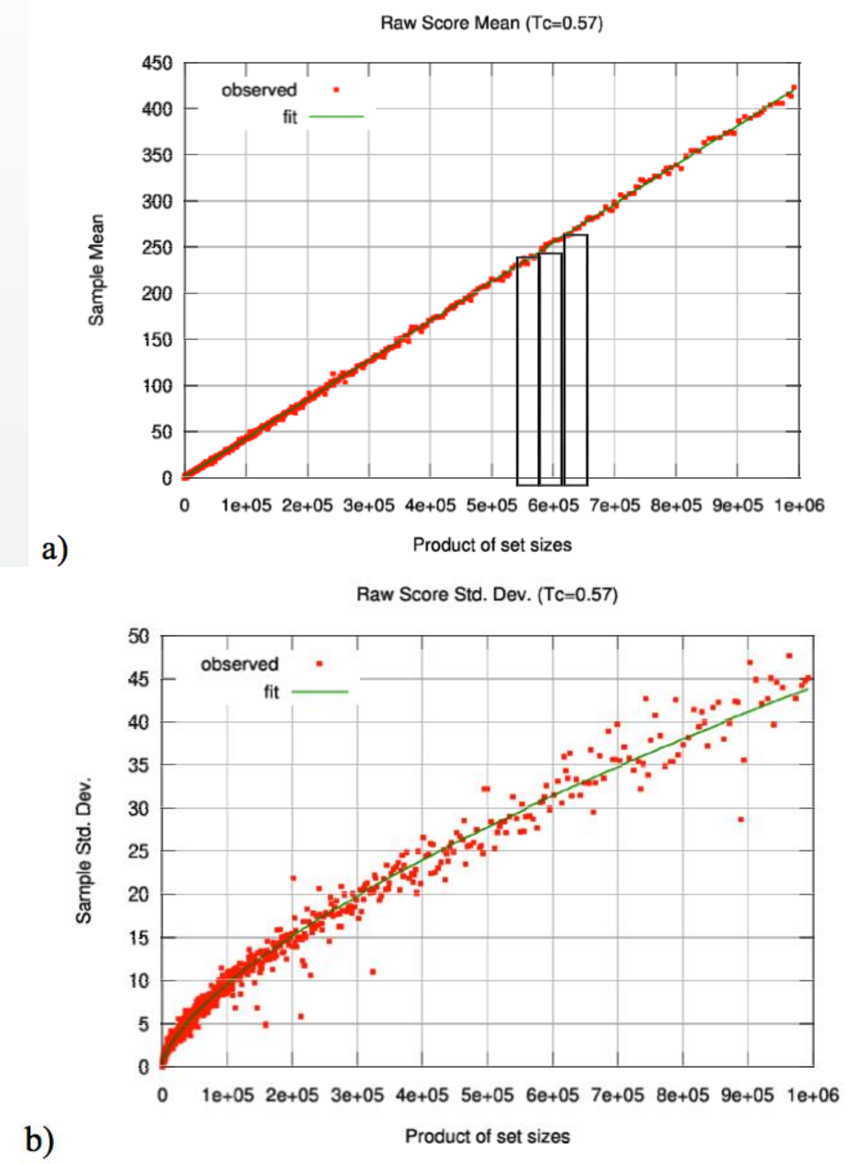


- Database：10W 分子
（Ligand Set size： S_{min}=10，S_{max} =300）
- 抽取1000个整数S_i ，范围在 $S_{min}*S_{min}$~$S_{max}*S_{max}$
- 对于每个S_i，抽取30个整数因子f_i ，建立 set a，set b 含有f_i ，S_i/f_i个分子
- 计算set a，set b的相似性(FP,TC)：$C_{a,b}$
- 相似性阈值t_i =0~1 (step：0.01)

$$r_{a,b}(t_i) = \sum C_{a,b} \quad (C_{a,b} > t_i)$$

In breif, the apparent raw score from the query comparison need to find the mean raw score and raw score standard deviation. Then the set comparison Z-scores were calculated as a function of the set raw scores, expected raw scores and s.d. The histogram of Z-scores (right) of the random sets conformed to an extreme value distribution, which underlied BLAST comparisons of protein and DNA sequences. Finally, the probability of the score being achieved by random chance alone, given the Z-score, was converted to an expectation value (E-value).

1. Calculating the parameters of the reference database(1)



对于每个t_i作图，共100次

横坐标：|set a| * |set b|
set size: S_i

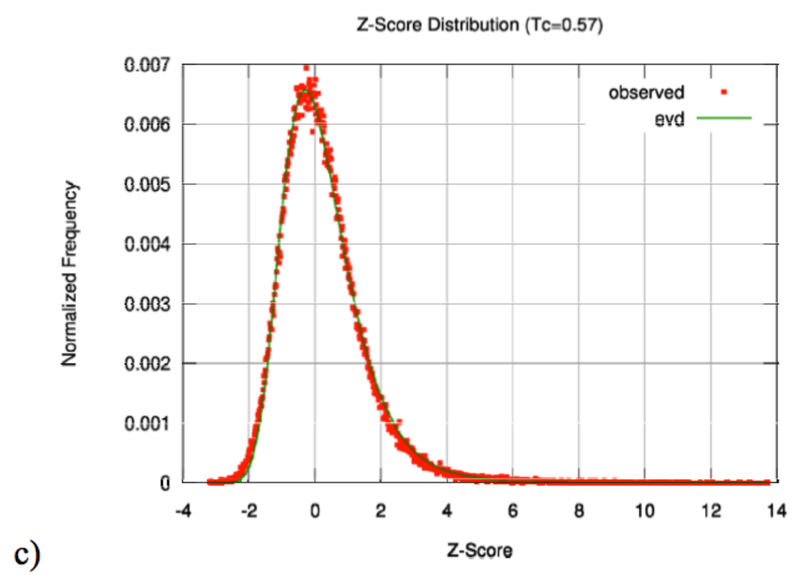
纵坐标：$r_{a,b}(t_i)$

拟合出均值 $y_\mu = mx^n + p$

沿着图a的横坐标取很多个bin，每个bin不少于5个点

根据均值计算标准差
拟合出标准差方程 $y_\sigma = qx^r + s$

1. Calculating the parameters of the reference database(2)



Transformation of Raw scores to Z-scores

$$z = (rs(S_1,S_2) - \mu(n(S_1,S_2))) / \sigma(n(S_1,S_2))$$

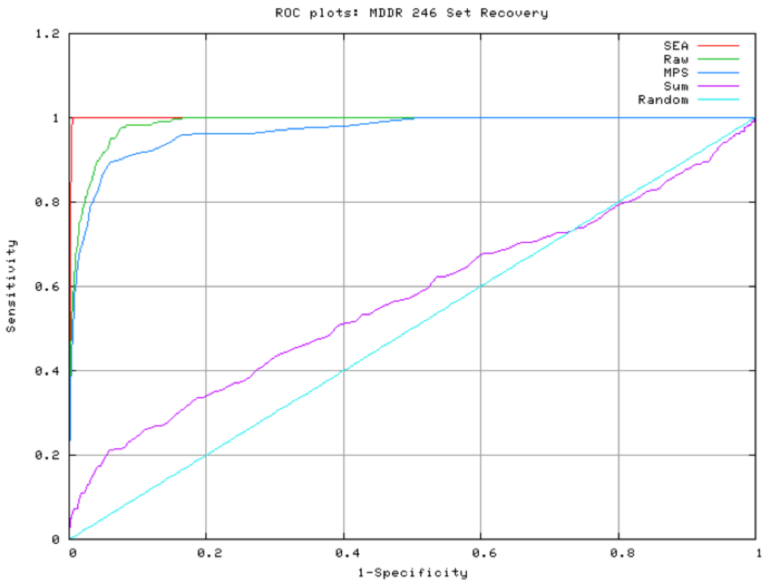
对于每个阈值t_i，做Z值的分布，
找出100个t_i中最接近极值分布的t_i值

图c 代表了随机抽取分子的相似性得分分布（0的概率最高）。

Z值越大，说明两个分子的相似性越大，随机产生的概率越小（越大越相似）

2. Calculating set-wise similarity ensemble

Supplementary Figure 4 Set recovery in database search over 246 MDDR classes



对所有的target set中进行两两比较

用上一步确定的t_i计算$r_{a,b}(t_i) = \Sigma C_{a,b}$ （$C_{a,b} >$

通过图a，b 对应size的均值和方差，转化为Z-score

$$z = (rs(S_1,S_2) - \mu(n(S_1,S_2))) / \sigma(n(S_1,S_2))$$

$$P(Z > z) = 1 - \exp(-e^{-z\pi/\mathrm{sqrt}(6) - \Gamma'(1)})$$

$$E(z) = P(z)N_{db}$$

图c 代表了随机抽取分子的相似性得分分布（0的概率最高）。

Z值越大，说明两个分子的相似性越大，随机产生的概率越小（越大越相似）

P值的意义：数据库随机搜索条件下，得到大于或等于Z值的概率 （越小，随机产生的概率越低）