J Comput Aided Mol Des (2010) 24:173–182
DOI 10.1007/s10822-009-9316-x

Spatial chemical distance based on atomic property fields

A. V. Grigoryan · I. Kufareva · M. Totrov · R. A. Abagyan

Received: 20 September 2009 / Accepted: 6 December 2009 / Published online: 13 March 2010

Abstract Similarity of compound chemical structures often leads to close pharmacological profiles, including binding to the same protein targets. The opposite, however, is not always true, as distinct chemical scaffolds can exhibit similar pharmacology as well. Therefore, relying on chemical similarity to known binders in search for novel chemicals targeting the same protein artificially narrows down the results and makes lead hopping impossible. In this study we attempt to design a compound similarity/distance measure that better captures structural aspects of their pharmacology and molecular interactions. The measure is based on our recently published method for compound spatial alignment with atomic property fields as a generalized 3D pharmacophoric potential. We optimized contributions of different atomic properties for better discrimination of compound pairs with the same pharmacology from those with different pharmacology using Partial Least Squares regression. Our proposed similarity measure was then tested for its ability to discriminate pharmacologically similar pairs from decoys on a large diverse dataset of 115 protein–ligand complexes. Compared to 2D Tanimoto and Shape Tanimoto approaches, our new approach led to improvement in the area under the receiver operating characteristic curve values in 66 and 58% of domains respectively. The improvement was particularly high for the previously problematic cases (weak performance of the 2D Tanimoto and Shape Tanimoto measures) with original AUC values below 0.8. In fact for these cases we obtained improvement in 86% of domains compare to 2D Tanimoto measure and 85% compare to Shape Tanimoto measure. The proposed spatial chemical distance measure can be used in virtual ligand screening.

Keywords Chemical distance · Chemical similarity · Quantitative structure-property relationship · Internal coordinates mechanics (ICM) · Atomic property fields (APF) · Spatial alignment

Electronic supplementary material The online version of this article (doi:10.1007/s10822-009-9316-x) contains supplementary material, which is available to authorized users.

A. V. Grigoryan · I. Kufareva · R. A. Abagyan (✉)
Department of Molecular Biology, TPC28, The Scripps Research Institute, 10550 N Torrey Pines Rd., La Jolla, CA 92037, USA
e-mail: rabagyan@ucsd.edu; abagyan@gmail.com

M. Totrov
Molsoft, LLC, 3366 N Torrey Pines Ct. Suite 300, La Jolla, CA 92037, USA

Introduction

Ligand-based approaches to protein family profiling has been widely studied and used for in silico pharmacology [1]. Similarity of compound chemical structures often leads to close pharmacological profiles, including binding to the same protein targets. By this reason, chemical similarity criterion is widely used for identification of novel lead molecules in the development of pharmaceuticals. A variety of chemical similar measures has been proposed. However, in many cases compounds with similar pharmacology escape correct recognition as they appear to be dissimilar by any existing measure.

In order to navigate in ligand space, one need to represent the compound using appropriate properties (descriptors) and then use a master equation to measure a distance between two compounds.

Descriptors are usually classified according to their dimensionality ranging from one-dimensional (1-D) to

three-dimensional (3D) properties [2, 3, 10]. Easy and fast to compute 1-D descriptors describe global properties which can be derived from chemical formula and classify compounds or ligands from various target families [3–5, 10]. To perform fast comparison 1-D linear representations of compounds are often used. The most popular of this kind of simplified string is the 'Simplified Molecular Input Line Entry System' or SMILES [3, 6, 10].

To improve discrimination, 2D topological descriptors are used. Graph-based methods, such as maximum common subgraph (MCS) [3, 7, 10] and fingerprint-based methods [3, 8, 10] are popular for substructure clustering chemical compounds into subfamilies. Subgraph isomorphism in large molecular databases is quite often time consuming to perform on large numbers of structures and it was for this reason that substructure screening was developed as a rapid method of filtering out those molecules that definitely do not contain the substructure of interest [10, 46]. The similarity between two molecules represented by 2D binary fingerprints is most frequently quantified using the Tanimoto coefficient, which gives a measure of the number of fragments in common between the two molecules [3, 9, 10].

It is well known that molecular recognition depends on the 3D structure and properties of molecule rather than the underlying substructure(s) [10]. 3D methods are computationally more expensive than 2D descriptor based methods, because they require consideration of conformational space of the molecule. These methods can be divided into methods that are alignment-independent and methods that require the molecules to be aligned in 3D space before similarity function is used [10].

Some computationally expensive alignment-independent methods use 3D geometrical descriptors represent them in a binary fingerprint and then use with the Tanimoto coefficient exactly as for 2D fingerprints [10, 11]. Other methods are 3D equivalent of the MCS [10, 12, 13]. Many 3D approaches are based on the use of distances matrices where the value of each element (i, j) equals the interatomic distance between atoms i and j [10, 14]. Also there are approaches where the pharmacophore points are used for similarity comparisons [10, 15–17].

Consideration of conformational flexibility of the molecules as well as their relative orientation is required for alignment dependent methods [10]. These methods devised to align the compared structures via maximization of the similarity function that is used [10, 45]. Many different ways have been developed to represent molecules and calculate similarity based on molecular shape and/or field [18–31, 45]. For reviews of molecular similarity methods, see refs [2, 10, 32–36].

The aim of this study is to design a spatial distance measure between two chemicals that optimizes recognition of their pharmacological similarity by using their 3D conformational ensembles and properties pertaining to molecular interactions. We recently introduced a novel spatial alignment method based on atomic property fields (APF) as a generalized 3D pharmacophoric potential [37]. APF is the representation of the ligand by a multi-component (vector) 3D potential, with the components corresponding to various physico-chemical atomic properties. In the present study, the APF alignment is used to measure spatial chemical similarity/distance between ligands.

A diverse benchmark of 99 proteins (see Supplementary Table 1 for details) and ligands co-crystallized with these proteins (with 6 ligands per protein on average) was used to train APF parameters for better discrimination of pharmacologically similar pairs from dissimilar ones. All possible combinations of pairs of ligands from the same receptors as well as for ligands co-crystallized with certain protein all possible combinations of pairs with ligands co-crystallized with 20 different randomly chosen from benchmark other proteins, were taken and APF representation of larger ligand was used as a reference to superimpose them. Distances between all superimposed pairs of ligands have been evaluated. Performance and results of proposed approach are reported and systematically compared to those obtained with standard 2D fingerprints (using 2D Tanimoto equation [38]) and 3D shape (using shape Tanimoto similarity measure [38]) based approaches (see Supplementary Table 2 for details).

Materials and methods

Atomic property field

Atomic property field (APF) [37] is the representation of the ligand by multi-component (vector) 3D grid potential $P_i(\vec{r})$, with the components i corresponding to various physico-chemical properties. Each property component of the APF determines whether the presence at any specific point $\vec{r}$ in space of an atom with that particular property is favorable or unfavorable. Pseudo-energy (score) of an atom j in this field is a dot product of its property vector ϕ_i^j and the APF potential at its position $\vec{r}^j$:

$$E_{\mathrm{APF}} = -\sum_i \phi_i^j P_i(\vec{r}^j). \quad (1)$$

The minima for this pseudo-energy for an atom with any specific property vector ϕ_i^j will be in the areas of space with similar APF potential components.

Assignment of property vectors ϕ_i^j for various atom types is carried out as described in [37], according to the general knowledge of their empiric physico-chemical behavior. Seven property field components were introduced:

hydrogen bond donor, hydrogen bond acceptor, sp^2 hybridized, lipophilic, size, charged, and electronegative/electropositive. Five of them (hydrogen bond donor, hydrogen bond acceptor, sp^2 hybridized, lipophilic and charged) are classic pharmacophoric types. Other two (size and electronegative/electropositive) are extension to make APF vectors differentiate certain atom types that are indistinguishable by the first five components: for example, aliphatic carbon (as in a methyl group) and large halogens are differentiated by electronegativity. For more details, see [37].

APF-based local ligand superimposition

Suboptimal ligand superimpositions were obtained for all pairs of compounds binding to the same protein by superimposing the protein binding pockets. These suboptimal superimpositions were improved by locally minimizing APF-correspondence between the two ligands. APF representation was generated from the larger ligand. The second ligand was flexibly minimized in the obtained field in a search for the local minimum of its APF-score calculated as described above. The local minimization procedure used the Newton method [39] if the number free variables of a given molecule are less than 100 and switched to the conjugate gradient method [39] if number of free variables exceeded 100. The maximal number of function evaluations was set to N^2, where N is number of free variables of a molecule. This local minimization procedure is implemented in the ICM software [40].

APF-based global ligand superimposition

For all ligand pairs that do not bind to the same protein, global ligand superimposition was performed by Monte Carlo search of the minimum of the APF-score of the smaller ligand in the APF of the larger. Similarly, the maximal number of function evaluations was set to N^2, where N is number of free variables of a given molecule. The global minimization procedure is implemented in the ICM software [40].

Dataset of protein–ligand complexes

The proposed approach to similarity measure was validated on a comprehensive benchmark of protein–ligand complexes from RCSB Protein Data Bank with high resolution and drug-like ligands (Kufareva et al. manuscript in preparation). For this validation, all ligands were paired up with one another to form the so-called *correct* and *decoy* pairs. A pair of ligands was considered *correct* if these ligands exhibited similar pharmacology exemplified by binding to the same protein target, as found in PDB. If two ligands were not found in co-crystal complexes with any single protein, they formed an *incorrect*, or *decoy* pair. As the number of decoy pairs was several order of magnitude larger than the number of correct pairs, we compressed the decoy set by pairing each ligand up with all ligands from 20 randomly chosen proteins (rather than all other proteins), excluding the cognate receptor.

The dataset was randomly split into a training set and a test set. The training set consisted of 2,538 correct and 9,819 decoy pairs, so that their numbers are in the same order of magnitude. The test set consisted of 1898 correct and 317,807 decoy pairs. For the training data set, correct ligand pairs were superimposed locally starting from their crystallographic poses as described above, while the decoy pairs were superimposed globally. For the test set all pairs were superimposed using global APF superimposition described above.

Partial least square regression

Partial least square (PLS) regression [41, 42] was used to design the proposed similarity measure. PLS is a recent technique that generalizes and combines features from principal component analysis and multiple regression. The purpose of PLS is to approximate a dependent (target) variable Y with a linear combination of independent variables X_i (descriptors). The output of the learning algorithm then consists of a linear coefficient c_i for each of the descriptors X_i and a free term c_0.

For each ligand pair in the training set, we assigned the value of Y equal to 1 for correct pairs and 0 for decoy pairs. PLS regression was trained to approximate these values as a linear combination of the seven descriptors $E^{2,i}_{\mathrm{APF}} \Big/ E^{1,i}_{\mathrm{APF}} \quad (i = 1, \ldots, 7)$. The optimal vector of coefficients, c_i, was used to access pairwise similarity if the compounds in the test set. The relative contributions w_i of seven descriptors, further referred as weights, were evaluated as $c_i \Big/ \sum_{i=1}^{7} c_i$. A recent implementation of PLS in ICM software [40] was used.

Design of novel spatial chemical distance measure

To design a novel and optimal spatial chemical distance measure we undertook the following approach (see Fig. 1). For each ligand pair, we performed the following steps. First, the larger ligand (greater number of heavy atoms) was locally relaxed from its bioactive conformation in its receptor and its APF representation was built. Second, the smaller ligand was flexibly superimposed onto the larger ligand using local and global APF-based superimposition for correct and decoy pairs, respectively. Third, the pseudo-energies, $E^{1,i}_{\mathrm{APF}}$ and $E^{2,i}_{\mathrm{APF}} \quad (i = 1, \ldots, 7)$, of the larger and the smaller ligands in the APF potential grid of the larger ligand were calculated.

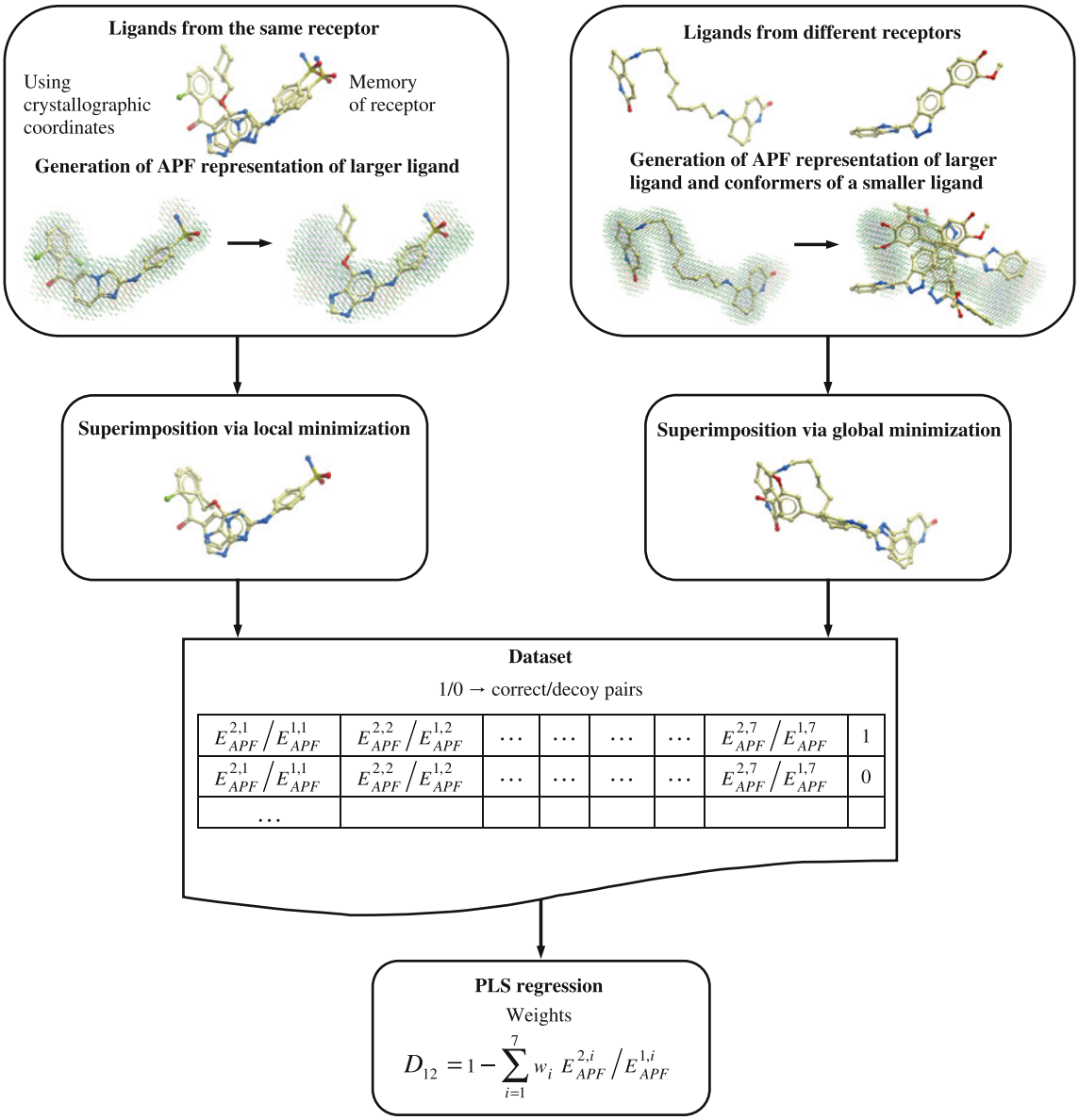

Fig. 1 Schematic outline of approach to design a novel spatial chemical distance measure

Further on, the seven ratios $E_{\mathrm{APF}}^{2,i}/E_{\mathrm{APF}}^{1,i}$ $(i = 1, \ldots, 7)$, were recorded as descriptors of the ligand pair, and PLS regression was trained to approximate the correct/decoy (1/0) value as a weighted sum

$$\sum_{i=1}^{7} w_i \frac{E_{\mathrm{APF}}^{2,i}}{E_{\mathrm{APF}}^{1,i}}. \tag{2}$$

The protocol is built on top of a well established and implemented in ICM APF concept and ICM gradient minimization.

The weights w_i $(i = 1, \ldots, 7)$ obtained by the PLS regression were used in the definition of the compound chemical distance as:

$$D_{12} = 1 - \sum_{i=1}^{7} w_i \frac{E_{\mathrm{APF}}^{2,i}}{E_{\mathrm{APF}}^{1,i}}. \tag{3}$$

2D tanimoto distance calculation

Compounds chemical similarity was evaluated as Tanimoto distance between their molecular fingerprints as implemented in ICM [40]. Briefly, given a molecule, all linear and non-linear fragments of different size were enumerated and hashed into a bit string called a *fingerprint*. The *Tanimoto coefficient*, *T*, for two fingerprints was calculated as the number of bits in which they differ divided

by the number of non-zero bits they have in common. The Tanimoto distance was defined as $1 - T$.

3D shape tanimoto distances

For each compound, the volume of its molecular envelope was calculated using ICM. For every pair of compounds superimposed by the APF-based algorithm, the volume of the smallest envelope enclosing both compounds, V, was calculated. The overlapping volume, V_{12}, was found by $V_{12} = V_1 + V_2 - V$. The 3D shape Tanimoto distance [38] between the two compounds was calculated as

$$\text{3D Shape Tanimoto Distance} = 1 - \frac{V_{12}}{V}, \quad (4)$$

where V_{12} is the volume overlap between two compounds, V_1 and V_2 are volumes of compounds.

Distance distribution analysis

Distance probability distribution curves were built by normalization (division by number of observations) of histogram plots of 2D Tanimoto, 3D Shape Tanimoto, and our newly designed distances for all correct and decoy pairs from the test set. Global discrimination abilities of 2D Tanimoto, 3D Shape Tanimoto, and our newly designed distance measure are evaluated by calculating an overlapping area between distance probability distribution curves for all correct and decoy pairs from the test set.

ROC AUC values

The new measure was evaluated for its ability to distinguish correct compound pairs from decoy pairs on the large test set. The performance of the measure was evaluated as the area under the ROC (receiver operating characteristic) curve, or ROC AUC. ROC curve analysis [43] describes the ability of a screening method to avoid false positives and false negatives. The ideal screening device demonstrates the ROC AUC value of 1, while a random selection performance corresponds to the ROC AUC value of 0.5.

ROC AUC analysis was performed for 2D Tanimoto, 3D Shape Tanimoto, and our newly designed chemical distance measure for the entire test set. Also ROC AUC analysis was performed for 2D Tanimoto, 3D Shape Tanimoto, and our newly designed chemical distance measure for each of the 115 protein ensembles in the testing set.

Statistical analysis

To evaluate the statistical significance of the obtained ROC AUC values, we assumed that they are distributed according to Gaussian distribution with mean 0.5. The uncertainty σ is calculated by performing 20 random statistical experiments in which the ranks are reshuffled, calculating individual ROC AUCs and calculating the root mean square deviation of those ROC AUCs. For each ROC AUC values obtained on a given subset of correct and decoy pairs, its P-value was calculated as probability to obtain the same AUC by random coincidence:

$$P\text{-value} = 1 - \frac{1}{2}\left[1 + \text{erf}\left(\frac{\text{ROC AUC value} - 0.5}{\sigma\sqrt{2}}\right)\right], \quad (5)$$

where erf is the error function [44] and σ is the standard deviation. Small P-values below 0.05 mean high statistical significance, while cases with large P-values are statistically less significant.

Software and hardware

Ligand preparation, superimpositions, APF generations, evaluation of pseudo-energies and PLS regression analysis were carried out with ICM 3.6 (Molsoft LLC, La Jolla, CA).

The hardware facility employed in the present study was Intel Core 2 Duo workstation (2.4 GHz with 2 GB of RAM memory).

Results and discussion

APF-based compounds chemical distance measure

In this study, we propose an algorithm for evaluation of compound spatial chemical distance that consists of five consecutive steps (Fig. 2): (1) the ligand with greater number of heavy atoms is locally relaxed from its bioactive coordinates (found in the co-crystal structure with its target protein); (2) APF potential grid maps are built from this ligand; (3) the smaller ligand is flexibly superimposed onto the larger ligand by minimizing its APF-score in the field of the larger ligand; (4) ratios $E^{2,i}_{\text{APF}} \Big/ E^{1,i}_{\text{APF}} \quad (i = 1, \ldots, 7)$ are calculated, where $E^{1,i}_{\text{APF}}$ are the pseudo-energies of the larger ligand in its own APF maps and $E^{2,i}_{\text{APF}}$ are the pseudo-energies of the smaller ligand in APF representation of the larger one; (5) the similarity score is calculated by (2) and the chemical distance between the compounds is found by (3).

The weights w_i in Eq. (2) were derived by Partial Least Squares regression optimizing the pairwise correct/decoy (1/0) discrimination value as a linear combination of the $E^{2,i}_{\text{APF}} \Big/ E^{1,i}_{\text{APF}} \quad (i = 1, \ldots, 7)$ ratios as described in "Materials and Methods". The results of regression are shown in the Table 1. The APF potential component corresponding to the atomic size was found to contribute the most to the compound similarity. This result indicates the fact that shape complementarity between compounds has a

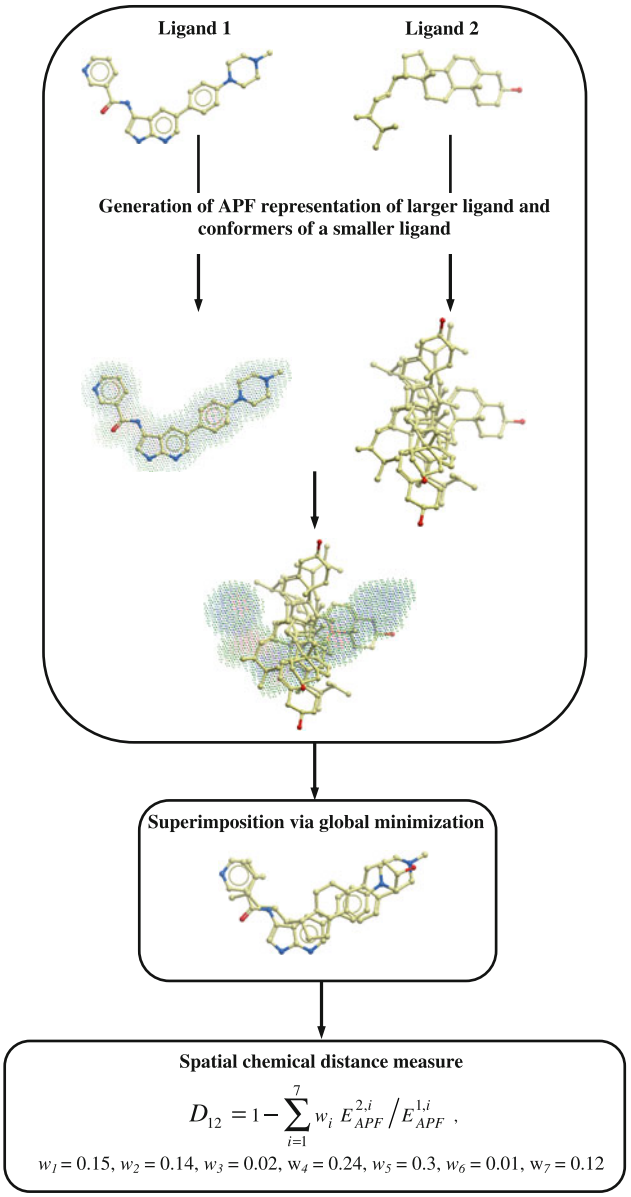

Fig. 2 Schematic flowchart of calculation of spatial chemical distance between two compounds

Table 1 Relative and absolute contribution of components of APF potential to spatial chemical distance measure

Components of APF potential	Relative contribution w_i (%)	Coefficient c_i
Hydrogen bond donor	15	0.5
Hydrogen bond acceptor	14	0.46
sp^2 hybridization	2	0.035
Lipophilicity	24	0.73
Size	30	−0.86
Charge	1	0.01
Electronegativity	12	0.37

fundamental role, so that compounds with similar shape often have similar pharmacological profiles.

According to these results, the spatial chemical distance between two compounds can be measured by

$$D_{12} = 1 - \left(0.15\frac{E_{\mathrm{APF}}^{2,1}}{E_{\mathrm{APF}}^{1,1}} + 0.14\frac{E_{\mathrm{APF}}^{2,2}}{E_{\mathrm{APF}}^{1,2}} + 0.02\frac{E_{\mathrm{APF}}^{2,3}}{E_{\mathrm{APF}}^{1,3}} + 0.24\frac{E_{\mathrm{APF}}^{2,4}}{E_{\mathrm{APF}}^{1,4}} + 0.3\frac{E_{\mathrm{APF}}^{2,5}}{E_{\mathrm{APF}}^{1,5}} + 0.01\frac{E_{\mathrm{APF}}^{2,6}}{E_{\mathrm{APF}}^{1,6}} + 0.12\frac{E_{\mathrm{APF}}^{2,7}}{E_{\mathrm{APF}}^{1,7}} \right). \quad (6)$$

Screening performance of novel spatial chemical distance measure

The screening performance of the proposed spatial chemical distance measure was tested on the diverse set of 1898 correct (from 115 proteins) and 317,807 decoy pairs (see Supplementary Table 2). Global discrimination abilities of different methods used in this study are presented in the Fig. 3. In this plot, the relative frequencies of three types of compound distances are shown for correct and decoy compound pairs. As expected, the distribution peaks are shifted with respect to each other in all three cases, reflecting the fact that on average, compounds in correct pairs tend to be closer to each other than compounds in decoy pairs. However, it is clear that the newly proposed chemical distance measure provides better separation of the peaks than either 2D fingerprint Tanimoto distance or 3D shape Tanimoto distance. Indeed, by our new measure, only 52% of the correct pairs appear in the distance region of decoy pairs (Fig. 3c). The corresponding number for the other two distance measures appear to be 64 and 56% (Fig. 3a, b). Discrimination abilities of different methods used in this study also can be seen in the Fig. 4, where ROC curves and overall correct/decoy discrimination ROC AUC values for the test set are presented. Our new proposed chemical distance measure gives ROC AUC value of 0.81, in contrast to 0.79 and 0.74 obtained by using 3D shape Tanimoto distance and 2D fingerprint Tanimoto distance. These results clearly shows an advantage of our new approach in lead/decoy discrimination on the benchmark compared to 2D Tanimoto and 3D Shape Tanimoto approaches.

It also can be seen in the Fig. 5 where the performance of the proposed measure was individually evaluated for each of the 115 protein targets from the test set. By using the novel spatial compounds chemical distance measure, the correct/decoy discrimination ROC AUC value was improved for 66% of the test set when compared to 2D Tanimoto, and 58% when compared to 3D Shape Tanimoto. The improvement was particularly high for the previously problematic cases (weak performance of the 2D Tanimoto and Shape Tanimoto measures) with original AUC values below 0.8. For these cases we obtained

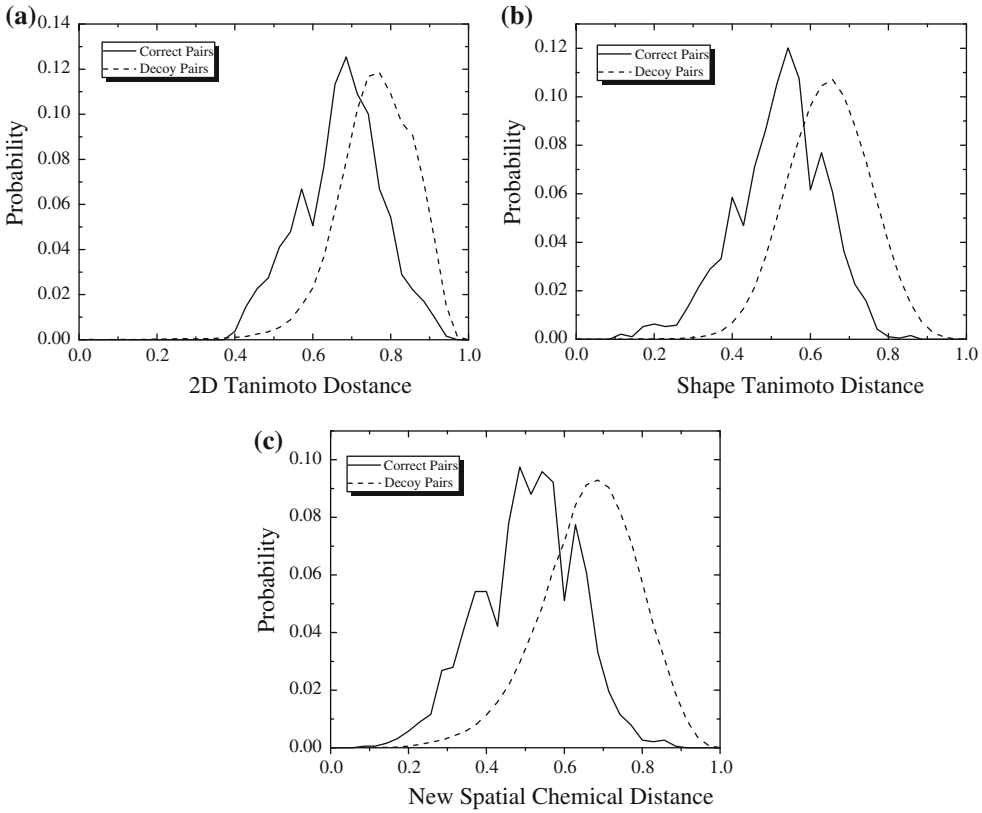

Fig. 3 Probability distribution for correct and decoy pairs: **a** 2D Tanimoto approach. Overlapped area is 64%; **b** Shape Tanimoto approach. Overlapped area is 56%; **c** New approach based on APF. Overlapped area is 52%

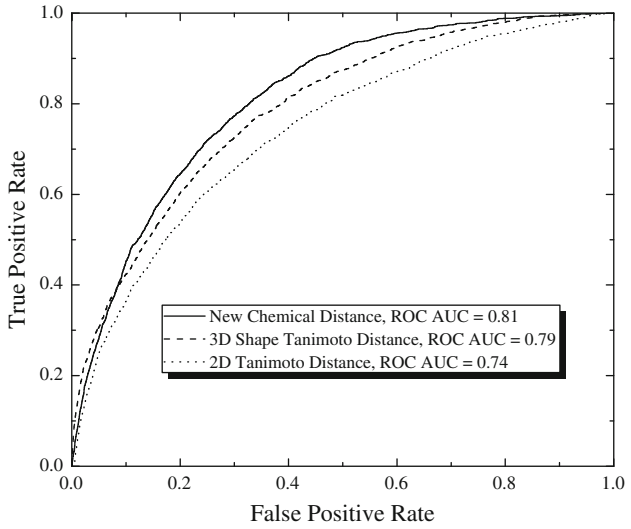

Fig. 4 ROC curves and overall correct/decoy discrimination ROC AUC values evaluated by using our novel chemical distance, 3D Shape Tanimoto distance and 2D Tanimoto fingerprint

improvement in 86% of domains compared to 2D Tanimoto measure and 85% compared to 3D Shape Tanimoto measure. The average ROC AUC value improvement for these cases equals 0.12 and 0.24 with standard deviation 0.15 and 0.24, respectively (Fig. 5a, b).

Compared to 2D Tanimoto approach, the best improvement (up to 0.92 in AUC units) was achieved for compounds with perfectly matching pharmacophores but no common substructure. An example is given in the Fig. 6, where the two ligands were taken from Geranyltranstransferase and globally superimposed as described above. The newly proposed distance measure returns the value of 0.28 for these two compounds, in contrast to 0.88 obtained by 2D Tanimoto measure. Consequently, using our measure for correct/decoy discrimination for this protein resulted in the ROC AUC value of 0.99, in contrast to 0.07 obtained by 2D Tanimoto approach. As another extreme, there are cases where 2D Tanimoto approach provides better recognition (Fig. 5a, AUCs based on 2D Tanimoto distance greater than 0.8). In general these are the cases where the two compounds have a common substructure but other parts of them are significantly different. However, even in these cases, our new approach performs reasonably accurate and provides systematically high AUCs with 0.85 on average and 0.11 standard deviation.

There is no doubt that ligand shape, and its complimentary to the shape of the protein interaction site are of fundamental importance for their bioactivity. However, comparison of ligands by shape causes compounds of dissimilar dimensions to be down-ranked during virtual screening. Such ligands, however, may have high affinity to a target as they may form additional interactions with the protein. The results presented in the Fig. 5b), namely up to 0.54 improvement in AUC, show that our new spatial chemical distance measure

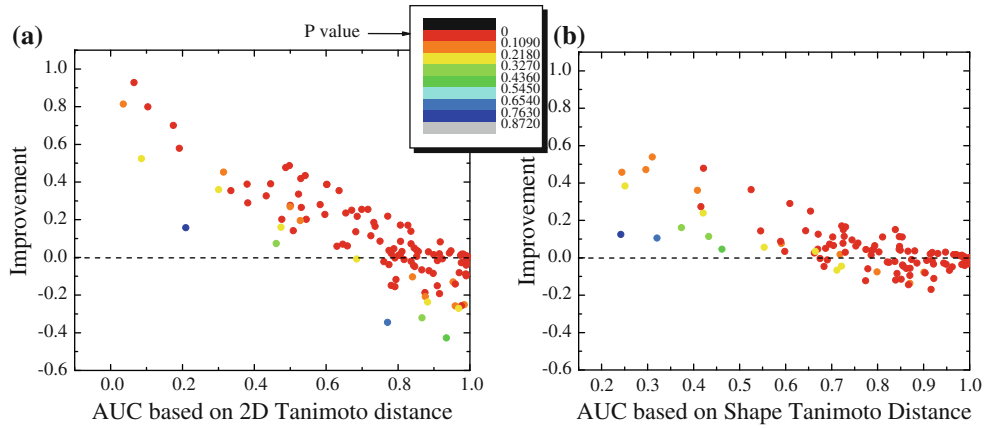

Fig. 5 Color-coded dependence of achieved improvement in ROC AUC values on original ROC AUC values obtained by 2D Tanimoto (**a**) and Shape Tanimoto (**b**) measures. Each *dot* represents one domain. The *dots* are colored in different colors spread evenly among *P*-values from 0.00 to 1.00 (smallest *P*-values, that is, most statistically significant are *red*, highest *P*-values are *blue*). **a** Overall improvement achieved in 66% of the cases in active/inactive ligand discrimination. The improvement is particularly high for the previously problematic cases with original AUC value below 0.8. In fact for these cases we obtained improvement in 86% of the proteins; **b** Overall improvement achieved in 58% of the cases in active/inactive ligand discrimination. The improvement is particularly high for the previously problematic cases with original AUC value below 0.8. In fact for these cases we obtained improvement in 85% of the proteins

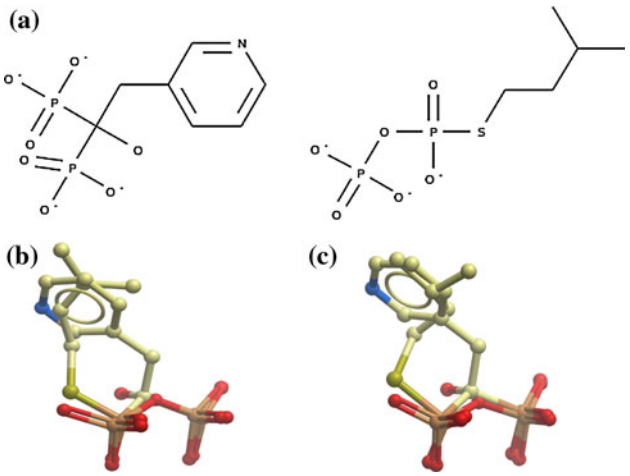

Fig. 6 Indicative example of weak performance of approach based on 2D Tanimoto similarity measure: **a** 2D structures of two ligands, which have taken from Geranyltranstransferase; **b** same ligands in their bioactive conformations; **c** ligands superimposed by approach suggested in this study. According to 2D Tanimoto similarity measure, distance between these two compounds equals 0.88, in contrast to measure introduced here, which gives distance 0.28. For this particular protein AUC based on 2D Tanimoto approach equals 0.07, in contrast to 0.99, obtained based on our new approach

is able to discover bioactive compounds that are down-ranked by the 3D-shape based measure.

Conclusions

In this study, we developed a spatial chemical distance measure between two chemicals by using their 3D conformational ensembles and properties pertaining to molecular interactions. The parameters for our novel measure were obtained by training a PLS regression to distinguish the correct ligand pairs from decoys on a large and sufficiently diverse set of correct pairs of ligands, which were taken from the same receptor and decoy pairs of ligands, which were taken from different receptors. The screening performance of the proposed spatial chemical distance measure was tested on a diverse and pharmaceutically relevant test set of correct and decoy pairs. Compared to 2D Tanimoto and Shape Tanimoto approaches, our new approach led to improvement in the area under the receiver operating characteristic curve values in 66 and 58% of domains respectively. The improvement was particularly high for the previously problematic cases (weak performance of the 2D Tanimoto and Shape Tanimoto measures) with original AUC values below 0.8. In fact for these cases we obtained improvement in 86% of domains compare to 2D Tanimoto measure and 85% compare to Shape Tanimoto measure.

The presented results suggest that our new spatial chemical distance measure can be successfully used in virtual ligand screening for novel chemical scaffolds targeting the existing proteins of therapeutic relevance.

Acknowledgments This work was supported by NIH grants 5-R01-GM071872 and 1-R01-GM074832. We are indebted to Manuel Rueda for helpful discussions.

References

1. Mestres J, Martin-Couce L, Gregori-Puigjane E, Cases M, Boyer S (2004) Ligand-based approach to in silico pharmacology: nuclear Receptor Profiling. J Chem Inf Model 46:2725–2736
2. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular information. Org Biomol Chem 2:3204–3218
3. Rognan D (2007) Chemogenomic approaches to rational drug design. Br J Pharmacol 152:38–52
4. Sadowski J, Kubinyi HA (1998) scoring scheme for discriminating between drugs and nondrugs. J Med Chem 41:3325–3329
5. Morphy R (2006) The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. J Med Chem 49:2969–2978
6. Weininger D (1988) SMILES 1. Introduction and encoding rules. J Chem Inf Comput Sci 28:31–36
7. Raymond JW, Blankley CJ, Willett P (2003) Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. J Mol Graph Model 21:421–433
8. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11:1046–1053
9. Willett P, Winterman V, Bawden D (1986) Implementation of nearest neighbor searching in an online chemical structure search system. J Chem Inf Comput Sci 38:983–996
10. Leach AR, Gillet VJ (2007) An introduction to cheminformatics. Springer, Dordrecht
11. Fisanick W, Lipkus AH, Rusinko A III (1992) Similarity searching on CAS registry substances. 1. Global molecular property and generic atom triangle geometry searching. J Chem Inf Comput Sci 32:664–674
12. Moon JB, Howe WJ (1990) 3D database searching and de novo construction methods in molecular design. Tetrahedron Comput Methodol 3:697–711
13. Pepperrell CA, Willett P (1991) Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. J Comput-Aided Mol Des 5:455–474
14. Pepperrell CA, Willett P, Taylor R (1990) Implementation and use of an atom-mapping procedure for similarity searching in databases of 3D chemical structures. Tetrahedron Comput Methodol 3:575–593
15. Good A, Lewis RA (1997) New methodology for profiling combinatorial libraries and screening sets: cleaning up the design with HARPick. J Med Chem 40:3926–3936
16. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Lebaudiniere RF (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. J Med Chem 42:3251–3264
17. Good AC, Mason JS, Green DVS, Leach AR (2001) Pharmacophore-based approaches to combinatorial library design. In: Ghose AK, Viswanadhan VN (eds) Combinatorial library design and evaluation. Principles, software tools and applications in drug discovery. Marcel Dekker, New York, pp 399–428
18. Kearsley SK, Smith GM (1990) An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. Tetrahedron Comput Methodol 3:615–633
19. Bemis GW, Kuntz ID (1992) A fast and efficient method for 2D and 3D molecular shape description. J Comput-Aided Mol Des 6:607–628
20. Wild D, Willett P (1996) Similarity searching in files of three-dimensional chemical structures. Alignment of molecular electrostatic potential fields with a genetic algorithm. J Chem Inf Comput Sci 36:159–167
21. Thorner D, Wild D, Willett P, Wright P (1996) Similarity searching in files of three-dimensional chemical structures: flexible field-based searching of molecular electrostatic potentials. J Chem Inf Comput Sci 36:900–908
22. Hahn M (1997) Three-dimensional shape-based searching of conformationally flexible compounds. J Chem Inf Comput Sci 37:80–86
23. Rush T, Grant J, Mosyak L, Nicholls A (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem 48:1489–1495
24. Labute P, Williams C, Feher M, Sourial E, Schmidt J (2001) Flexible alignment of small molecules. J Med Chem 44:1483–1490
25. Pitman MC, Huber WK, Horn H, Kramer A, Rice JE, Swope WC (2001) FLASHFLOOD: a 3D field-based similarity search and alignment method for flexible molecules. J Comput-Aided Mol Des 15:587–612
26. Putta S, Landrum G, Penzotti J (2005) Conformation mining: an algorithm for finding biologically relevant conformations. J Med Chem 48:3313–3318
27. Tervo A, Ronkko T, Nyronen T, Poso A (2005) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. J Med Chem 48:4076–4086
28. Rönkkö T, Tervo AJ, Parkkinen J, Poso A (2006) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. J Comput-Aided Mol Des 20:227–236
29. Cheeseright T, Mackey M, Rose S, Vinter A (2006) Molecular field extrema as descriptors of biological activity: definition and Valida-tion. J Chem Inf Model 46:665–676
30. Good AC (2007) Novel DOCK clique driven 3D similarity database search tools for molecule shape matching and beyond: adding flexibility to the search for ligand kin. J Mol Graph Model 26:656–666
31. Kirchmair J, Distinto S, Markt P, Schuster D, Spitzer GM, Liedl KR, Wolber G (2009) How to optimize shape-based virtual screening: choosing the right query and including chemical information. J Chem Inf Model 49:678–692
32. Lemmen C, Lengauer T (2000) Computational methods for the structural alignment of molecules. J Comput-Aided MolDes 14:215–232
33. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today 12:225–233
34. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983–996
35. Maldonado AG, Doucet JP, Petitjean M, Fan B-T (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers 10:39–79
36. Good AC, Richards WG (1998) Explicit calculation of 3D molecular similarity. Perspect Drug Discov Design 9/10/11, 321–338
37. Totrov M (2008) Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. Chem Biol Drug Des 71:15–27
38. Fontaine F, Bolton E, Borodina Y, Bryant SH (2007) Fast 3D shape screening of large chemical databases through alignment-recycling. Chem Cent J 1:12

39. Leach AR (2001) Molecular modeling. Principles and applications. Principles and Applications, Edinburgh
40. Abagyan R (2008) ICM Manual v. 3.6
41. Geladi P, Kowalski B (1986) Partial least squares regression: a tutorial. Anal Chim Acta 185:1–17
42. Hand D, Mannila H, Smyth P (2001) Principles of data mining. The MIT Press, Cambridge
43. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
44. Milton Abramowitz M, Stegun IA (eds) (1972) Handbook of mathematical functions with formulas, graphs, and mathematical tables. Dover, New York
45. Vainio MJ, Puranen JS, Johnson MS (2009) ShaEP: molecular overlay based on shape and electrostatic potential. J Chem Inf Model 49:492–502
46. Brown N (2009) Cheminformtics–an introduction for computer scientists. ACM Comput Surv 41:1–38