European Journal of Medicinal Chemistry 59 (2013) 296–303

Contents lists available at SciVerse ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

MolShaCS: A free and open source tool for ligand similarity identification based on Gaussian descriptors

Luis Antônio C. Vaz de Lima [a], Alessandro S. Nascimento [a,b,*]

[a] Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas, Universidade Federal do ABC, R. Santa Adélia, 166, Bangu, Santo Andre, Sao Paulo 09210-170, Brazil
[b] Grupo de Biotecnologia Molecular, Instituto de Fisica de São Carlos, Av. Trabalhador saocarlense, 400, Centro, Sao Carlos, Sao Paulo 13560-970, Brazil

ARTICLE INFO

Article history:
Received 23 July 2012
Received in revised form
7 November 2012
Accepted 8 November 2012
Available online 17 November 2012

Keywords:
Molecular similarity
Gaussian descriptors
MolShaCS

ABSTRACT

Molecular similarity evaluation is an important step in most drug development strategies, since molecular similarity is usually related to functional similarity. Here, we developed a method based on the Gaussian description of molecular shape and charge distribution for molecular similarity identification. The method was evaluated using the Directory of Useful Decoys (DUD) and a retrospective test. Enrichment factors computed for DUD targets showed that the proposed method performs very well in recognizing molecules with similar physicochemical properties and dissimilar topologies, reaching an average AUC of 0.63 and enrichment factor of 10 at 0.5% of decoys. A retrospective test also showed that nine mineralocorticoid ligands were ranked among the top ten molecules in a search of a database of approved drugs for molecules similar to aldosterone. Altogether, these data show that the Gaussian-based description of molecular shape and charge distribution implemented in the program MolShaCS is an efficient method for molecular similarity identification. The program is publicly available at the address http://www.ifsc.usp.br/biotechmol.

1. Introduction

Similar small molecules are usually assumed to produce similar biological activities, taken that they bind similarly to the same biological receptor [1–3]. This assumption makes the correct evaluation of the similarity degree among two molecules an outstanding step in drug development, especially for the 'ligand-based' strategies [4]. Different methods have been proposed for molecular similarity evaluation using two dimensional or three dimensional molecular data. Two-dimensional methods typically use molecular fingerprints that encode structural features in arrays of strings for further comparison. A molecule can thus be described as a graph where the atoms are nodes and the bonds are connections among the nodes. Some implementations of 2D methods can be found in the programs DAYLIGHT FINGERPRINTS [5], MACCS keys [6] and SEA [7]. 3D methods, on the other hand, make use of molecular properties computed in the 3D space for molecular comparison. An example of a 3D descriptor is the molecular electrostatic potential. Some implementations of 3D methods include the programs SHAFTS [8], ELECTROSHAPE [9], SHAEP [10] and ROCS [11–13]. A more comprehensive review of 2D and 3D molecular descriptors can be found in Ref. [14].

3D methods, in principle, should be able to readily recognize similar molecules beyond the molecular topology or bond connectivity. When enough information is available from well chosen structural descriptors, those methods can recognize similarity among bioisosteres, for example [15]. Two chemical groups are said bioisosteres when despite their topological dissimilarities they have similar physicochemical properties and produce similar biological response. Carboxylate and tetrazole are a well known example of bioisosteric groups. Since the 3D descriptors are designed to reproduce physicochemical properties, the methods based on these descriptors should be able to recognize the groups as similar, improving the accuracy of molecular similarity assignment. However, since 3D descriptors, such as electrostatic or van der Waals potential, are usually computed in the Cartesian space, the molecules must be correctly posed in the space prior to comparison computation. The misalignment or errors during the alignment stage have been recognized as the major source of error in 3D-QSAR methods [16,17], where this same alignment problem exists.

The measurement of the degree of similarity among two molecules has been shown to be an outstanding tool in the scope of the off-target interaction identification in recent years [18–20]. In this approach, the prediction of cross-interactions among existing drugs

* Corresponding author. Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas, Universidade Federal do ABC, R. Santa Adélia, 166, Bangu, Santo Andre, Sao Paulo 09210-170, Brazil. Tel.: +55 11 4996 8264; fax: +55 11 4996 0100.
E-mail addresses: asnascimento@if.sc.usp.br, alessandro.nascimento@ufabc.edu.br (A.S. Nascimento).

0223-5234/$ – see front matter © 2012 Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.ejmech.2012.11.013

and drug candidates is becoming increasingly important, especially considering the adverse reactions that these cross-interactions can cause. A cross-interaction can be defined as the ability of a certain molecule designed to recognize a given pharmacological receptor or macromolecule to bind a second target, unintended [21].

Cross-interactions are especially important when drug side effects are originated due to an unintended recognition among the drug and a biological receptor. In the United States, the number of events associated to adverse drug reactions (ADR) reaches 2 million yearly, causing a hundred thousand deaths. The most serious consequences of the ADRs are responsible for 1.8–6.2% of all cases of hospitalization in US and a burden of 136 million dollars annually, more than what is spent in diabetes or cardiovascular diseases management [22].

The number of events associated with ADRs suggests that the chemotherapy paradigm where a drug binds to one receptor and leads to a unique biological effect, is still a goal to be reached, as previously observed [23]. Actually, Keiser and coworkers came to the conclusion that about 35% of the drugs already in the market can interact with at least two pharmacological receptors [7]. The unintended interaction can result in the desired pharmacological response when additive or synergic effects take place, or in undesired adverse effects, when the unintended interaction triggers a harmful reaction to the individual.

Given the relevance of cross-interactions in biological systems and their potential to trigger adverse reactions, it is important to recognize candidate cross-interactions as early as possible in the drug development pipeline, preferably, still in the design stage. In this scope, computational methods that accurately assign candidate cross-interactions are important tools for both academy and industry. Here, we show how a computational tool based on a Gaussian description of molecular shape and charge can be useful for molecular overlay and as a measurement of molecular similarity. An evaluation of the performance of the method using the DUD database shows that it is equally fast and accurate when compared to methods previously proposed. A retrospective examination of a database against mineralocorticoid receptor (MR) ligands also shows that our method can find 9 MR ligands among the 10 top scored molecules, demonstrating its ability to correctly recognize molecules with similar activity using structural data solely.

2. Results and discussion

2.1. Overall performance

In order to be able to probe large databases of small molecules, the methods for molecular similarity evaluation must be fast and accurate. Our method showed a good performance in large scale pairwise similarity comparison, with more than 5 molecules compared per second in a single thread on average on a 2.4 GHz Intel i7 processor running 8 threads simultaneously, with most of the computation time spent in the overlay optimization process. This performance is, in part, due to the fact that the objective function (shown in equation (2)) loops only once over the atoms in both molecules instead of the total number of grid points, as in the grid-based methods, for example. Then, for small molecules, the performance encourages virtual screening strategies for similarity identifications.

Different algorithms for molecular overlay optimization were also evaluated within MolShaCS. The current version of the program has seven different optimization algorithms implemented. Using the derivative-free Augmented Lagrangian algorithm [24–26], we were able to speed up the computation to compare over 21 molecules per second (about 46 ms per molecule) in the COX-2 target. Vainio and coworkers reported an average time per molecule of 135 ms for this target in ShaEP and 22 ms for ROCS [10]. These results show that MolShaCS can be very efficient for similarity evaluation, compared to previously reported methods.

Regarding molecular overlay, Fig. 1 shows an example of the performance of our method. The figure shows the starting conformation (Fig. 1A) and the aligned conformation (Fig. 1B) of the 30 top ranked molecules of the mineralocorticoid receptor retrospective test. The starting conformations have different degrees of difficulty to reach a good molecular overlay, but the good matching in the steroid scaffold for those structures shows that our method performs fairly well in molecular alignment, which is an outstanding prerequisite for a 3D based structure comparison.

2.2. DUD enrichment test

We then assessed the similarity evaluation of the method using a well established benchmark for comparison of small molecules. The Directory of Useful Decoys (DUD) dataset, although conceived originally as a benchmark for a structure-based method (docking), has been proven very useful for ligand-based methods as well [8–10,27]. As previously noted, DUD release 2 has some inaccuracies in charge computations [9]. Armstrong and coworkers noted in their analysis of charge spreading that the differences in charges among decoys are much larger than what would be expected, making the partial charges of DUD decoys useless for our purposes [9]. For this reason, we recomputed partial charges for the entire DUD database using AM1 charge method, as implemented in the program ANTECHAMBER [28]. Some molecules that did not converge in charge fitting were removed from the database. Table 1 shows the final

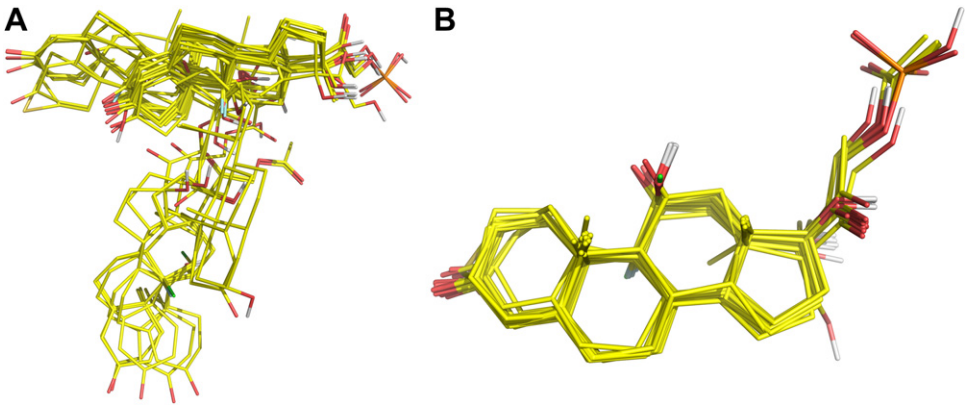

Fig. 1. Starting structures (A) and aligned structures (B) of the top 30 molecules from the mineralocorticoid retrospective test.

Table 1
Targets used in DUD enrichment test.

DUD target	Number of actives	Number of decoys	Average AUC	AUC*	Average $EF_{0.5\%}$	$EF^{*}_{0.5\%}$	Average $EF_{1\%}$	$EF^{*}_{1\%}$	Average $ROCE_{0.5\%}$	$ROCE^{*}_{0.5\%}$	Average $ROCE_{1\%}$	$ROCE^{*}_{1\%}$
ace	49	1797	0.54 (±0.110)	0.66	12.02	25.15	8.72	19.38	21.98	73.46	12.95	38.77
ache	107	3891	0.60 (±0.080)	0.698	11.82	19.18	8.09	13.89	18.09	38.47	10.55	21.49
ada	39	927	0.685 (±0.069)	0.751	11.88	17.06	8.95	13.65	25.3	53.21	14.63	29.07
alr2	26	995	0.557 (±0.131)	0.699	13.06	21.47	8.85	14.76	20.99	46.14	11.83	23.07
ampc	21	786	0.655 (±0.184)	0.768	17.64	24.47	12.77	19.23	38.54	66.66	21.55	37.71
ar	79	2853	0.789 (±0.077)	0.820	14.42	20.91	11.23	15.61	25.27	46.49	16.62	26.32
cdk2	72	2073	0.596 (±0.088)	0.629	9.59	19.19	6.67	14.56	16.14	52.08	9.18	27.49
comt	11	468	0.507 (±0.144)	0.692	25.12	31.18	16.56	23.82	66.9	108.21	28.06	51.45
cox1	25	911	0.568 (±0.066)	0.505	8.99	16.64	6.2	11.52	12.44	29.63	7.67	16.33
cox2	421	13,288	0.789 (±0.107)	0.868	16.52	24.13	13.42	19.86	48.17	93.42	27.83	50
dhfr	390	8364	0.582 (±0.052)	0.658	8.66	12.96	6.51	10.65	15.21	30.68	9.41	18.06
egfr	465	15,993	0.593 (±0.061)	0.663	9.47	19.7	6.86	14.59	14.87	44.2	9.21	24.42
er_agonist	67	2567	0.723 (±0.075)	0.786	15.88	26.88	11.97	20.75	32.3	83.58	19.14	42.85
er_antagonist	39	1448	0.614 (±0.050)	0.629	12.49	20.35	8.57	13.27	19.85	42.73	11.36	19.91
fgfr1	120	4550	0.560 (±0.055)	0.589	11.9	21.09	8.17	15.42	18.43	45	10.59	24.75
fxa	145	5736	0.634 (±0.118)	0.782	14.48	28.1	10.76	21.78	35.15	89.64	18.93	45.74
gart	40	839	0.605 (±0.081)	0.615	11.52	16.12	8.62	12.72	28.59	58.51	15.59	28.95
gpb	50	2136	0.583 (±0.140)	0.713	16.56	28.78	11.77	22.41	31.62	85.35	17.93	45.1
gr	78	2947	0.657 (±0.062)	0.750	13.8	22.7	9.61	17.68	23.49	53.84	13.32	31.73
hivpr	62	2038	0.710 (±0.088)	0.779	11.27	18.49	7.88	16.1	18.78	39.49	10.96	31.26
hivrt	43	1519	0.531 (±0.142)	0.719	8.92	16.96	6.35	16.16	12.98	31.29	8.13	28.47
hmga	35	1480	0.695 (±0.082)	0.824	15.21	26.45	10.34	20.09	26.70	66.85	14.91	36.77
hsp90	37	978	0.807 (±0.08)	0.919	16.16	21.98	12.85	18.89	56.85	105.98	30.14	58.28
inha	80	3266	0.575 (±0.068)	0.513	11.25	20.28	7.73	14.22	16.76	38.46	9.83	21.04
mr	15	636	0.821 (±0.063)	0.858	19.94	24.84	14.42	19.76	39.71	56.72	22.69	35.46
na	49	1874	0.643 (±0.104)	0.686	14.18	20.72	9.57	13.57	23.69	42.5	13.06	20.2
p38	453	9141	0.605 (±0.082)	0.710	11.13	16.2	8.72	13.39	32.23	66.96	17.32	35.02
parp	35	1349	0.646 (±0.136)	0.632	18.41	26.45	13.38	20.52	41.77	78.43	22.9	41.6
pde5	88	1978	0.622 (±0.091)	0.506	9.44	16.61	6.86	12.83	19.26	54.54	10.67	27
pdgfrb	170	5979	0.583 (±0.066)	0.656	12.73	25.25	9.29	19.76	24.67	81.16	14.48	42.35
pnp	50	1036	0.651 (±0.122)	0.771	12.45	17.71	9.6	14.94	32.74	91.67	18.11	45.83
ppar_gamma	85	3126	0.561 (±0.099)	0.678	16.43	23.78	11.59	17.63	32.23	62.27	17.56	32.08
pr	27	1041	0.721 (±0.100)	0.75	21.65	30.23	17.1	24.42	75.43	123.43	39	61.72
rxr_alpha	20	746	0.7 (±0.068)	0.783	19.82	26.52	14.68	21.28	48.58	94.34	26.17	46.73
sahh	33	1345	0.786 (±0.066)	0.769	17.07	24.61	12.92	19.31	30.37	58.27	18.89	34.96
src	159	6319	0.541 (±0.055)	0.665	10.17	21.61	6.84	16.63	14.67	45.28	8.51	27.4
thrombin	71	2455	0.564 (±0.109)	0.739	12.16	24.82	8.76	19.98	23.42	81.25	13.55	44.62
tk	22	891	0.564 (±0.078)	0.616	12.78	20.29	8.3	14.04	18.75	35.71	10.4	19.8
trypsin	49	1663	0.589 (±0.118)	0.713	12.07	19.42	8.03	13.73	20.1	42.5	10.83	22
vegfr2	88	2905	0.517 (±0.072)	0.563	7.54	18.55	5.4	14.28	11.23	40.1	7.08	24.1
Mean	35.3 ± 6.95 decoys/ligand		0.63 ± 0.08	0.70 ± 0.10	13.67 ± 3.85	21.95 ± 4.29	9.87 ± 2.91	16.93 ± 3.54	28.36 ± 14.56	61.96 ± 23.94	15.79 ± 7.10	33.50 ± 11.49

number of actives and decoys for each of the 40 DUD targets. On average, there are about 35 decoys for each active in the dataset. The molecules filtered out from the working dataset are listed in Supplementary material.

Since our method computes pairwise similarities, for each target in DUD, calculations were performed using each of the active molecules against all actives and decoy molecules. The similarity indexes were used to rank the final results and compute ROC curves and enrichment factors. Comparing the area under curve (AUC) values for ROC plots, the methods implemented in MolShaCS showed a good performance in terms of enrichment (Table 1 and Fig. 2). For each target, $N_{ligands}$ simulations were performed and an average AUC was computed for each target. Averaging over all 40 DUD targets, we found an average AUC of 0.63 ± 0.08 (Table 1), very similar to the values already reported for well established methods implemented in SHAEP, ROCS [12,13] and ElectroShape [9,10]. In 19 out of the 40 targets, MolShaCS reached better average AUC values than SHAEP [10]. Compared to ROCS, MolShaCS reached better average AUC values in 16 targets (Supplementary material).

Liu and coworkers compared the performance of some methods for ligand comparison using $ROCE_{x\%}$. $ROCE_{x\%}$ values are defined as the ratio of active rate and decoy rate at some x value of decoy rate [8]. Using a single 'reference' molecule for each DUD target, corresponding to the crystallographic pose of a reference structure and on the complex of all DUD targets, they found average $ROCE_{0.5\%}$ values of 65, 53, 42 and 15 for SHAFTS, ROCS, ShaEP and PharmMapper, respectively [8]. Using the selected (top-ranked) ligands, we found a $ROCE_{0.5\%}$ of 62 using MolShaCS, showing that the methods implemented in MolShaCS are comparable in reliability with standard methods used for molecular comparison. It is worthy to note that ROCE takes the ratios of actives and decoys instead of their counts, facilitating the comparison among different targets, since it is independent of the target dataset size. The meaning of the ROCE values found here is that, on average, MolShaCS ranks 62 times more ligands than decoys among the top ranked 0.5% of the decoys and 33 times more ligands at 1% of the decoys (Table 1). If we compare the averages instead of the selected molecules, MolShaCS still ranks 28 and 16 times more ligands than decoys among the top ranked molecules (0.5% and 1.0% of the decoys, respectively), as shown in Table 1.

Armstrong and coworkers evaluated the average enrichment factors for the DUD targets using three different methods: Ultrafast Shape Recognition (USR), Chiral Shape Recognition (CSR) and ElectroShape. They found an enrichment factor at 1% of the decoys ($EF_{1\%}$) of 7.4 and 7.7 for USR and CSR, respectively, and 10.8 for ElectroShape using AM1 charge method [9,29]. Here, we found an average $EF_{1\%}$ of 9.9 for MolShaCS, comparable to the previously reported methods. An $EF_{1\%}$ of 10 means that 10 times more ligands were found than would be expected upon random picking at 1% of decoys. At 0.5% of the decoys, we found with MolShaCS an $EF_{0.5\%}$ of 13.7 on average for the 40 DUD targets (Table 1).

Fig. 2 shows the ROC plots for the 40 DUD targets, where some outstanding early enrichments can be visualized. The targets PR, COMT, RXRα and MR are the top scored targets when the average $EF_{1\%}$ values are compared across the DUD data set. Surprisingly, three of the four top-scored targets are nuclear receptors. A possible explanation to the good performance within the nuclear receptors is the buried binding pocket with well defined polar and nonpolar regions, that imposes some polarity distribution among the ligands caught by the computational method.

We selected some targets for a deeper examination. COMT showed the best results in terms of enrichment factors, while HSP90 is very well ranked in AUC. One could consider the high degree of topology similarity as an explanation for the good recognition of actives in the targets. We evaluated this feature using OpenBabel [30] to measure the Tanimoto coefficients among actives in COMT and HSP90. On average, COMT actives show a Tanimoto coefficient of 0.387, while HSP90 show a coefficient of 0.491. These indices are not high enough to explain the good EFs found in these targets. On the other hand, VEGFR2 was the worst target in terms of $EF_{0.5\%}$, $ROCE_{0.5\%}$ and the second worst in terms of AUC. We also compared the Tanimoto coefficients for the VEGFR2 ligands and found an average Tanimoto coefficient of 0.338. Although this average coefficient is smaller than those observed for COMT and HSP90, the difference is small and the performance may not be due to the fingerprint similarity but rather to the structural similarity. Altogether, the results shown here indicate that the physicochemical properties, rather than molecular topology, are being evaluated in our method.

The weights for shape and charge are also important parameters in molecular overlay and similarity computation. For this work, we set these parameters to 1.0. However, increasing the charge distribution weight tends to improve the accuracy in similarity recognition for ligands binding to polar sites. In order to test the effect of shape and charge weights on DUD AUC values, we scanned the weight parameters for the targets ACE and MR. ACE (angiotensin-converting enzyme) is an enzyme with a known polar binding pocket and that binds polar and charged ligands. MR (mineralocorticoid receptor), on the other hand, is a steroid receptor with a buried and hydrophobic pocket. Increasing the charge weight (w_2 in equation (2)) resulted in an increase of the average AUC value for ACE from 0.54 ($w_2 = 1.0$) to 0.61 ($w_2 = 20.0$), as shown in Table 2. Increasing the shape weight (w_1 in equation (2)) did not change the average AUC values (0.54 for $w_1 = 2.0$ and 0.53 for $w_1 = 20.0$). For the hydrophobic MR ligands, the reciprocal behavior was observed. Increasing the weight for charge distribution leads to a decrease in the average AUC value from 0.82 ($w_2 = 1.0$) to 0.73 ($w_2 = 20.0$), as shown in Table 3. Again, increasing the weight for shape did not change substantially the average AUC value. Since the term for shape overlap is the dominant term in the objective function shown in equation (2), an increase in w_1 is not expected to cause a significant change. However, the results shown for these two targets reveal that the charge distribution weight parameter can be fine-tuned in a context dependent manner to improve the results obtained by the program.

Molecular flexibility is another important parameter for structural comparisons algorithms. To account for the molecular flexibility effect, we used FROG2 [31] to generate ensembles of conformations of DUD ligands and decoys for three of the DUD targets. We chose the targets ACE, MR and FGFR1, as examples of average ranked, well ranked and badly ranked target among the 40 targets in DUD database. The conformations generated with FROG2 were used directly in MolShaCS for molecular comparison. Again ROC curves were computed for each active, choosing the best active or decoy conformer as the molecule representative. Compared to original conformers, the flexible approach showed slightly reduced average enrichment factors for ACE and MR (10.7 and 16.03, compared to 12.02 and 19.94, respectively), reflecting the fact that some decoys found better ranked conformers. However, for FGFR1 the opposite scenario was found. The average enrichment factor increased from 11.9 to 13.09. These results show that molecular flexibility plays an important role, as expected, in this three dimensional method for molecular alignment, but also show that accurate results still can be reached using a fast and simple rigid approach. It should be taken into account that DUD molecules are energy minimized already, and reflect low energy, good starting conformations.

2.3. Mineralocorticoid receptor retrospective study

Given the encouraging results in the virtual screening enrichment tests shown in the section above, we decided to study a real

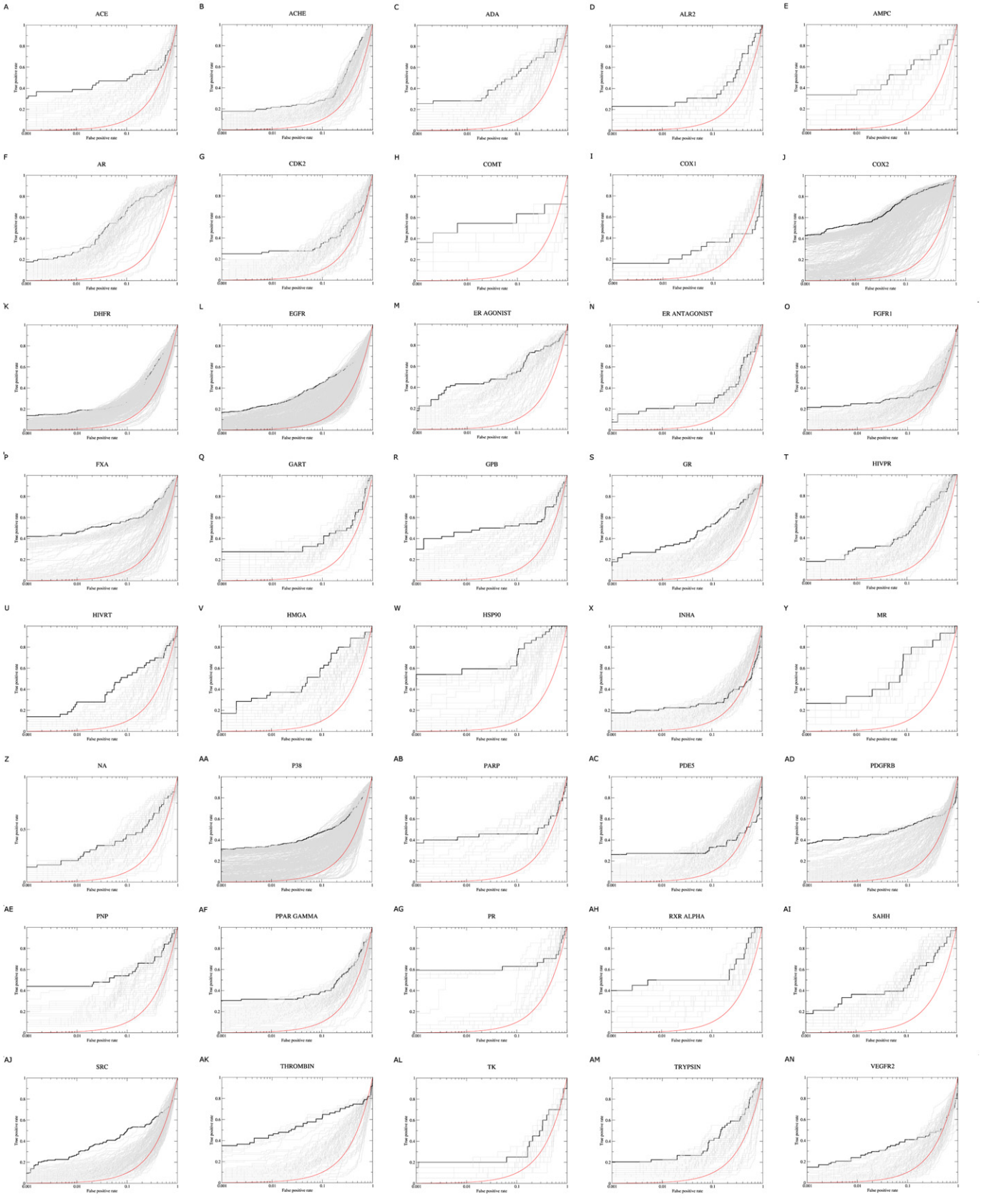

Fig. 2. Enrichment plots. Bold lines show selected runs for which AUC and enrichment factors are computed and listed in Table 1. Red lines show the pattern expected for random assignment. Plots are shown for ACE (A), ACHE (B), ADA (C), ALR2 (D), AMPC (E), AR (F), CDK2 (G), COMT (H), COX1 (I), COX2 (J), DHFR (K), EGFR (L), ER AGONIST (M), ER AGONIST (N), FGFR1 (O), FXA (P), GART (Q), GPB (R), GR (S), HIVPR (T), HIVRT (U), HMGA (V), HSP90 (W), INHA (X), MR (Y), NA (Z), P38 (AA), PARP (AB), PDE5 (AC), PDGFRB (AD), PNP (AE), PPAR GAMMA (AF), PR (AG), RXR ALPHA (AH), SAHH (AI), SRC (AJ), THROMBIN (AK), TK (ALI), TRYPSIN (AM), VEGFR2 (AN). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Effect of weights on AUC values for target ACE. Curiously, the increase in the average AUC value does not reflect an increase in the enrichment factors, or in the ROCE values (not shown). This means that the shape match is probably the most important feature for the molecules with high similarity.

w_1 (shape weight)	w_2 (charge weight)	Average AUC	Standard deviation	$EF_{0.5\%}$	Standard deviation
1.0	2.0	0.555	0.106	12.36	5.58
1.0	3.0	0.561	0.108	11.71	5.11
1.0	4.0	0.569	0.104	11.25	4.59
1.0	5.0	0.578	0.100	11.01	4.06
1.0	10.0	0.597	0.090	10.02	3.74
1.0	20.0	0.615	0.077	10.04	3.19
1.0	30.0	0.613	0.085	10.04	3.87
1.0	40.0	0.621	0.074	10.06	3.27
1.0	50.0	0.625	0.076	9.98	3.05
2.0	1.0	0.537	0.118	12.06	6.96
3.0	1.0	0.531	0.116	12.35	7.05
4.0	1.0	0.528	0.116	12.08	7.04
5.0	1.0	0.528	0.113	12.47	7.22
10.0	1.0	0.531	0.118	12.07	7.23
20.0	1.0	0.532	0.118	12.20	7.31
30.0	1.0	0.530	0.113	12.27	7.09
40.0	1.0	0.535	0.119	12.25	7.21
50.0	1.0	0.531	0.121	12.18	7.17

case of ligand identification in a database of approved compounds. We chose the mineralocorticoid receptor (MR) as the test set. This receptor was chosen because of the good enrichment results in the section above and because it is somewhat familiar to us. MR is well known for binding both glucocorticoids and mineralocorticoids with high affinity [32,33]. Since this receptor strongly regulates sodium retention, some important issues for patients under glucocorticoid therapy (for chronic inflammatory diseases, for example) are water retention and increase in blood pressure. This polypharmacology of glucocorticoids and mineralocorticoids makes MR an interesting target for a retrospective test in the scope of similarity evaluation.

Table 3
Effect of weights on AUC values for target MR.

w_1 (shape weight)	w_2 (charge weight)	Average AUC	Standard deviation	$EF_{0.5\%}$	Standard deviation
1.0	2.0	0.813	0.061	19.34	5.64
1.0	3.0	0.825	0.054	19.76	5.92
1.0	4.0	0.806	0.070	20.19	5.44
1.0	5.0	0.800	0.072	18.49	6.15
1.0	10.0	0.789	0.094	20.37	5.98
1.0	20.0	0.734	0.107	19.65	6.01
1.0	30.0	0.698	0.114	18.48	5.53
1.0	40.0	0.701	0.112	18.97	5.76
1.0	50.0	0.713	0.104	18.49	7.10
2.0	1.0	0.813	0.068	20.43	6.50
3.0	1.0	0.820	0.057	20.86	6.02
4.0	1.0	0.813	0.075	20.49	6.14
5.0	1.0	0.812	0.071	20.32	6.03
10.0	1.0	0.824	0.064	20.95	5.86
20.0	1.0	0.819	0.052	20.19	6.41
30.0	1.0	0.825	0.061	21.51	6.14
40.0	1.0	0.818	0.067	21.38	5.27
50.0	1.0	0.822	0.056	19.98	5.31

Aldosterone, MR's natural ligand, was chosen as the reference compound. In this section we used AMSOL charges, as available in ZINC molecules [34]. The drug database, a subset of the ZINC containing 6869 purchasable compounds worldwide approved, was used for similarity identification. These molecules were obtained from ZINC website as mol2 files with AMSOL charges and used with no prior modifications.

The entire calculation took 14 min, i.e., 8 molecules processed per second on average, and the results were sorted by similarity index. The top 10 molecules (0.14% of the database) were selected and analyzed for their known properties. As shown in Table 4, at least 7 of the top ranked ten molecules are actual MR agonists. 11-Deoxycorticosterone (DOC) appears in the top of the list and again in the bottom of the list (10th molecule in the ranked list) in the acetate form. DOC is a potent MR agonist known for its mineralocorticoid side effects, such as hypertension, cardiac hypertrophy and cardiac fibrosis [35]. Three synthetic glucocorticoids with known mineralocorticoid activity also appear in the ranked list. Fludrocortisone, prednisolone and triamcinolone are MR agonists with important activities [36]. From these, fludrocortisone was showed to be 10 times more potent than aldosterone in CV1 cells [33]. Two other glucocorticoids, cortisone and prednisone, appear in the list, but are inactive on MR. Cortisone seems to bind in the receptor binding site but is not able to elicit a response. The effects of cortisone are very well known in MR_{S810L} mutants, where it acts as an agonist [37]. Both cortisone and prednisone are sometimes used as competitive MR inhibitors. 11-Deoxycortisol also appears in the list. It is a cortisol precursor known for its MR activity [38]. Finally, Table 2 lists cortisol, a physiological MR agonist together with aldosterone and diflucortolone. No mineralocorticoid activity has been reported for the latter, as far as these authors are aware.

Using ShaEP for the same study, similar results were obtained (Supplementary material). Five of the 10 molecules found by MolShaCS were also identified by ShaEP among the top 0.14% of the database (deoxycorticosterone, cortisol, fludrocortisone, 11-deoxycortisol and diflucortolone). ShaEP also identified five molecules not listed among the top scored molecules in MolShaCS: corticosterone, fluocortolone, desoxymetasone, progesterone and isoflupredone. Corticosterone is a known MR antagonist [39]. Isoflupredone is a strong MR agonist that activates CV-1 cells 7 times more potently than aldosterone [33]. Progesterone has negligible activity on MR (less than 1% of aldosterone activity on

Table 4
Results of MolShaCS run using aldosterone as the reference molecule against ZINC drug database.

Rank	ZINC ID	Similarity	Common name	Biological function on MR	Reference
Query	ZINC03833824	1.0	Aldosterone	MR agonist	Arriza et al., 1987 [32]
1	ZINC03833823	0.99037	Deoxycorticosterone	MR agonist	Rickard et al., 2006 [35]
2	ZINC04097304	0.98664	Fludrocortisone	MR agonist	Druce et al., 2008 [36]
3	ZINC03833821	0.98058	Prednisolone	MR agonist	Druce et al., 2008 [36]
4	ZINC03882036	0.97424	Triamcinolone	MR agonist	Grossmann et al., 2004 [33]
5	ZINC13540519	0.97300	Cortisol	MR natural agonist	Arriza et al., 1987 [32]
6	ZINC03875357	0.967232	Prednisone	Inactive	Grossmann et al., 2004 [33]
7	ZINC03860956	0.96867	11-Deoxycortisol	MR agonist	Bureik et al., 2005 [38]
8	ZINC04212939	0.96767	Diflucortolone		
9	ZINC04083557	0.96688	Cortisone	Inactive	Grossmann et al., 2004 [33]
10	ZINC04428527	0.95928	Deoxycorticosterone acetate	MR agonist	Rickard et al., 2006 [35]

CV1 cells [33]). Fluocortolone is a pure synthetic glucocorticoid with no mineralocorticoid activity [40]. Desoxymetasone is also a glucocorticoid with mineralocorticoid activity to be established. ShaEP and MolShaCS were very efficient in the identification of steroid ligands, although ShaEP ranked among the top-scored molecules a number of glucocorticoids ligands with no mineralocorticoid activity.

Taken together the results of this retrospective study show that 70% of the top ranked list of molecules are MR agonists and 90% are MR ligands, considering cortisone and prednisone as competitive binders. This result highlights the efficiency of MolShaCS and its Gaussian-based methods for virtual screening with different strategies, from cross-interaction studies to new ligands prospective study.

3. Conclusion

Computational tools are now an integral part of the drug development process in academia and industry contexts. Among the different strategies pursued, similarity identification has an important role, since most of times, structural similarity leads to functional similarity. Here we described a method for molecular overlay and similarity identification using molecular shape and charge distribution in a Gaussian-Based function. The method was shown very effective in the DUD enrichment test and performed very well when compared to similar methods currently available. A retrospective study with mineralocorticoid receptor showed that our method was able to recover at least nine MR ligands among the top ranked 10 molecules using aldosterone as the reference molecule. Taken together, these results show that the Gaussian-Based 3D descriptors used here are useful for molecular structural and functional similarity identification.

Besides its use as a virtual screening tool, MolShaCS can be used for the identification of cross interactions in biological systems or the 'off-target' interactions. Keiser and coworkers used the 2D similarity method SEA [7] to relate ligands to their targets and identified a number of interesting off-target interactions among the molecules studied. The efficiency of MolShaCS compared to 2D methods makes its use encouraging for cross-interactions initiatives. As an evidence of this capability, we used the angiotensin-converting enzyme (ACE) inhibitor enalapril as a reference molecule and the Drugbank set of approved drugs, as available in ZINC, and found Aprindine as the highest scored molecule after enalapril itself (Supplementary material). This anti-arrhythmic drug was previously described to increase the specificity for ACE synthetic substrates [41].

Our software, MolShaCS, is available free of charge at the web address http://www.ifsc.usp.br/biotechmol.

4. Experimental

4.1. Shape and charge modeling and overlay optimization

The molecules used in this study were parametrized with atom types from the General Amber Force Field (GAFF) [42]. Atomic charges were computed with AM1 method, as implemented in the program ANTECHAMBER [28], except where indicated. Atomic charges were recomputed for the entire set of DUD actives and decoys avoiding the problems in charge assignment previously observed in the database [9].

Molecular shape was described as a Gaussian function:

$$\rho(\mathbf{r}) = p_i \exp\left\{\left[-\pi\left(\frac{3p_i}{4\pi\sigma_i^3}\right)^{\frac{2}{3}}\right](\mathbf{r}-r_i)^2\right\} \quad (1)$$

for the ith atom, with Gaussian amplitude p_i defined as $2\sqrt{2}$ and, r_i as the atomic coordinate for the ith atom and σ_i defined as van der Waals radius [10,12,13].

Gaussian-based descriptors, as implemented in the program ROCS [11–13], have been used for over 20 years with good performance in terms of pose prediction and similarity identification [8,9,20,21]. In MolShaCS, however, we implemented an empirical objective function where the distribution of charges in the molecule is taken into account within the molecular overlay optimization. The program ShaEP used a similar strategy [6], but here, we modeled the charge distribution using a Gaussian description as well. The advantage of Gaussians for modeling is that the objective function tends to be approximately quadratic, favoring convergence of optimization algorithms besides the fact that shape and charge descriptors can be computed in the same loop over the atoms in the system, as shown below.

In MolShaCS, the polar or charged atoms are not just 'colored', but quantitatively modeled using the same Gaussian descriptors as used for molecular shape modeling. The molecular charge distribution $\phi(\mathbf{r})$ is modeled according to a Gaussian function similar to equation (1), but with σ_i taken as a function of the atomic charge. Overlay optimization was then computed as the sum of the overlay of individual atoms shape and charge:

$$V_{AB} = w_1 \sum_i \sum_j \int \mathrm{d}r \rho_i(r)\rho_j(r) + w_2 \sum_i \sum_j \int \mathrm{d}r \varphi_i^{\mathrm{pos}}(r)\varphi_j^{\mathrm{pos}}(r) + w_2 \sum_i \sum_j \int \mathrm{d}r \varphi_i^{\mathrm{neg}}(r)\varphi_j^{\mathrm{neg}}(r) \quad (2)$$

where ρ_i and ρ_j are the atomic Gaussian description for volume and ϕ_i and ϕ_j are the Gaussian description for charge distribution, computed separately for positive and negative atomic charges. w_1 and w_2 are weights for shape and charge terms, respectively. In this work, w_1 and w_2 were both set to 1.0.

Similarity indexes were computed as a Hodgkin's index, given by [43,44]:

$$\mathrm{SI}_{AB} = \frac{2V_{AB}}{V_{AA}+V_{BB}} \quad (3)$$

The overlay optimization was performed by computer optimization of equation (2), taking one molecule as a 'reference' molecule and the second molecule as a 'searching' molecule, able to rotate and translate in the Cartesian space as a rigid body. Optimization was achieved by using the method of moving asymptotes [45] algorithm as implemented in NLOPT library [46]. All these methods were implemented in a C/C++ program with a Qt graphical interface called *MolShaCS*, available at http://www.ifsc.usp.br/biotechmol. The graphical interface allows the user to select the 'query' molecule and the 'reference' molecules from files, and set all the necessary parameters (optimization tolerance, weights, choice of minimization algorithm, etc) in a user-friendly environment (Supplementary material).

4.2. DUD enrichment test

For the enrichment tests, we chose the 40 targets from the Directory of Useful Decoys (DUD) database [47]. A list of the targets used is shown in Table 1 together with the respective number of actives and decoys. For each target, coordinate files for actives and decoys were obtained from DUD website (http://dud.docking.org) and, after GAFF atom type assignment and refitting of atomic charges, pairwise similarity calculation was performed with MolShaCS. Molecules that did not succeed, for any reason, in atom type assignment or in charge refitting were filtered out. Each active was

compared against the list of actives and decoys for that target using an optimization relative tolerance of 0.01, i.e., an optimum point is found if $|\Delta f_i|/|\Delta x_i| < 0.01$, where f is the objective function and x is one of the six parameters of the function. The final results of each run were sorted according to the similarity indexes computed and receiver operating characteristic (ROC) data was generated and plotted using (selected decoys rate) *versus* (selected actives rate) for fractions of the set for each target. A different plot was generated for each active and for each target. Enrichment factors (EF) were computed as [47] $EF_{Subset} = (ligands_{selected}/N_{subset})/(ligands_{total}/N_{total})$. For the sake of comparison, we also computed ROC enrichment data (ROCE), taken as the ratio of active rate to the decoy rate for a particular subset of decoys [8].

4.3. Retrospective test: human mineralocorticoid receptor

As a retrospective test, we chose the human mineralocorticoid receptor (MR) as a target. Here, we selected MR natural ligand aldosterone as the reference molecule and the Zinc Drug Database (ZDD) from ZINC as the comparing molecules [34,48]. For these molecules atomic charges were used as provided by standard ZINC protocol together with GAFF atomic radii.

Acknowledgements

LACVL thanks the ABC Federal University for his PDPD stipend. ASN thank CNPq and FAPESP for the financial support (grants 476606/2010-1 and 2010/15376-8, respectively).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejmech.2012.11.013.

References

[1] A. Bender, R.C. Glen, Molecular similarity: a key technique in molecular informatics, Org. Biomol. Chem. 2 (2004) 3204–3218.
[2] D.E. Patterson, R.D. Cramer, A.M. Ferguson, R.D. Clark, L.E. Weinberger, Neighborhood behavior: a useful concept for validation of 'molecular diversity' descriptors, J. Med. Chem. 39 (1996) 3049–3059.
[3] H. Eckert, J. Bojorath, Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches, Drug Discov. Today 12 (2007) 225–233.
[4] A.J. McMahon, P.M. King, Optimization of Carbo molecular similarity index using gradient methods, J. Comput. Chem. 18 (1997) 151–158.
[5] Daylight Fingerprints, Version 4.62, Daylight Chemical Information Systems, Laguna Niguel, CA, 1999.
[6] MACCS Structural Keys, Symyx Software, San Ramon, CA, 2010.
[7] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, B.K. Shoichet, Relating protein pharmacology by ligand chemistry, Nat. Biotechnol. 25 (2007) 197–206.
[8] X. Liu, H. Jiang, H. Li, SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening, J. Chem. Inf. Model. 51 (2011) 2372–2385.
[9] M.S. Armstrong, G.M. Morris, P.W. Finn, R. Sharma, L. Moretti, R.I. Cooper, W.G. Richards, ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics, J. Comput. Aided Mol. Des. 24 (2010) 789–801.
[10] M.J. Vainio, J.S. Puranen, M.S. Johnson, ShaEP: molecular overlay based on shape and electrostatic potential, J. Chem. Inf. Model. 49 (2009) 492–502.
[11] ROCS, OpenEye Scientific Software, Santa Fe, NM, 2012.
[12] J.A. Grant, M.A. Gallardo, B.T. Pickup, A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape, J. Comput. Chem. 17 (1996) 1653–1666.
[13] J.A. Grant, B.T. Pickup, A Gaussian description of molecular shape, J. Phys. Chem. 99 (1995) 3503–3510.
[14] I. Muegge, I. Enyedy, Virtual screening, in: Burger's Medicinal Chemistry and Drug Discovery, John Wiley & Sons, Inc., 2003.
[15] P.W. Kenny, Hydrogen bonding, electrostatic potential, and molecular design, J. Chem. Inf. Model. 49 (2009) 1234–1244.
[16] P. Bultinck, Computational Medicinal Chemistry for Drug Discovery, Marcel Dekker, New York, 2004.
[17] P.v.R. Schleyer, Encyclopedia of Computational Chemistry, John Wiley, Chichester, New York, 1998.
[18] J.B. Brown, Y. Okuno, Systems biology and systems chemistry: new directions for drug discovery, Chem. Biol. 19 (2012) 23–28.
[19] J. von Eichborn, M.S. Murgueitio, M. Dunkel, S. Koerner, P.E. Bourne, R. Preissner, PROMISCUOUS: a database for network-based drug-repositioning, Nucleic Acids Res. 39 (2011) D1060–D1066.
[20] L. Yang, K. Wang, J. Chen, A.G. Jegga, H. Luo, L. Shi, C. Wan, X. Guo, S. Qin, G. He, G. Feng, L. He, Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome – clozapine-induced agranulocytosis as a case study, PLoS Comput. Biol. 7 (2011) e1002016.
[21] N.P. Tatonetti, T.Y. Liu, R.B. Altman, Predicting drug side-effects by chemical systems biology, Genome Biol. 10 (2009).
[22] J. Scheiber, B. Chen, M. Milik, S.C.K. Sukuru, A. Bender, D. Mikhailov, S. Whitebread, J. Hamon, K. Azzaoui, L. Urban, M. Glick, J.W. Davies, J.L. Jenkins, Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis, J. Chem. Inf. Model. 49 (2009) 308–317.
[23] A. Flemming, Chemoinformatics: where 'magic bullets' go astray, Nat. Rev. Drug Discov. 8 (2009) 933.
[24] E.G. Birgin, J.M. Martinez, Improving ultimate convergence of an augmented Lagrangian method, Optim. Methods Softw. 23 (2008) 177–195.
[25] R. Andreani, E.G. Birgin, J.M. Martinez, M.L. Schuverdt, On augmented Lagrangian methods with general lower-level constraints, SIAM J. Optim. 18 (2007) 1286–1309.
[26] R. Andreani, E.G. Birgin, J.M. Martinez, M.L. Schuverdt, Augmented Lagrangian methods under the constant positive linear dependence constraint qualification, Math. Program. 111 (2008) 5–32.
[27] A. Jahn, G. Hinselmann, N. Fechner, A. Zell, Optimal assignment methods for ligand-based virtual screening, J. Cheminform. 1 (2009) 14.
[28] J. Wang, W. Wang, P.A. Kollman, D.A. Case, Automatic atom type and bond type perception in molecular mechanical calculations, J. Mol. Graph. Model. 25 (2006) 247–260.
[29] M.S. Armstrong, P.W. Finn, G.M. Morris, W.G. Richards, Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension, J. Comput. Aided Mol. Des. 25 (2011) 785–790.
[30] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminform. 3 (2011) 33.
[31] M.A. Miteva, F. Guyon, P. Tuffery, Frog2: efficient 3D conformation ensemble generator for small compounds, Nucleic Acids Res. 38 (2010) W622–W627.
[32] J.L. Arriza, C. Weinberger, G. Cerelli, T.M. Glaser, B.L. Handelin, D.E. Housman, R.M. Evans, Cloning of human mineralocorticoid receptor complementary DNA: structural and functional kinship with the glucocorticoid receptor, Science 237 (1987) 268–275.
[33] C. Grossmann, T. Scholz, M. Rochel, C. Bumke-Vogt, W. Oelkers, A.F. Pfeiffer, S. Diederich, V. Bahr, Transactivation via the human glucocorticoid and mineralocorticoid receptor by therapeutically used steroids in CV-1 cells: a comparison of their glucocorticoid and mineralocorticoid properties, Eur. J. Endocrinol. 151 (2004) 397–406.
[34] J.J. Irwin, B.K. Shoichet, ZINC – a free database of commercially available compounds for virtual screening, J. Chem. Inf. Model. 45 (2005) 177–182.
[35] A.J. Rickard, J.W. Funder, P.J. Fuller, M.J. Young, The role of the glucocorticoid receptor in mineralocorticoid/salt-mediated cardiac fibrosis, Endocrinology 147 (2006) 5901–5906.
[36] L.A. Druce, C.M. Thorpe, A. Wilton, Mineralocorticoid effects due to cortisol inactivation overload explain the beneficial use of hydrocortisone in septic shock, Med. Hypotheses 70 (2008) 56–60.
[37] M.E. Rafestin-Oblin, A. Souque, B. Bocchi, G. Pinon, J. Fagart, A. Vandewalle, The severe form of hypertension caused by the activating S810L mutation in the mineralocorticoid receptor is cortisone related, Endocrinology 144 (2003) 528–533.
[38] M. Bureik, N. Bruck, K. Hubel, R. Bernhardt, The human mineralocorticoid receptor only partially differentiates between different ligands after expression in fission yeast, FEMS Yeast Res. 5 (2005) 627–633.
[39] P. Galuppo, J. Bauersachs, Mineralocorticoid receptor activation in myocardial infarction and failure: recent advances, Eur. J. Clin. Invest. 42 (2012) 1112–1120.
[40] J. Born, A. Zwick, G. Roth, G. Fehmwolfsdorf, H.L. Fehm, Differential-effects of hydrocortisone, fluocortolone, and aldosterone on nocturnal sleep in humans, Acta Endocrinol. 116 (1987) 129–137.
[41] L.V. Kulemina, D.A. Ostrov, Prediction of off-target effects on angiotensin-converting enzyme 2, J. Biomol. Screen. 16 (2011) 878–885.
[42] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, Development and testing of a general amber force field, J. Comput. Chem. 25 (2004) 1157–1174.
[43] E.E. Hodgkin, W.G. Richards, Molecular similarity based on electrostatic potential and electric-field, Int. J. Quantum Chem. (1987) 105–110.
[44] A.C. Good, E.E. Hodgkin, W.G. Richards, Similarity screening of molecular-data sets, J. Comput. Aided Mol. Des. 6 (1992) 513–520.
[45] K. Svanberg, A class of globally convergent optimization methods based on conservative convex separable approximations, SIAM J. Optim. 12 (2001) 555–573.
[46] S.G. Jhonson, The NLopt nonlinear-optimization package, in: The NLopt nonlinear-optimization package, http://ab-initio.mit.edu/nlopt.
[47] N. Huang, B.K. Shoichet, J.J. Irwin, Benchmarking sets for molecular docking, J. Med. Chem. 49 (2006) 6789–6801.
[48] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, ZINC: a free tool to discover chemistry for biology, J. Chem. Inf. Model. (2012).