Journal of Molecular Graphics and Modelling 44 (2013) 278–285

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

A combined molecular docking-based and pharmacophore-based target prediction strategy with a probabilistic fusion method for target ranking

Guo-Bo Li [a], Ling-Ling Yang [a,b], Yong Xu [a], Wen-Jing Wang [a], Lin-Li Li [a,c], Sheng-Yong Yang [a,*]

[a] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Sichuan 610041, China
[b] College of Chemical Engineering, Sichuan University, Sichuan 610041, China
[c] West China School of Pharmacy, Sichuan University, Sichuan 610041, China

ARTICLE INFO

Article history:
Accepted 12 July 2013
Available online 23 July 2013

Keywords:
Drug target prediction
Molecular docking
Pharmacophore
A probabilistic fusion method

ABSTRACT

Herein, a combined molecular docking-based and pharmacophore-based target prediction strategy is presented, in which a probabilistic fusion method is suggested for target ranking. Establishment and validation of the combined strategy are described. A target database, termed TargetDB, was firstly constructed, which contains 1105 drug targets. Based on TargetDB, the molecular docking-based target prediction and pharmacophore-based target prediction protocols were established. A probabilistic fusion method was then developed by constructing probability assignment curves (PACs) against a set of selected targets. Finally the workflow for the combined molecular docking-based and pharmacophore-based target prediction strategy was established. Evaluations of the performance of the combined strategy were carried out against a set of structurally different single-target compounds and a well-known multi-target drug, 4H-tamoxifen, which results showed that the combined strategy consistently outperformed the sole use of docking-based and pharmacophore-based methods. Overall, this investigation provides a possible way for improving the accuracy of in silico target prediction and a method for target ranking.

1. Introduction

Drug target identification is extremely important not only for determining mechanism of action of active agents but also for anticipating their side effects or exploring possible new therapeutic indications of old drugs [1–5]. The most direct methods for the target identification correspond to those based on chemical biology [6–8]. However, these methods often require many expensive and time consuming wet experiments. In order to reduce the cost and save time, various computational methods [9], which are generally much cheaper and faster, have been involved in this kind of task. Because the predicted targets by computational methods still need further confirmation by wet experiments, a hybrid mode of target identification has been widely adopted at present, in which computational methods are first used to predict the potential targets, followed by validation by wet experiments. In this mode, the target prediction ability of computational methods is fairly important for the final success of target identification [9,10].

Currently a number of sophisticated computational methods have been established for the target prediction, which mainly include molecular docking-based, pharmacophore-based, molecular similarity-based, and others. A molecular docking-based method tries to dock a query compound to a panel of known target proteins to determine which one is the most likely interaction partner according to the scoring function. The representative examples of this method are INVDOCK [11] and TarFisDock [12]. A pharmacophore-based method finds the best mapping poses of the query molecule against a set of predefined pharmacophore models, in which each one corresponds to a target, and outputs the top best-fitted hits as the target candidates. PharmMapper is one of the typical representatives [13]. A molecular similarity-based method simply compares a query compound with a database of compounds whose targets are known. If the query compound is similar in structure with some compounds in the database, the targets of these compounds are considered as the target candidates of the query compound. This method is relatively simple and has more applications in recent years [5,14]. Other methods such as machine learning-based [15,16] and biochemical network-based [17–19] have also been developed recently.

Though each method has its own inherent advantages and disadvantages, which have been discussed in literature [10,20,21],

* Corresponding author. Tel.: +86 28 85164063; fax: +86 28 85164060.
E-mail address: yangsy@scu.edu.cn (S.-Y. Yang).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.07.005

these methods have some common problems. Of which the biggest problem for all of these methods is the poor target prediction ability. In finding a solution to this problem, we thought of a combined strategy of these methods, which has been used successfully in virtual screening by us [22,23] as well as other groups [24–26]. We thus, in this investigation, proposed a combined molecular docking-based and pharmacophore-based target prediction strategy. Here we chose the combination of the molecular docking-based and pharmacophore-based methods mainly because the two methods are apparently complementary. For example, the scoring function and the protein flexibility problems are obsessions in the docking-based method [20], whereas they are not a problem anymore in the pharmacophore-based method. The pharmacophore-based method often lacks consideration of receptor structural information [21], while it is a strong point of docking-based method. Even so, there is still a problem when using the combined strategy in target prediction, namely, how to sort the targets predicted by these methods. Here, we adopted a probabilistic fusion method for target ranking, which is based on Belief Theory (also known as Dempster–Schafer Theory) [27–29].

2. Methods

2.1. The target database

To construct a comprehensive potential target database (TargetDB), we first collected potential drug targets as many as possible from several public databases, including Therapeutic Target Database (TTD) [30], Potential Drug Target Database (PDTD) [31], DrugBank [32], and RSCB Protein Data Bank (PDB) [33]. Only those protein targets whose protein–ligand complex structures are known were selected. A total of 1105 different targets were deposited in TargetDB. Meanwhile, we also noticed that many of these targets have two or more crystal structures in the PDB database (see Supplementary Fig. S1). Thus, for some targets, several crystal structures are included; these structures have a relatively large difference. The finally formed TargetDB contains 1481 crystal structures covering the selected 1105 drug targets. These targets were annotated with biochemical type, therapeutic disease and development state.

2.2. The binding site database and the pharmacophore database

Based on TargetDB, we further constructed a binding site database and a pharmacophore database. Before the compilation of these databases, all the structures in TargetDB were prepared by utilizing DS 3.1 (Discovery Studio 3.1, Accelrys, Inc., San Diego, CA) software package. Operations for the preparation included: (i) removing water molecules and buffers, but preserving pivotal enzyme cofactor and metal cations; (ii) assigning CHARMM force filed [34]; (iii) for the structures with homopolymers, only one monomer was reserved; (iv) for the structures determined by NMR with multiple conformations, only the first conformation was remained.

The commercial molecular docking program GOLD [35] (CCDC, Cambridge CB2 1EZ, UK) was used in the docking-based target prediction; GOLD was chosen since it is one of the most widely used docking programs and has shown a better performance in virtual screening. Accordingly, the binding site database was created using GOLD, in which a binding site was defined as a sphere that contains all the residues around the ligand in the complex structure. A configuration file (gold.cfg) for each crystal structure including the absolute path of the corresponding protein target file and the 3-D coordinates of the binding site center was also recorded and saved for later use.

The pharmacophore database, which will be used in the pharmacophore-based target prediction method, was constructed using the module 'Receptor–Ligand Phamacophore Generation' implemented in the DS 3.1 software package. Six pharmacophore features, including hydrogen-bonding acceptor, hydrogen-bonding donor, aromatic ring, hydrophobic feature, positive charge center, and negative charge center, were considered in the model building process. Other parameters for the program were set as default. The program generated ten pharmacophore models for each complex, and the model with the highest score was selected to stay in the pharmacophore database. Overall, we finally obtained a binding site database containing 1481 binding sites and a pharmacophore database comprising 1481 pharmacophore models.

2.3. The docking-based and pharmacophore-based target prediction protocols

The GOLD program was taken as the docking engine in the docking-based target prediction method. The protocol or workflow for the docking-based target prediction method can be briefly described as follows: (i) preparing the query compound; (ii) docking the query compound to each binding site in the binding site database using GOLD, and calculating two scoring functions: Chemscore (empirical) [36] and Goldscore (force field-based) [37]; (iii) preserving the best docking pose for each target, and extracting the corresponding scoring values; (iv) prioritizing the targets according to the scoring values of Chemscore and Goldscore, respectively. The top-ranking targets are supposed to be the most potential targets of the query compound.

The Catalyst program [38] implemented in DS 3.1 software package was used in the pharmacophore-based target prediction method. The protocol or workflow for the pharmacophore-based target prediction method can be simply described as follows: (i) generating conformers of the query compound using the 'fast conformer generation' approach with 20 kcal/mol being set as the energy cutoff and 250 as the maximum number of conformers; (ii) mapping the generated conformers onto each pharmacophore model in the pharmacophore database using a grid-fitting method; (iii) calculating the fitness value, which is used to define how well a given compound is mapped to a pharmacophore model, according to the following formula (Eq. (1)) [39]:

$$Fitness = \frac{\sum_n [1 - \sum (d/t)^2]}{n} \quad (1)$$

where n denotes the number of pharmacophore features, d represents the displacement of the feature from the center of the location constraint, t is the radius of the location constraint sphere for the feature (tolerance); (iv) prioritizing all the pharmacophore models (actually they correspond to targets) in the pharmacophore database according to the fitness values. The top best-fitted hits are considered as the target candidates of the query compound.

2.4. The probabilistic fusion method

To provide a reasonable ranking order for the targets in the combined docking-based and pharmacophore-based target prediction method, we introduced a probabilistic fusion method, which is based on Belief Theory (also known as Dempster–Schafer Theory) [27]. The basic requirement of Belief Theory is that quantifiable probabilities of an event being true can be obtained. For satisfying this requirement, we created a training set to construct probability assignment curves (PACs), which are empirically derived functions that can translate a measure (e.g. Chemscore) into a probability of true prediction by this measure. The training set contains 20 protein targets, which cover a variety of biochemical types (see Supplementary Table S1). For each target, 200 known ligands or actives

Table 1
Sigmoidal curve parameters for the probability assignment curves in Fig.2.

Measure	Parameters in Eq. (3)			
	F_{min}	F_{max}	Slope	SC_{50}
Chemscore	0.018	0.5	0.76	4.05
Goldscore	0.026	0.5	2.19	1.85
Fitness	0.023	0.5	1.28	3.01

(IC_{50} < 10 μM) were chosen from the BindingDB database [40], and 4000 decoys [41] were sampled judiciously from the ZINC database [42]. The selected decoys have similar physiochemical properties but different topological structures with the known ligands (see Supplementary Fig. S2). The selection method for the decoys is given in the Supplementary Methods.

The construction process for the PACs of the proposed measures, including Chemscore, Goldscore, and Fitness, is briefly described as follows. First, the specific set of actives and decoys for each target in the training set were docked to the corresponding target, and the scoring functions, Chemscore and Goldscore, were calculated. Meanwhile, the actives and decoys for each target were mapped to the corresponding pharmacophore model, and the fitness values were calculated (see Eq. (1)). Second, actives and decoys for the entire 20 targets together with their corresponding Chemscore, Goldscore and Fitness values were gathered together to form a training set. Third, for each measure (namely, Chemscore, Goldscore and Fitness), the values for all the actives and decoys in the training set were transformed to Z-score values using the following equation (Eq. (2)), which purpose is to normalize the values to a same scale:

$$Z\text{-score} = \frac{x - \mu}{\sigma} \quad (2)$$

where x is the measure value, μ is the mean value for the decoys, and σ is the standard deviation of the distribution of values across the decoys. Fourth, the Z-score values for each measure were divided into small equal intervals. In each interval, we calculated the fraction of actives within this interval. The resulting plots of Z-score value versus fraction active closely resemble standard dose–response curves (see Fig. 2), and the data were therefore fit to sigmoidal curves of the following formula (Eq. (3)), which is very similar to that proposed by Hajduk et al. [28,29].

$$P_i = \frac{F_{max}}{1 + e^{(SC_{50} - x_i) \times \text{slope}}} + F_{min} \quad (3)$$

where x_i is the Z-score value calculated from the i-th measure, P_i is the probability of a compound binding to a target predicted by the i-th measure given x_i, F_{max} is the maximum value for the fraction active, F_{min} is the minimum value for the fraction active, SC_{50} (by analogy to the IC_{50}) is the cutoff value at which 50% of the maximum fractional active is observed, and slope is the steepness of the curve. The fitted sigmoidal curve parameters are given in Table 1. Hereafter the fitted sigmoidal curves will be termed as probability assignment curves (PACs).

The probabilistic fusion method was then adopted to combine the probabilities produced by Chemscore, Goldscore and Fitness, which is based on Belief Theory. Belief Theory can provide the framework for the combination of multiple probabilities from different sources using the conjunctive rule (Eq. (4)) [28,29]:

$$C\text{-value} = 1 - \prod_{i=1}^{n} (1 - P_i) \quad (4)$$

where n is the number of the proposed measures, P_i is the probability derived from the PAC of the i-th measure, and C-value is the cumulative probability.

3. Results and discussion

3.1. Profile of the comprehensive target database (TargetDB)

To carry out target prediction, we first constructed a comprehensive target database, called TargetDB. Currently, TargetDB contains a total of 1105 different potential drug targets. Fig. 1A–C shows the distributions of these targets according to biochemical type, therapeutic disease and development state, respectively. In terms of the biochemical types of these targets, they can be roughly classified into four categories: enzyme (including kinase, hydrolase, transferase, oxidoreductase, ligase, isomerase, and synthase), receptor (including nuclear hormone receptor, G-protein coupled receptor, and ionotropic receptor), protein (including binding protein, transport protein, viral protein, signaling protein, cell cycle protein, structural protein, and apoptosis protein), and others (including transcription regulator, cell adhesion molecule, chaperone, cytokine, lectin, toxin, and undefined classes). In terms of the disease types, the targets in TargetDB are associated with more than 405 kinds of diseases, which can be roughly categorized into 10 classes (see Fig. 1B). In addition, the development state of drugs related to the targets in TargetDB have also be investigated, and it was found that there are 126 successful targets, 330 clinical trial targets, and 649 researching targets (see Fig. 1C).

3.2. Development and evaluation of the probabilistic fusion method

To give a reasonable ranking order for the targets in the combined docking-based and pharmacophore-based target prediction method, we introduced a probabilistic fusion method, which is based on Belief Theory. Here three measures, namely Chemscore, Goldscore (both are from molecular docking), and Fitness (obtained in pharmacophore mapping) were calculated. To obtain quantifiable probabilities for the three measures, their corresponding probability assignment curves (PACs) were first constructed, which are shown in Fig. 2. The fitted curve parameters are given in Table 1. The derived probabilities of the three measures (PACs) are finally combined using a conjunctive rule (see Eq. (4)), producing a cumulative probability (C-value) that is a fuse performance of the three measures.

To assess the performance of the probabilistic fusion method as well as individual measure, we first adopted the receiver operating characteristic (ROC) curves. Accuracy was measured using the area under the ROC curve (AU-ROC). An AU-ROC value of 1 represents a perfect test; an AU-ROC value of 0.5 represents a worthless test. Fig. 3 depicts AU-ROC values for Chemscore, Goldscore, Fitness, and C-value on each target in the training set. The average of AU-ROC values among the targets using Chemscore is 0.634, Goldscore is 0.647, and Fitness is 0.681, while the average AU-ROC value by C-value is 0.766. Then, we calculated the enrichment factor at 1% of the ranked molecules for each target in the training set. The calculated enrichment factors are summarized in Supplementary Table S2, from which one can see that C-value performs best. Among the entire 20 targets in the training set, the average enrichment factor is 5.57 for Chemscore, 4.52 for Goldscore, and 8.21 for Fitness, whereas the average enrichment factor is 9.34 for C-value. Thus, we can conclude that this probabilistic fusion method consistently outperforms any single measure.

The performance of the probabilistic fusion method used here was also compared with that of alternative fusion methods such as mean rank and mean Z-score. The results are presented in Supplementary Table S2. The enrichment factor obtained by using the mean rank fusion is lower than that by the probabilistic fusion method and is even lower than that by the Fitness measure alone, which may stem from the fact that the mean rank fusion treats

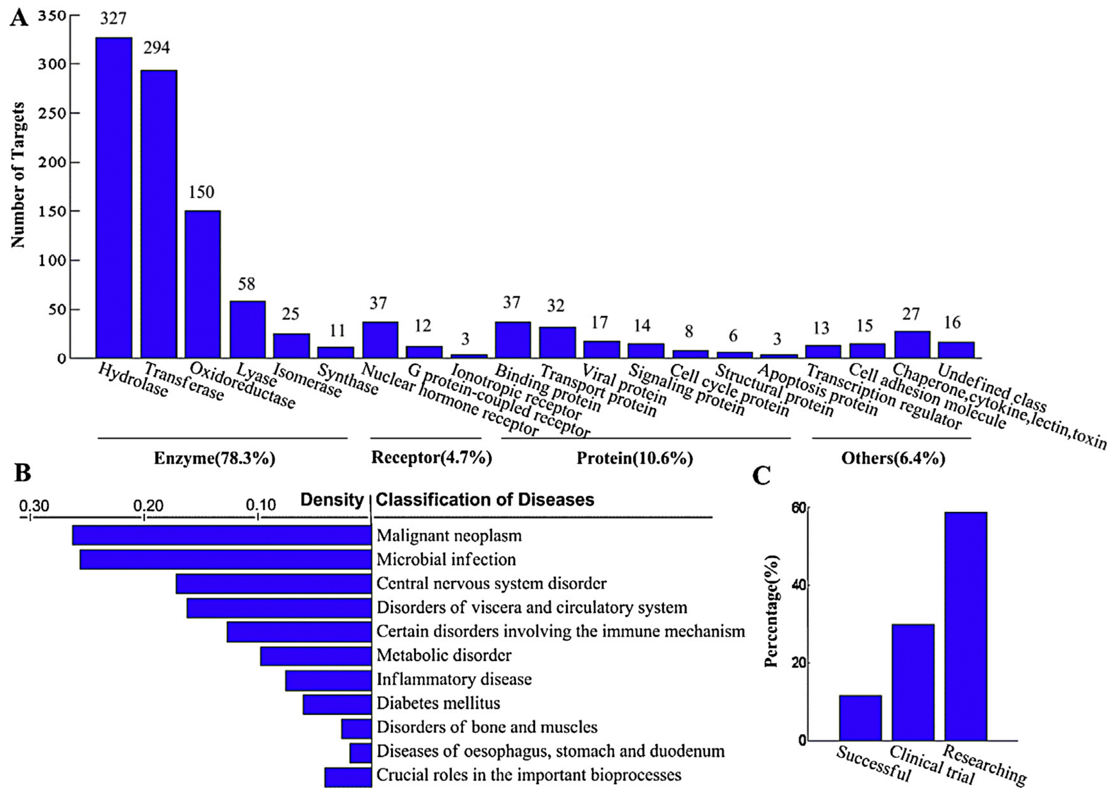

Fig. 1. Distribution of drug targets according to the (A) biochemical type, (B) therapeutic diseases, and (C) development state.

all measures equally. The mean Z-score fusion shows comparable performance to probabilistic fusion method. However, only the probabilistic fusion method can return a quantitative expectation of a compound being active, which can provide effective basis for further experimental validation.

3.3. Workflow for the combined target prediction strategy

Fig. 4 schematically depicts the workflow for the proposed target prediction method. A brief description for this workflow is given as follows. First, a query compound is docked into each binding site in the binding site database, and Chemscore and Goldscore scoring functions are calculated. Meanwhile, the query compound is mapped onto each pharmacophore model in the pharmacophore database, and Fitness values are calculated according to Eq. (1). We now obtain three values for each target in TargetDB: Chemscore, Goldscore, and Fitness. Second, for each target, the three values are translated into their corresponding probabilities according to the PACs (see Fig. 2). Third, the individual probabilities of the three measures for each target are combined into a cumulative probability (C-value) using the conjunctive rule (Eq. (4)). Fourth, all the targets in TargetDB are ranked in descending order according to their C-values.

3.4. Performance evaluation of the combined strategy on the target prediction for a set of single-target compounds

In this section, we shall assess the performance of the docking-based, the pharmacophore-based and the combined strategy on the prediction of true targets for a set of structurally different compounds whose targets are known. We selected 10 marketed drugs from DrugBank database and 90 compounds from the BindingDB database. As far as we known, to date, each of the selected compounds has only one pharmacological target reported; their targets cover a wide variety of biological types (see Supplementary Table S3).

First, the docking-based and pharmacophore-based methods were individually employed to retrieve the targets for the selected

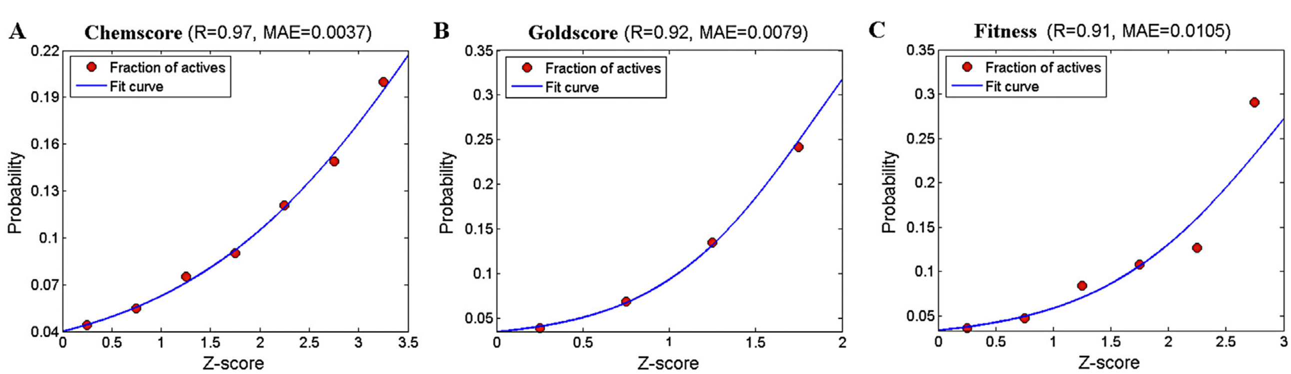

Fig. 2. Probability assignment curves (PACs) and fitted sigmoidal curves for (A) Chemscore, (B) Goldscore, and (C) Fitness.

Table 2
Retrieval of the true targets for the 100 selected compounds using Chemscore, Goldscore, Fitness and *C*-value.[a]

Cpd. id[b]	Target name	PDB code	Chemscore rank	Goldscore rank	Fitness rank	*C*-value rank
1	Thrombin	4ax9	317	49	116	18
2	Penicillin-binding proteins 1	2ex6	60	123	78	21
3	Estrogen receptor	1r5k	49	16	85	37
4	Histamine H1 receptor	3rze	25	7	49	9
5	Beta-1 adrenergic receptor	2y02	37	107	234	156
6	Cytochrome P450 51	2wuz	61	1	1	1
7	Human immunodeficiency virus 1 reverse transcriptase	1c1c	2	28	42	19
8	3-Hydroxy-3-methylglutaryl-coenzyme A reductase	1hw8	30	1	102	2
9	Renin	3d91	1	1	138	2
10	Dihydropteroate synthase	2y5s	127	259	56	33
11	Human immunodeficiency virus 1 protease	1ajv	109	1	244	201
12	Glutamate carboxypeptidase 2	2or4	83	38	22	16
13	Androgen receptor	1e3g	116	79	68	46
14	Hepatitis C virus RNA polymerase	2giq	91	442	179	243
15	Renin	2i4q	17	93	219	107
16	Frnesyl pyrophosphate synthase	3n5h	133	122	16	3
17	Cyclin-dependent kinase 2	3pxy	3	64	33	8
18	Tyrosine-protein kinase lck	3mpm	13	79	230	111
19	11-Beta-hydroxysteroid dehydrogenase	3g49	79	32	87	65
20	Heat shock protein 90	2h55	20	79	73	43
21	Pteridine reductase 1	3mcv	22	1	74	9
22	Polo-like kinase 1	2rku	346	121	260	176
23	Factor Xa	2xbv	282	5	261	52
24	Corticotropin releasing factor receptor type 1	3ehu	139	763	989	687
25	Dipeptidyl peptidase 4	3g0b	100	89	42	23
26	Thymidylate synthase	1ci7	731	40	240	94
27	Human immunodeficiency virus 1 reverse transcriptase	1c1c	26	55	64	5
28	Protein farnesyltransferase	2zis	23	195	137	68
29	Liver carboxylesterase 1	1ya4	55	143	127	71
30	Carbonic anhydrase 2	2qp6	100	295	92	85
31	Renin	2i4q	100	40	310	169
32	Checkpoint kinase 1	2br1	39	275	171	82
33	Farnesyl pyrophosphate synthase	3n5h	35	50	9	1
34	Cyclin dependent kinase 2	1jvp	303	502	33	30
35	Serine/threonine-protein kinase b-raf	1uwh	23	31	148	39
36	Insulin-like growth factor 1 receptor	3lw0	286	411	116	94
37	Mitogen-activated protein kinase 8	2g01	6	334	198	143
38	Tyrosine-protein kinase c-src	2hwp	479	760	10	26
39	Glycogen phosphorylase a	1l5r	133	114	9	4
40	Thymidine phosphorylase	1uou	166	102	33	26
41	Dihydrofolate reductase	3jwf	110	1	117	81
42	Androgen receptor	2hvc	1	63	28	3
43	Renin	2i4q	30	26	234	146
44	Aldose reductase	1pwl	45	6	270	39
45	Uracil-DNA glycosylase	3fck	112	263	45	39
46	Coagulation factor viii	3hnb	859	462	24	107
47	Acetylcholinesterase	1e66	27	41	46	1
48	Poly (ADP-ribose) polymerase 1	2rd6	63	13	55	3
49	Beta-lactamase	1y54	620	581	23	34
50	Leukotriene a4 hydrolase	3chp	25	3	93	28
51	Angiotensin converting enzyme	2x93	343	156	92	32
52	Biotin carboxylase	2w6p	86	18	19	8
53	Phosphoinositide 3 kinase gamma	2chx	5	119	338	114
54	Protein tyrosine phosphatase 1b	1c84	536	6	706	19
55	Dihydroorotate dehydrogenase	3i65	198	238	80	69
56	Purine nucleoside phosphorylase	3e0q	162	181	70	61
57	Thrombin	1jwt	303	122	44	14
58	Polo-like kinase 1	2rku	103	210	219	174
59	Hepatitis C virus RNA polymerase	2hai	94	86	289	70
60	Aldose reductase	1pwl	4	6	24	2
61	Glycogen synthase kinase-3 beta	1q41	202	84	56	36
62	Protein kinase b	3cqw	346	166	126	68
63	Xanthine dehydrogenase	3nvy	123	106	209	96
64	Cathepsin k	2auz	291	239	68	46
65	Dihydroorotate dehydrogenase	1d3h	29	120	91	25
66	Phospholipase a2	1q7a	166	179	2	23
67	Matrix metalloproteinase 13	1xuc	6	26	161	73
68	Casein kinase 2	1m2r	9	21	52	16
69	Mandelate racemase	2p8b	3	48	41	26
70	Acetylcholinesterase	2whr	1	3	115	8
71	Neuraminidase	2ht8	33	13	95	72
72	Aldose reductase	1pwm	7	130	224	198
73	Phosphodiesterase 5	1tbf	46	74	190	110
74	Thrombin	1k21	102	50	161	142

Table 2 (*Continued*)

Cpd. id[b]	Target name	PDB code	Chemscore rank	Goldscore rank	Fitness rank	C-value rank
75	3-Hydroxy-3-methylglutaryl-coenzyme A reductase	1hw8	405	21	207	155
76	Carbonic anhydrase 2	3caj	46	558	72	59
77	Dipeptidyl peptidase 1	2djf	83	52	85	57
78	Nitric oxide synthase	1d1w	5	202	6	1
79	Leukocyte elastase	1h1b	130	105	59	19
80	Adenosine kinase	2i6b	22	90	151	93
81	Phospholipase a2	3l30	175	182	51	12
82	Glycogen phosphorylase	1l5r	386	292	53	42
83	Liver carboxylesterase 1	1ya4	40	266	177	93
84	Acetylcholinesterase	2whr	1	1	76	1
85	Tyrosine-protein kinase c-src	3g5d	92	51	95	15
86	Progesterone receptor	1e3k	313	419	30	35
87	DNA topoisomerase 1	1seu	279	9	220	118
88	Glutamate carboxypeptidase 2	2or4	66	9	28	16
89	Adenosine kinase	2i6b	152	77	116	88
90	3-Hydroxy-3-methylglutaryl-coenzyme A reductase	1hw9	601	1	230	175
91	Carbonic anhydrase 2	2qp6	119	301	121	122
92	Endothelin-converting enzyme 1	3dwb	237	2	225	181
93	Phosphodiesterase 4	1xoz	34	251	33	6
94	Matrix metalloproteinase	3f17	503	132	20	21
95	Trypanothione reductase	2wpf	62	146	121	65
96	Poly (ADP-ribose) polymerase 1	2pax	99	91	33	5
97	Tyrosine-protein kinase lck	1qpd	67	57	469	158
98	Cyclooxygenase 2	4cox	59	24	324	60
99	Ephrin type-b receptor 4	2vwx	444	567	2	58
100	Metabotropic glutamate receptor 5	3lmk	22	302	181	71
	Average ranking order		143	140	130	67

[a] In each case, two independent target prediction studies were carried out, of which the mean values were used to rank the targets in TargetDB.
[b] Compounds 1–10 were selected from DrugBank database; compounds 11–100 were selected from the BindingDB database.

Table 3
Retrieval of 12 protein targets of 4H-tamoxifen using the Chemscore, Goldscore, Fitness, and *C*-value.[a]

Target name	PDB code	Chemscore rank (%)	Goldscore rank (%)	Fitness rank (%)	C-value rank (%)
17-Beta-hydroxysteroid dehydrogenase	1xf0	17	33	10	3
3-Alpha-hydroxysteroid dehydrogenase	2ipj	186	84	49	34
Alcohol dehydrogenase	2ao0	41	170	126	84
Calmodulin	1lin	499	1012	13	23
Cyclooxygenase-2	4cox	119	68	105	102
Dihydrofolate reductase	1ia1	131	12	122	100
Estrogen receptor alpha	1r5k	5	8	27	1
Estrogen receptor beta	1qkn	14	29	77	13
Glutathione transferase	1gsf	182	115	48	35
Human fibroblast collagenase	3ayk	294	90	314	222
Immunoglobulin	3fo9	20	3	78	7
Protein kinase c	2i0e	401	128	134	158

[a] In each case, two independent target prediction studies were carried out, of which the mean values were used to rank the targets in TargetDB.

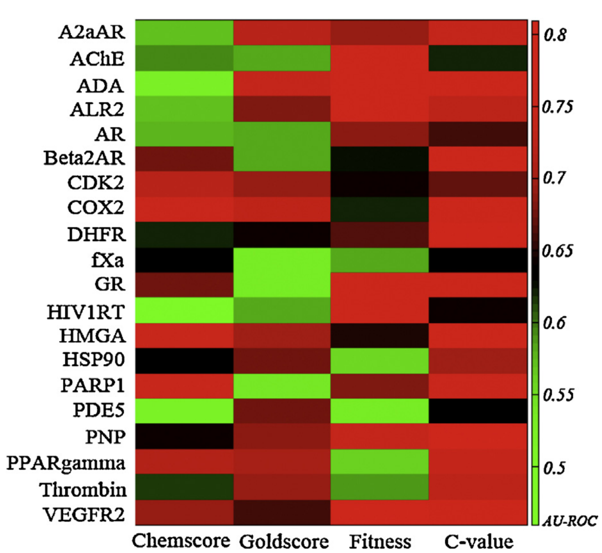

Fig. 3. Heatmap plot showing the performance of Chemscore, Goldscore, Fitness, and *C*-value for each target in the training set.

drugs. The prediction results are shown in Table 2. From Table 2, we can see that the docking-based and pharmacophore-based methods exhibit different performance on the retrieval of true targets for different compounds. The average ranking orders for the real targets of the selected 100 compounds are 143, 140, and 130 by Chemscore, Goldscore, and Fitness, respectively.

Subsequently, the combined strategy was used to predict the true targets for the selected drugs, which results are also shown in Table 2. From Table 2, we can see that, for the most of the selected drugs, the ranking order of its true target was brought forward significantly. The average ranking order by *C*-value for the true targets of the selected drugs is 67, which is much better compared with the results by Chemscore, Goldscore, and Fitness alone. A statistically significant difference exists between any single measure and *C*-value ($p < 0.01$, *t*-test).

3.5. Further evaluation of the combined target prediction method using 4H-tamoxifen

4H-tamoxifen, which is a marketed drug for the treatment of breast cancer clinically, was further used to validate the combined

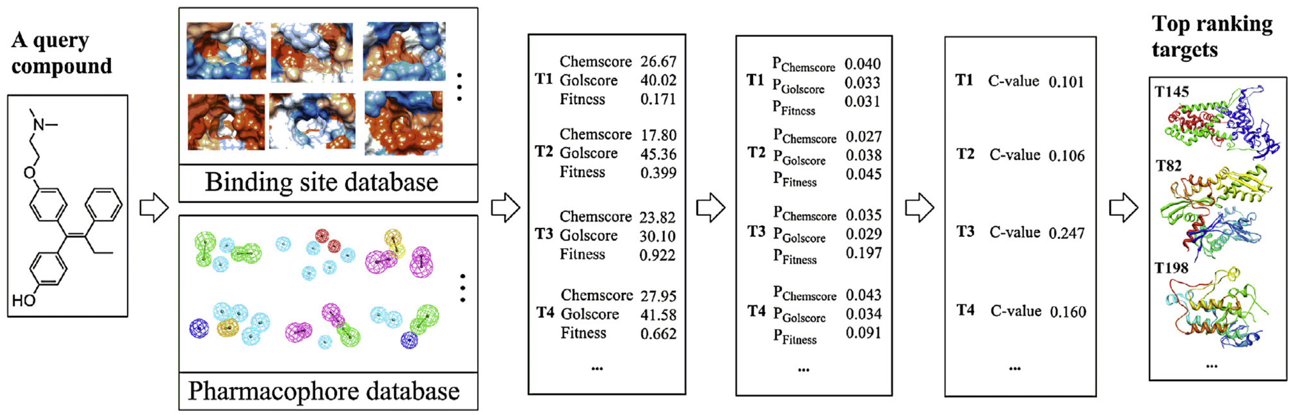

Fig. 4. The workflow for the combined target prediction strategy.

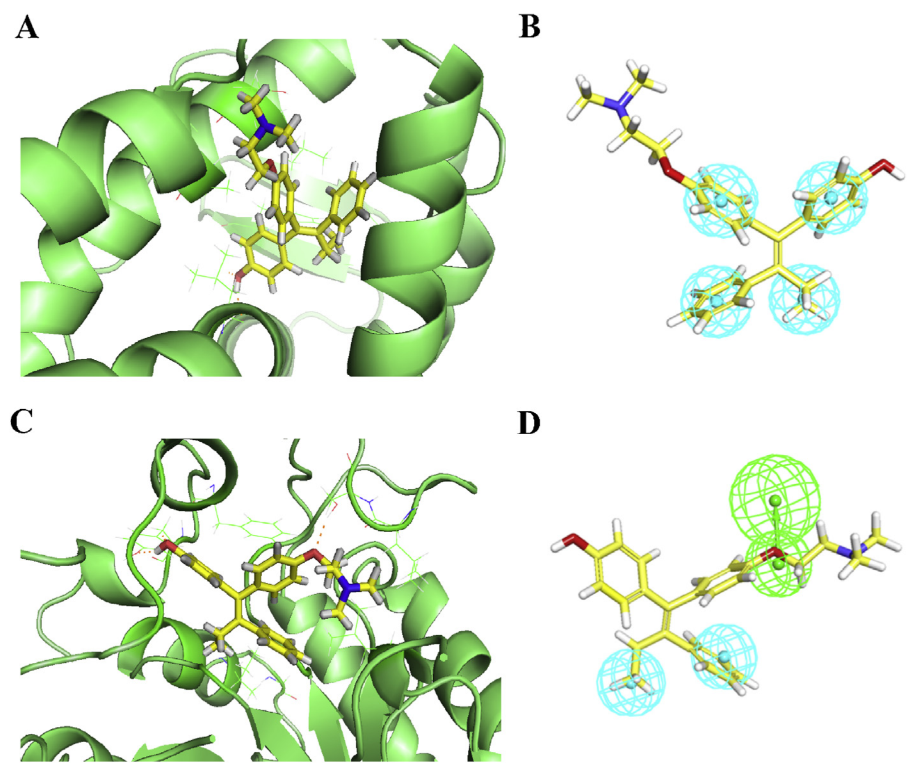

Fig. 5. (A) The binding mode of 4H-tamoxifen within the binding site of estrogen receptor alpha. (B) 4H-tamoxifen mapped with the pharmacophore model corresponding to estrogen receptor alpha. (C) The binding mode of 4H-tamoxifen within the binding site of 17-beta-hydroxysteroid dehydrogenase. (D) 4H-tamoxifen mapped with the pharmacophore model corresponding to 17-beta-hydroxysteroid dehydrogenase.

target prediction method; 4H-tamoxifen was chosen since it is a famous multi-target drug. Up to now, more than 12 proteins have been identified as its targets [12,13,43]. Table 3 shows the predicted results by the docking-based method, the pharmacophore-based method, and the combined method. The detailed target names of the top-ranking targets predicted by the combined method are presented in Supplementary Table S4.

From Table 3, we can see that, according to the *C*-value, 4 targets of the 12 known targets are ranked in the top 1% of the 1481 entries in TargetDB, and 10 targets are in the top 10%. The average ranking for all the targets is 65. Compared with the combined method, the docking-based and pharmacophore-based methods have a relatively worse performance; the average ranking orders for all the targets are 159, 146, and 92 for Chemscore, Goldscore, and Fitness, respectively. Fig. 5 shows the molecular docking and pharmacophore mapping results of 4H-tamoxifen with the top 2 targets (estrogen receptor alpha and 17-beta-hydroxysteroid dehydrogenase) predicted by the combined strategy. From Fig. 5, we can see that 4H-tamoxifen can be perfectly docked into the binding sites of estrogen receptor alpha and 17-beta-hydroxysteroid dehydrogenase; their Chemscore values are 47.37 and 44.13, respectively, and their Goldscore values are 54.85 and 51.04, respectively. 4H-tamoxifen was also mapped very well with the pharmacophore models corresponding to the two targets; the Fitness values are 0.9607 and 0.9852, respectively.

4. Conclusions

We have presented a combined molecular docking-based and pharmacophore-based target prediction strategy, in which a probabilistic fusion method is used for target sorting. Evaluation results against a set of selected drugs whose targets are known and 4H-tamoxifen, which is a multi-target drug, showed that the combined strategy consistently outperformed the sole use of docking-based and pharmacophore-based methods. This work highlights the value of using multiple, complementary measures for target prediction, and the advantages of the probabilistic fusion method, which maximizes the use of different computational measures. It is expected that this type of combined target prediction method could be useful in drug discovery process.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (81172987), 973 project (2013CB967204), and the 863 Hi-Tech Program (2012AA020301, 2012AA020308).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2013.07.005.

References

[1] A.L. Hopkins, Network pharmacology, Nature Biotechnology 25 (2007) 1110–1111.
[2] M. Bantscheff, D. Eberhard, Y. Abraham, S. Bastuck, M. Boesche, S. Hobson, T. Mathieson, J. Perrin, M. Raida, C. Rau, V. Reader, G. Sweetman, A. Bauer, T. Bouwmeester, C. Hopf, U. Kruse, G. Neubauer, N. Ramsden, J. Rick, B. Kuster, G. Drewes, Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors, Nature Biotechnology 25 (2007) 1035–1044.
[3] B.L. Roth, Drugs and valvular heart disease, New England Journal of Medicine 356 (2007) 6–9.
[4] M. Campillos, M. Kuhn, A.C. Gavin, L.J. Jensen, P. Bork, Drug target identification using side-effect similarity, Science 321 (2008) 263–266.
[5] M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, N.H. Jensen, M.B. Kuijer, R.C. Matos, T.B. Tran, R. Whaley, R.A. Glennon, J. Hert, K.L. Thomas, D.D. Edwards, B.K. Shoichet, B.L. Roth, Predicting new molecular targets for known drugs, Nature 462 (2009) 175–181.
[6] U. Rix, G. Superti-Furga, Target profiling of small molecules by chemical proteomics, Nature Chemical Biology 5 (2009) 616–624.
[7] M. Bantscheff, A. Scholten, A.J. Heck, Revealing promiscuous drug-target interactions by chemical proteomics, Drug Discovery Today 14 (2009) 1021–1029.
[8] L. Sleno, A. Emili, Proteomic methods for drug target discovery, Current Opinion in Chemical Biology 12 (2008) 46–54.
[9] M. Schenone, V. Dancik, B.K. Wagner, P.A. Clemons, Target identification and mechanism of action in chemical biology and drug discovery, Nature Chemical Biology 9 (2013) 232–240.
[10] J.L. Jenkins, A. Bender, J.W. Davies, In silico target fishing: predicting biological targets from chemical structure, Drug Discovery Today 3 (2006) 413–421.
[11] Y.Z. Chen, Z.R. Li, C.Y. Ung, Computational method for drug target search and application in drug discovery, Journal of Theoretical and Computational Chemistry 01 (2002) 213–224.
[12] H. Li, Z. Gao, L. Kang, H. Zhang, K. Yang, K. Yu, X. Luo, W. Zhu, K. Chen, J. Shen, X. Wang, H. Jiang, TarFisDock. A web server for identifying drug targets with docking approach, Nucleic Acids Research 34 (2006) W219–W224.
[13] X. Liu, S. Ouyang, B. Yu, Y. Liu, K. Huang, J. Gong, S. Zheng, Z. Li, H. Li, H. Jiang, PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach, Nucleic Acids Research 38 (2010) W609–W614.
[14] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, B.K. Shoichet, Relating protein pharmacology by ligand chemistry, Nature Biotechnology 25 (2007) 197–206.
[15] Nidhi, M. Glick, J.W. Davies, J.L. Jenkins, Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases, Journal of Chemical Information and Modeling 46 (2006) 1124–1133.
[16] F. Mohd Fauzi, A. Koutsoukas, R. Lowe, K. Joshi, T.P. Fan, R.C. Glen, A. Bender, Chemogenomics approaches to rationalizing the mode-of-action of traditional Chinese and Ayurvedic medicines, Journal of Chemical Information and Modeling 53 (2013) 661–673.
[17] A. Zhang, H. Sun, B. Yang, X. Wang, Predicting new molecular targets for rhein using network pharmacology, BMC Systems Biology 6 (2012) 20.
[18] T.I. Oprea, A. Tropsha, J.L. Faulon, M.D. Rintoul, Systems chemical biology, Nature Chemical Biology 3 (2007) 447–450.
[19] A.L. Hopkins, Network pharmacology: the next paradigm in drug discovery, Nature Chemical Biology 4 (2008) 682–690.
[20] D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications, Nature Reviews Drug Discovery 3 (2004) 935–949.
[21] S.Y. Yang, Pharmacophore modeling and applications in drug discovery: challenges and recent advances, Drug Discovery Today 15 (2010) 444–450.
[22] G.B. Li, L.L. Yang, S. Feng, J.P. Zhou, Q. Huang, H.Z. Xie, L.L. Li, S.Y. Yang, Discovery of novel mGluR1 antagonists: a multistep virtual screening approach based on an SVM model and a pharmacophore hypothesis significantly increases the hit rate and enrichment factor, Bioorganic and Medicinal Chemistry Letters 21 (2011) 1736–1740.
[23] J.X. Ren, L.L. Li, R.L. Zheng, H.Z. Xie, Z.X. Cao, S. Feng, Y.L. Pan, X. Chen, Y.Q. Wei, S.Y. Yang, Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model, pharmacophore, and molecular docking, Journal of Chemical Information and Modeling 51 (2011) 1364–1375.
[24] S. Distinto, F. Esposito, J. Kirchmair, M.C. Cardia, M. Gaspari, E. Maccioni, S. Alcaro, P. Markt, G. Wolber, L. Zinzula, E. Tramontano, Identification of HIV-1 reverse transcriptase dual inhibitors by a combined shape-, 2D-fingerprint- and pharmacophore-based virtual screening approach, European Journal of Medical Chemistry 50 (2012) 216–229.
[25] D. Vidovic, S.A. Busby, P.R. Griffin, S.C. Schurer, A combined ligand- and structure-based virtual screening protocol identifies submicromolar PPARgamma partial agonists, ChemMedChem 6 (2011) 94–103.
[26] T. Tuccinardi, S. Taliani, M. Bellandi, F. Da Settimo, E. Da Pozzo, C. Martini, A. Martinelli, A virtual screening study of the 18 kDa translocator protein using pharmacophore models combined with 3D-QSAR studies, ChemMedChem 4 (2009) 1686–1694.
[27] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Annals of Mathematical Statistics 38 (1967) 325–339.
[28] S.W. Muchmore, D.A. Debe, J.T. Metz, S.P. Brown, Y.C. Martin, P.J. Hajduk, Application of belief theory to similarity data fusion for use in analog searching and lead hopping, Journal of Chemical Information and Modeling 48 (2008) 941–948.
[29] S.L. Swann, S.P. Brown, S.W. Muchmore, H. Patel, P. Merta, J. Locklear, P.J. Hajduk, A unified, probabilistic framework for structure- and ligand-based virtual screening, Journal of Medicinal Chemistry 54 (2011) 1223–1232.
[30] F. Zhu, B. Han, P. Kumar, X. Liu, X. Ma, X. Wei, L. Huang, Y. Guo, L. Han, C. Zheng, Y. Chen, Update of TTD: therapeutic target database, Nucleic Acids Research 38 (2010) D787–D791.
[31] Z. Gao, H. Li, H. Zhang, X. Liu, L. Kang, X. Luo, W. Zhu, K. Chen, X. Wang, H. Jiang, PDTD: a web-accessible protein database for drug target identification, BMC Bioinformatics 9 (2008) 104.
[32] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Research 36 (2008) D901–D906.
[33] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Research 28 (2000) 235–242.
[34] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM: a program for macromolecular energy, minimization, and dynamics calculations, Journal of Computational Chemistry 4 (1983) 187–217.
[35] M.L. Verdonk, J.C. Cole, M.J. Hartshorn, C.W. Murray, R.D. Taylor, Improved protein–ligand docking using GOLD, Proteins 52 (2003) 609–623.
[36] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, R.P. Mee, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, Journal of Computer-Aided Molecular Design 11 (1997) 425–445.
[37] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, Journal of Molecular Biology 267 (1997) 727–748.
[38] P. Sprague, Automated chemical hypothesis generation and database searching with Catalyst®, Perspectives in Drug Discovery and Design 3 (1995) 1–20.
[39] Q. Huang, L.L. Li, S.Y. Yang, PhDD: a new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility, Journal of Molecular Graphics and Modelling 28 (2010) 775–787.
[40] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, M.K. Gilson, D.B. Binding, A web-accessible database of experimentally determined protein–ligand binding affinities, Nucleic Acids Research 35 (2007) D198–D201.
[41] N. Huang, B.K. Shoichet, J.J. Irwin, Benchmarking sets for molecular docking, Journal of Medicinal Chemistry 49 (2006) 6789–6801.
[42] J.J. Irwin, B.K. Shoichet, ZINC—a free database of commercially available compounds for virtual screening, Journal of Chemical Information and Modeling 45 (2005) 177–182.
[43] S.L. Kinnings, R.M. Jackson, ReverseScreen3D: a structure-based ligand matching method to identify protein targets, Journal of Chemical Information and Modeling 51 (2011) 624–634.