Supporting Information

Enhancing the Enrichment of Pharmacophore-Based Target Prediction for the Polypharmacological Profiles of Drugs

Xia Wang[1], Chenxu Pan[2], Jiayu Gong[2], Xiaofeng Liu[1,*], Honglin Li[1,2,*]

[1]Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

[2]School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Table of contents

Appendix 1. PharmMapper Server.

PharmMapper server is a freely accessed web server designed to identify potential target candidates for the given small molecules (drugs, natural products or other newly discovered compounds with unidentified binding targets) using pharmacophore mapping approach. PharmMapper hosts a large, in-house repertoire of pharmacophore database (namely PharmTargetDB) annotated from all the targets information in TargetBank, BindingDB[1], DrugBank[2] and potential drug target database, including over 7000 receptor-based pharmacophore models (covering over 1500 drug targets information). PharmMapper automatically finds the best mapping poses of the query molecule against all the pharmacophore models in PharmTargetDB and lists the top N best-fitted hits with appropriate target annotations, as well as respective molecule's aligned poses are presented.

Pharmacophore Databases Construction. PharmMapper requires a sufficient number of available pharmacophore models describing the binding modes of known ligands at the binding sites of protein targets. The target protein structures co-complexed with small molecules were carefully selected from DrugBank, BindingDB, PDBBind[3] and our PDTD[4] databases. Only those proteins with available 3D crystal structures were selected and used for pharmacophore model extraction.

LigandScout[5] was used in the process of pharmacophore model derivation. Each ligand binding site was manually analyzed after generation of corresponding pharmacophore model and the corresponding shape was characterized by several excluded volumes centered at each residue of the binding pocket. All the small ligands with molecular weight lower than 100, such as solvents, buffers and metal cations, and all the cofactors with molecular weight over 600, such as CoAs, polypeptides and nucleic acids were regarded as ‘environment atoms’ instead of binding ligands. For the proteins existing as homopolymers, only one monomer was reserved for analysis. For the proteins determined by NMR with multiple structure models, only the first model was selected for pharmacophore generation. As a result, we generated 7302 pharmacophore models and deposited them in PharmTargetDB.

Pharmacophore Mapping Algorithm. PharmMapper is based on semi-flexible fitting strategy for efficiency, multiple conformers of the given ligand are necessary. The conformational generation process can be performed either online (by in-house conformational analysis program, Cyndi[6]) or offline (by uploading user-specified conformational ensemble generated by third-party conformational analysis methods). In addition, the ligand is assigned with a set of physico-chemical features corresponding to the seven pharmacophore features characterized in PharmTargetDB according to predefined topological-based atom typing rules. In the second stage, each triplet of feature points (e.g., H-H-H, H-HBA-HBD) is enumerated and stored in a hash table for both pharmacophore models and ligand conformers. For vector features (HBA, HBD and A), the projected points of the feature atoms are also involved. The three feature types of the vertexes in the triangle and the corresponding length of each edge are encoded as the searching index in the triangle hashing table. This method is similar to the concept of “geometric hashing” which has been widely used in pharmacophore mapping and molecular alignment[7]. In the third stage, each query triangle from ligand is checked against the pre-computed triangle ensemble of the target pharmacophore models in the TriHash table to verify if the features of corresponding three vertexes are identical; meanwhile, the lengths of the triangle edges are queried against the hash table to check whether their differences are within the adjustable tolerance threshold. Therefore, the feature triangles can be efficiently compared to check if they are congruent. For each pair of matched triangles, a Kabsch algorithm[8] was implemented to achieve the optimal alignment between the two triangles. For each alignment, the fit value score is calculated to evaluate the pairwise alignment between the ligand pose and pharmacophore model. In the last stage, all the binding poses are ranked according to respective fit scores in descending order and only the top one pose was reserved, which wrapped up a mapping process for one target candidate.

To evaluate the alignments between pharmacophores and ligand conformations, a distance-dependent score is defined. The score, which only depends on the relative positions of each pair of feature points from the pharmacophore and ligand respectively, is the weighted sum of point score and vector score in the form of equations (1) and (2) as follows:

$$\text{PointScore} = \sum_{i=1}^{n} w_i F(p)_i = \sum_{i=1}^{n} w_i \begin{cases} 1-\left[\dfrac{d(p)_i}{T(p)_i}\right]^2 & ,d(p)_i \le T(p)_i \\ 0 & ,d(p)_i > T(p)_i \end{cases} \qquad (1)$$

The fit score between the ligand's point feature p and corresponding one in pharmacophore model is normalized to the range of [0, 1], according to whether the distance d(p)i between these two points exceeds the corresponding matching tolerance value T(p)i.

The point score is essentially obtained by calculating the Euclidian distances between each pair of points with the same type, and is scaled to the range of 0 to 1 with a binomial function. The vector score, on the other hand, is the same type of point score calibrated by the direction difference between the aligned vectors as depicted in equation (2):

$$\text{VectorScore} = \sum_{i=1}^{n} w_i F(v)_i * \cos\theta \qquad (2)$$

Where F(v)i has the same format as F(p)i and θ is the angle between the two vectors from the matching pair of vector features in the ligand and pharmacophore.

The penalty score is calculated in the same way as Point Score does. It’s a weighted sum of point scores between the heavy atoms in the ligand and the excluded volumes in the pharmacophore as calculated with equation (1). By default, all the weights of exclude volumes are identical if we don’t know which residues are more flexible upon the ligand’s binding.

The final fit score is a weighted summation of PointScore and VectorScore subtracting the penalty score. The algorithm uses default weight values for each feature type (1.0 for both point and vector features), unless other values are provided by the user.

Appendix 2. DrugBank Test Set Performance.

The results of Z'-score and fit score for DrugBank test set are shown in Table S6. It can be observed that fit score was able to correctly predict 289 of 2080, or 13.9%, of the drug-target associations at top 1% rank stage, while Z'-score produced a higher accuracy at the same rank stage (379/2080), agreeing with the conclusion obtained through the previous examples that Z'-transformation performs much better than simple fit score, particularly at early stages. Moreover, for all the 2080 drug-target pairs, among which 1295 pairs were improved by Z'-score, while 767 pairs saw a decrease in their rankings. This might be due to the reason that the number of valid PharmMapper fit scores for some targets in the ligand-target matrix is not large enough to reflect the distribution of the corresponding target in the background database.

Detailed analysis of specific drugs further highlights the potential of Z'-score. For example, quinolinic acid (QA, DrugBank ID DB01796) is a metabolite of tryptophan with a possible role in neurodegenerative disorders. The crystal structure of quinolinic acid phosphoribosyltransferase (QAPRTase) bound with QA (PDB code 1QAP)[9] is present in our pharmacophore database. With Z'-score, the ranking of QAPRTase was improved significantly comparing with fit score (from 275 to 30). 2,4-Diamino-6-[N-(3',4',5'-Trimethoxybenzyl)-N-Methylamino]Pyrido[2,3-D]Pyrimidine (DB02919), another molecule in DrugBank test set, was reported to be an inhibitor of dihydrofolate reductase (DHFR). Likewise, the protein structure of DHFR in complex with DB02919 (PDB code 1MVT)[10] was involved in PharmTargetDB.

The ranking of DHFR was also improved a lot by Z'-score comparing with fit score (from 204 to 13).

Appendix 3. Targets Distribution in Background Database.

For testing the Gaussianity of target *T* distribution we propose to use the statistical analysis tool Jarque-Bera test (JB test)[11]. The JB test evaluates the hypothesis that the input data vector *X* has a normal distribution with unspecified mean and variance against the alternative that *X* does not follow a normal distribution. The test statistic JB is defined by:

$$\mathrm{JB} = \frac{n}{6}(s^2 + \frac{(k-3)^2}{4})$$

Where *n* is the sample size, *s* and *k* stand for skewness and kurtosis, respectively.

The *p* values for the jbtest were evaluated in MATLAB using its standard procedure "jbtest", which is described as [h,p] = jbtest(x,alpha), where x: the sample column vector; alpha: significance level of the hypothesis test. If the returned test decision (*h*) is 1, reject hypothesis H_0, and 0 otherwise.

Histograms and normal probability plots were also used to visually measure if the data fits a normal density function. The former plots a histogram of values in data using the number of bins equal to the square root of the number of elements in data, where the axis *X* means the Z'-scores of the specific target *T*, and the axis *Y* means the frequency of Z'-scores. The latter depicts the empirical cumulative distribution of the sample data versus the theoretical cumulative distribution function of a normal distribution. The horizontal axis plots the sorted Z'-score, and the vertical axis plots the normal order statistic medians. If the data has a normal distribution, then the plot appears linear. Distributions other than normal would introduce curvature in the plot.

Table S7 shows the results of MATLAB jbtest function for the Z-scores of target *T*. As can be seen, 505 targets are normally distributed and this fact supports visual observations (histograms and normal probability plots for individual targets can be accessed via PharmMapper http://lilab.ecust.edu.cn/pharmmapper/download.php). For the rest targets of PharmTargetDB, JB tests produce significantly low *p* values, indicating that all of which are not normally distributed. However, although Z-scores vector of most targets fail to satisfy the JB test, it can still reflect the distribution of the background database to some extent (see the histograms and normal probability plots), and thus could be used to discriminate the simple fit scores.

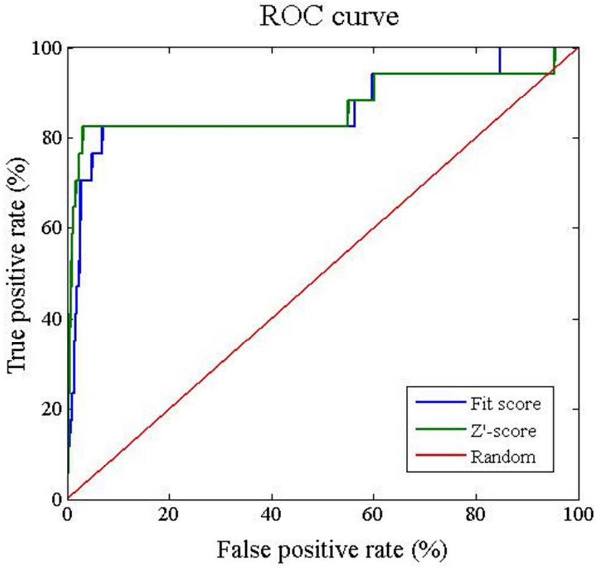

Figure S1. Comparison of ROC curves of fit score and Z'-score for target identification of S-Adenosyl-L-Homocysteine.

Table S2. Targets identified for S-Adenosyl-L-homocysteine by Z'-score and fit score at 0.5% FPR.

Z'-score			**Fit score**		
Rank	**PDB ID**	**FPR%**	**Rank**	**PDB ID**	**FPR%**
3	1AQJ	0.03	4	1NW7	0.05
12	1NW7	0.16	16	1AQJ	0.22
18	1BOO	0.24	/	/	/
27	1PJT	0.36	/	/	/
28	1L1E	0.36	/	/	/
29	1MXI	0.36	/	/	/

Table S3. Details of ROCE of Z'-score and fit score in target identification of the DrugBank subset at different stages. The step size of rank was set to 0.1%.

Rank%	Z'-score			Fit score		
	TPR%	**FPR%**	**ROCE**	**TPR%**	**FPR%**	**ROCE**
0.1	4.47	0.10	45.11	2.84	0.10	28.52
0.2	7.55	0.20	38.02	4.95	0.20	24.88
0.3	9.76	0.30	32.74	6.49	0.30	21.72
0.4	11.63	0.40	29.25	7.69	0.40	19.30
0.5	12.88	0.50	25.90	9.38	0.50	18.82
0.6	14.28	0.60	23.91	10.38	0.60	17.36
0.7	15.48	0.70	22.21	11.35	0.70	16.26
0.8	16.54	0.80	20.76	12.26	0.80	15.37
0.9	17.50	0.90	19.52	13.08	0.90	14.57
1	18.22	1.00	18.28	13.89	1.00	13.92
2	25.19	2.00	12.63	20.72	2.00	10.37
3	29.95	2.99	10.00	26.39	3.00	8.81
4	33.13	3.99	8.29	29.66	4.00	7.42
5	36.78	4.99	7.37	32.84	5.00	6.57
6	39.33	5.99	6.56	34.86	6.00	5.81
7	41.44	6.99	5.93	38.08	7.00	5.44
8	44.33	7.99	5.55	40.96	8.00	5.12
9	46.97	8.99	5.22	42.55	9.00	4.73

10	49.90	9.99	4.99	44.38	10.00	4.44
25	73.46	24.99	2.94	69.04	25.01	2.76
50	90.82	49.99	1.82	90.10	50.03	1.80
75	98.61	75.00	1.31	98.46	75.05	1.31
100	100.00	99.00	1.01	100.00	99.08	1.01

Table S4. Area under curves achieved by Z'-score, fit score and random way.

Rank %	*$AUC_{Z'}$ %*	*AUC_F %*	*$AUC_{Z'}/AUC_r$*	*AUC_F/AUC_r*	*$AUC_{Z'}/AUC_F$*
0.1	0.00	0.00	90.47	57.13	1.58
0.2	0.01	0.01	60.70	39.19	1.55
0.3	0.02	0.01	48.81	31.91	1.53
0.4	0.03	0.02	42.09	27.59	1.53
0.5	0.05	0.03	37.29	25.19	1.48
0.6	0.06	0.04	33.86	23.28	1.45
0.7	0.08	0.05	31.22	21.75	1.44
0.8	0.09	0.07	29.09	20.49	1.42
0.9	0.11	0.08	27.32	19.43	1.41
1	0.13	0.09	25.78	18.52	1.39
2	0.35	0.27	17.69	13.38	1.32
3	0.63	0.51	14.10	11.28	1.25
4	0.95	0.79	11.93	9.87	1.21
5	1.31	1.10	10.47	8.84	1.19
6	1.69	1.44	9.40	8.03	1.17
7	2.09	1.81	8.57	7.39	1.16
8	2.52	2.20	7.90	6.90	1.15
9	2.98	2.62	7.38	6.48	1.14
10	3.47	3.06	6.95	6.12	1.13
25	12.93	11.69	4.14	3.74	1.11

50	33.80	32.07	2.71	2.57	1.05
75	57.99	56.08	2.06	1.99	1.03
100	82.99	81.08	1.66	1.62	1.02

Table S5. Details and ROC curve and number of true positives of Z'-score and fit score in target identification of the ChEMBL subset at different stages. The step size of rank was set to 0.1%.

Rank%	Z'-score				Fit score			
	TPR%	FPR%	ROCE	TP	TPR%	FPR%	ROCE	TP
0.1	5.36	0.10	53.95	135	4.13	0.10	41.50	104
0.2	7.27	0.20	36.49	183	5.92	0.20	29.69	149
0.3	9.21	0.30	30.82	232	7.23	0.30	24.16	182
0.4	10.41	0.40	26.09	262	8.38	0.40	21.00	211
0.5	11.40	0.50	22.85	287	9.53	0.50	19.10	240
0.6	12.51	0.60	20.90	315	10.44	0.60	17.44	263
0.7	13.74	0.70	19.67	346	11.08	0.70	15.86	279
0.8	14.50	0.80	18.16	365	12.11	0.80	15.17	305
0.9	15.45	0.90	17.20	389	13.03	0.90	14.50	328
1	16.04	1.00	16.07	404	13.90	1.00	13.92	350
2	22.00	2.00	11.01	554	19.42	2.00	9.72	489
3	26.29	3.00	8.77	662	23.71	3.00	7.91	597
4	29.63	4.00	7.41	746	26.81	4.00	6.71	675
5	33.28	5.00	6.66	838	29.98	5.00	6.00	755
6	35.90	6.00	5.99	904	32.49	6.00	5.42	818
7	38.17	7.00	5.46	961	34.51	7.00	4.93	869
8	40.47	8.00	5.06	1019	36.46	8.00	4.56	918

9	42.89	9.00	4.77	1080	39.08	9.00	4.34	984
10	45.47	10.00	4.55	1145	40.91	10.00	4.09	1030
25	68.23	24.99	2.73	1718	64.85	25.00	2.59	1633
50	89.56	50.00	1.79	2255	87.49	50.00	1.75	2203
75	99.56	75.00	1.33	2507	99.17	75.00	1.32	2497
100	100.00	100.00	1.00	2518	100.00	100.00	1.00	2518

REFERENCES

(1) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198-D201.

(2) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901-D906.

(3) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977-2980.

(4) Gao, Z.; Li, H.; Zhang, H.; Liu, X.; Kang, L.; Luo, X.; Zhu, W.; Chen, K.; Wang, X.; Jiang, H. PDTD: a web-accessible protein database for drug target identification. *BMC Bioinf.* **2008**, *9*, 104.

(5) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160-169.

(6) Liu, X.; Bai, F.; Ouyang, S.; Wang, X.; Li, H.; Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinf.* **2009**, *10*, 101.

(7) Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Deterministic pharmacophore detection via multiple flexible alignment of drug-like molecules. *J. Comput. Biol.* **2008**, *15*, 737-754.

(8) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta*

Crystallogr. **1976**, *32*, 922-923.

(9) Eads, J. C.; Ozturk, D.; Wexler, T. B.; Grubmeyer, C.; Sacchettini, J. C. A new function for a common fold: the crystal structure of quinolinic acid phosphoribosyltransferase. *Structure* **1997**, *5*, 47-58.

(10) Cody, V.; Galitsky, N.; Luft, J. R.; Pangborn, W.; Gangjee, A. Analysis of two polymorphic forms of a pyrido[2,3-d]pyrimidine N9-C10 reversed-bridge antifolate binary complex with human dihydrofolate reductase. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 654-661.

(11) Jarque, C. M.; Bera, A. K. A test for normality of observations and regression residuals. *Int. Stat. Rev.* **1987**, 163-172.