DOI 10.1007/s10822-008-9253-0

Comparison of structure fingerprint and molecular interaction field based methods in explaining biological similarity of small molecules in cell-based screens

Pekka Tiikkainen · Antti Poso · Olli Kallioniemi

Received: 4 September 2008 / Accepted: 15 November 2008 / Published online: 3 December 2008

Abstract In this work, we calculated the pair wise chemical similarity for a subset of small molecules screened against the NCI60 cancer cell line panel. Four different compound similarity calculation methods were used: Brutus, GRIND, Daylight and UNITY. The chemical similarity scores of each method were related to the biological similarity data set. The same was done also for combinations of methods. In the end, we had an estimate of biological similarity for a given chemical similarity score or combinations thereof. The data from above was used to identify chemical similarity ranges where combining two or more methods (data fusion) led to synergy. The results were also applied in ligand-based virtual screening using the DUD data set. In respect to their ability to enrich biologically similar compound pairs, the ranking of the four methods in descending performance is UNITY, Daylight, Brutus and GRIND. Combining methods resulted always in positive synergy within a restricted range of chemical similarity scores. We observed no negative synergy. We also noted that combining three or four methods had only limited added advantage compared to combining just two. In the virtual screening, using the estimated biological similarity for ranking compounds produced more consistent results than using the methods in isolation.

P. Tiikkainen (✉) · O. Kallioniemi
VTT Medical Biotechnology and University of Turku,
P.O. Box 106, 20521 Turku, Finland
e-mail: Pekka.tiikkainen@utu.fi

O. Kallioniemi
e-mail: olli.kallioniemi@vtt.fi

A. Poso
Department of Pharmaceutical Chemistry, University
of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland
e-mail: antti.poso@uku.fi

Keywords Ligand-based virtual screening · NCI-60 · Data fusion · Chemical similarity

Introduction

Chemical similarity is the most important factor explaining functional similarity of small molecules [1]. Usually the chemical similarity is calculated as function of structural features present/absent in the small molecules being compared [2]. Another approach for measuring chemical similarity is via comparison of molecules interaction fields (MIF) [3]. 2D structures are sufficient for the former approach whereas three dimensional conformations are required for the latter. Similar MIF patterns can arise from very dissimilar structures as long as important functional groups can be overlaid in the 3-dimensional space with sufficient shape complementarity, i.e., scaffold hopping.

Here, we applied two MIF-based methods, alignment-free GRIND descriptors [4] and BRUTUS [5, 6], for evaluation of chemoresponse data from a panel of cancer cell lines. Also Unity (Tripos, Inc. http://www.tripos.com) and Daylight fingerprints (Daylight Chemical Information Systems, Inc. http://www.daylight.com) were employed for detailed comparison of the methods. The objective was to determine how well the MIF methods perform in comparison to structure fingerprints in explaining biological correlation of the compounds in a cell-based screen of the 60 cancer cell lines. Analogously we studied how well chemical similarity could be explained as function of biological similarity. We also wanted to study the synergy of the methods by combining two or more chemical similarity methods [7]. Specific attention was paid on functionally similar compound pairs which MIF methods ranked as similar while the fingerprints did not (“scaffold-hopping cases”).

An approach analogous to ours has been used by Muchmore et al. [8]. They had related the chemical similarity of a compound library calculated with 10 different metrics to the biological data measured in-house for the same compounds. In addition to this they had applied the relationship data as a data fusion method. Their results show that this way one can have similar gains in synergy as with the SUM rule for data fusion. Their approach also allows for a quantitative estimate for the probability that two molecules have similar biological activity.

The biological similarity data used here came from the publicly available data of a systematic screen of small molecule compounds against a panel of 60 cancer cell lines data produced at the National Cancer Institute (Developmental Therapeutics Program. http://dto.nci.nih.gov/index.html) (NCI-60 dataset). A subset of 16,222 compounds was used in the present work.

The biological end point variable considered here is GI50, the molar concentration of the given compound where growth of the treated cells is inhibited by 50% compared to an untreated control sample. Combining the GI50 values from different cell lines gives a detailed cytotoxicity profile for each compound. These profiles have been widely investigated earlier. For example, the chemoresponse data have been correlated with pre-treatment gene expression profiles and genetic alterations targeting specific genes or pathway [9–12]. The chemoresponse data have been previously evaluated with structure-based Daylight fingerprints by Wallqvist et al. [13], but to our knowledge this is the first time this dataset has been explored using the MIF-based methods.

Directory of useful decoys (DUD) [14] is a collection of compounds originally designed for benchmarking docking software. The set contains two types of small molecules, actives with defined molecular targets and decoy molecules which resemble the actives on their physicochemical properties such as molecular weight but are assumed to be inactive. In the current work, we used DUD to see if the relationship between small molecules’ biological similarity and the combination of their GRIND and Brutus chemical similarity could be used as a data fusion method and therefore to lead improved enrichment of known DUD ligands when compared to the use of individual methods.

Materials and methods

Daylight fingerprints and biological data

Tanimoto similarity indices based on Daylight fingerprints and the cytotoxicity profile correlation data were kindly donated by Anders Wallqvist. The detailed explanation of how these results were calculated can be found in Wallqvist’s article [13]. In short, the original GI50 values across 60 cell lines were log transformed and profiles lacking data from more than 20 cell lines and with too little covariance were excluded. The two molecules were compared by the Pearson correlation of their cytotoxicity profiles. This is also the definition of biological similarity used throughout the present article.

Out of the molecule pairs provided by Wallqvist, 59,489,275 were used in the present work. These were the molecule pairs for which a chemical similarity could be calculated with at least one of the other methods described below. Chemical similarity for all possible compound pairs couldn’t be exhaustively calculated mainly due to the demands on computational time this would have required. Instead, we elected to pick a random subset of the compound pairs for each method and work with those.

Unity fingerprints

To compare the performance of Daylight fingerprints with another 2D method, Tanimoto similarities of the molecules were calculated using Unity fingerprints. This was done with Molecule Spreadsheet part of Sybyl 8.0 modelling software (Tripos, Inc. http://www.tripos.com)). Similarities were calculated for 38,332,428 pairs representing a total of 15,653 individual molecules.

GRIND descriptors

GRIND descriptors are designed to compare molecular interaction fields of small molecules without the need to overlay the molecules first. Originally, the descriptors were designed for building 3D QSAR models for structurally diverse compounds [15–17] but they are also applicable for virtual screening. The details of the GRIND method can be found elsewhere [5].

First a set of 3D structures was generated for each compound. For this, we used Corina Version 3.2 [18] with the following command:

```
corina -i t=sdf,sdfi2n=NSC -o t=mol2,nodummies -d newtypes,r2d,rs,wh,rc,stergen,names,de=20<input SD file> <output file>
```

The command generates all possible stereoisomers of the compound that are within 20 kJ/mol of the lowest energy conformation.

GRIND descriptors were then calculated with Almond 3.3.0 (Molecular Discovery Ltd. http://www.moldiscovery.com/soft_almond.php). Three probes were used to represent non-covalent interactions: DRY (hydrophobic interactions), O (hydrogen bond acceptor) and N1 (hydrogen bond donor). Also the TIP probe was employed to take the shape of the molecule in account [19]. Ten correlograms were generated with 122 descriptors in each of them. All the other settings were left to the program defaults. The correlograms were exported into a text file and merged in a single GRIND profile. At the end, we had a set of GRIND profiles for each molecule, corresponding to all the different 3D conformations. Compounds with more than 40 profiles (conformations) were excluded to save computation time. This led to exclusion of 520 compounds.

Two compounds were compared by calculating the Pearson correlation of their GRIND profiles. The correlation was calculated for all the combinations of conformations the molecules had. The largest correlation was chosen to represent similarity of the two molecules. Similarity data was calculated for 48,868,066 pairs representing 14,720 distinct molecules.

Brutus

The algorithm employed in the calculation of GRIND descriptors loses some information when the lower affinity grid pairs within a distance bin are discarded. Therefore, we also wanted to use another 3D method which should not lose any information. For this, we chose Brutus [5, 6] which superimposes two molecules based on their electrostatic and steric properties. A score is then calculated to measure the quality of the superposition.

Since running Brutus analysis for all the compound pairs would have been very expensive computationally, we decided to use it only for a subset of the pairs. We chose the pairs by random among those in the Daylight set. After filtering out structures with features not accepted by Brutus, we had 3,018,315 compound pairs for which we could calculate the similarity. These pairs represent 12,767 individual compounds.

3D structures for the molecules were generated as for the GRIND descriptors but only one conformation for each stereoisomer was retained. The conformations were further minimized with a custom made Sybyl script using the MMFF94s force field with MMFF94 point charges. Next a systematic search was conducted to generate a set of conformations for each stereoisomer.

When two molecules were compared, a single minimized conformation of each stereoisomer of one molecule was used as a template against all the conformations of the other molecule. From the set of scores that resulted, the highest one was retained as the similarity value for the two molecules.

Chemical versus biological similarity

Here we use similar mathematical notation as used by Wallqvist et al. [13]. Definition

$$N(c \geq a; r \geq b) \tag{1}$$

gives the number of compound pairs with a chemical similarity of a or higher (using a chemical similarity c) and a biological similarity of at least b (using a biological similarity r). This leads to the definition of

$$F(c = a|r = b) \equiv \frac{N(c \geq a; r \geq b)}{N(c \geq a; r \geq -1.0)} \tag{2}$$

which gives us a ratio of Eq. 1 (numerator) divided by the number of all compound pairs with chemical similarity of a or higher (denominator). Equation 2 can be easily extended for two or more types of chemical similarity. For example, we can write $F(g = 0.90,\ b = 2.50|r = 0.80)$. This gives the ratio of compound pairs whose biological similarity is at least 0.80 and chemical similarities are at least 0.90 and 2.50 as given by GRIND and Brutus, respectively.

Similarly, we can define

$$G(c = a|r = b) \equiv \frac{N(c \geq a; r \geq b)}{N(c \geq \text{chem\_min};\ r \geq b)} \tag{3}$$

where chem_min is the minimum value for chemical similarity with a given method. The equation gives the share of compound pairs with biological similarity of b or greater that also have a chemical similarity of a or greater. Again, Eq. 3 can be trivially extended for more than one chemical similarity method.

Synergy calculation

Using the notation introduced above, we can define relative synergy for a combination of two methods as

$$S_{\text{rel}}(x = i, y = j|r = b) \equiv \frac{F(x = i, y = j|r = b)}{\max\left[F(x = \text{chem\_min}_{\text{x}}, y = j|r = b); F(x = i, y = \text{chem\_min}_{\text{y}}\,|r = b)\right]} \tag{4}$$

where chem_min$_{\text{x}}$ and chem_min$_{\text{y}}$ are the smallest values for chemical similarity using methods x and y, respectively. Using the numerator and denominator from the equation above, we can define absolute synergy S_{abs} as the difference of the two.

Work with the DUD dataset

Directory of Useful Decoys [14] consists of 40 sets of ligands against 39 different protein targets. For each known ligand, there are on average 33 decoy molecules which are assumed not to bind the target. Originally created for evaluation of docking software, we also used the dataset for testing the algorithms in ligand-based virtual analysis. The objective was to test if the biological similarity values could be used to enrich known ligands of a target class using another known ligand as template.

Using Brutus and GRIND, each known ligand's chemical similarity was calculated for all the other known and decoy molecules of its target set. This resulted in 2,805 lists for both GRIND and Brutus.

Decoy ligands for each set are chosen so that they are chemically distinct from all the active ligands of the set.

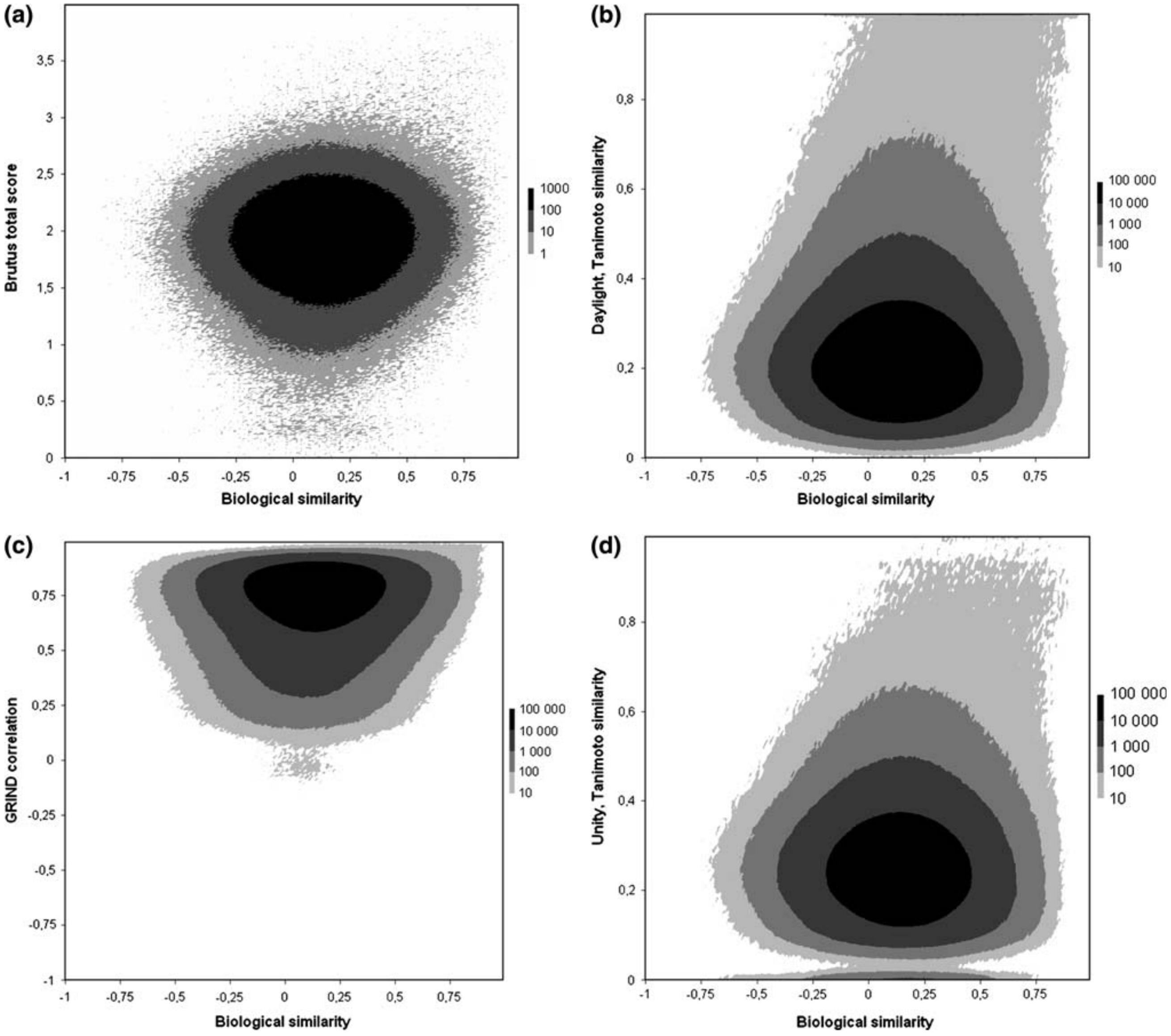

Fig. 1 **a** Distribution of molecule pairs as function of biological similarity (*x*-axis) and BRUTUS total score (*y*-axis). Skewness of the distribution towards the *top right corner* is less pronounced than with other measures of chemical similarity. This is due to smaller sample size for which BRUTUS score was calculated. Number of pairs in the illustration is 3,018,315. **b** Distribution of molecule pairs as function of biological similarity (*x*-axis) and Daylight Tanimoto similarity (*y*-axis).Here the distribution's skewness towards the top right corner is pronounced. Number of pairs in the illustration is 59,489,275. **c** Distribution of molecule pairs as function of biological similarity (*x*-axis) and GRIND chemical similarity (*y*-axis). Number of pairs in the illustration is 48,868,066. **d** Distribution of molecule pairs as function of biological similarity and UNITY Tanimoto score. As the Tanimoto score becomes larger, also the biological similarity tends to be larger than average. Number of pairs in the illustration is 38,322,428

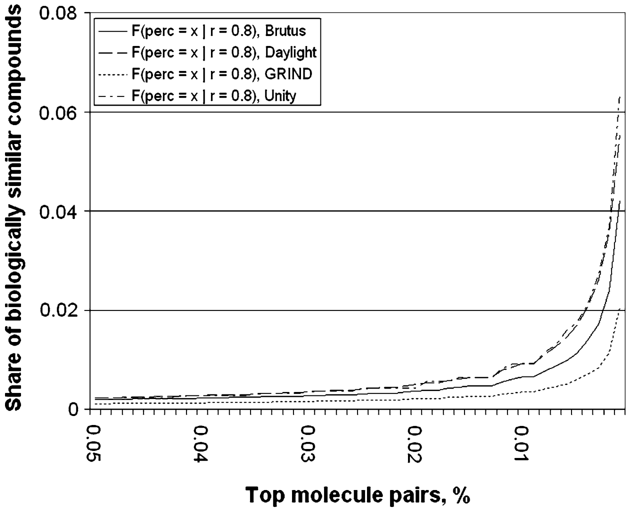

Fig. 2 Enrichment of biologically similar molecule pairs when the pairs are ranked by their chemical similarity with four methods. The threshold for biological similarity here is 0.8

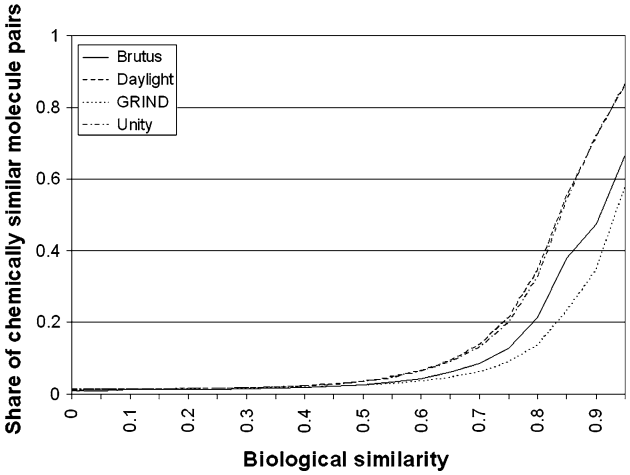

Fig. 3 The figure illustrates how the share of chemically similar pairs rises sharply as we approach the most biologically similar compound pairs. Analogously with results in Fig. 2, biological similarity best explains the 2D fingerprint methods (Daylight and UNITY) while the curves for GRIND and BRUTUS rise less sharply

For this the authors of DUD had used CACTVS type 2 fingerprints and chosen Tanimoto similarity of 0.9 as the threshold for chemical similarity. This leads to bias if another fingerprint method is used to enrich the known ligands. This is why we excluded UNITY fingerprints from the analysis.

A number of ranked lists were generated for each template molecule. The first two were based on the Brutus and GRIND similarities, respectively. For the third list, the probability of biological similarity was used to rank the molecules. The data for this came from a table calculated using Eq. 2 extended for two methods (GRIND and Brutus in this case). The table contained the probabilities of

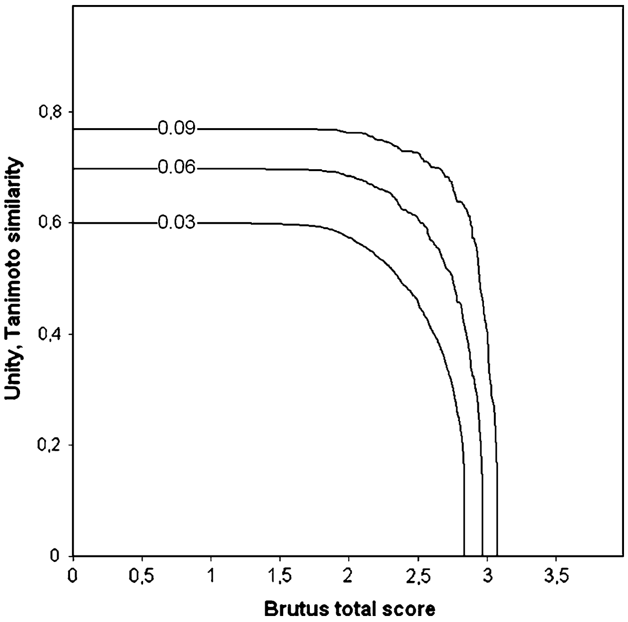

Fig. 4 The contour plot in the *top* visualizes share of biologically similar compound pairs when Brutus and UNITY scores are considered together. *Numbers* on the *contour lines* denote the share of biologically similar molecule pairs with the given BRUTUS and Unity similarity. *Curved* parts of the lines are those where synergy is gained from combining the two methods

biological similarity calculated for bins of Brutus and GRIND similarities using a set of thresholds for biological similarity (0.4; 0.5; 0.6; 0.7; 0.8 and 0.9). GRIND and Brutus similarities between the template ligand and the database ligand were used to query the table for the probability of biological similarity for the two molecules. This approach therefore serves as a data fusion method for combining the two chemical similarity types.

As reference data fusion methods, maximum and summation of rescaled chemical similarities were used [7, 20]. For both methods, all ranked GRIND and Brutus lists were rescaled from 0 to 1 using Eq. 5 where a denotes the original chemical similarity value, chem_max and chem_min the maximum and minimum similarity scores of the method, respectively. In the maximum data fusion method, the larger of the two rescaled values was chosen for each molecule. For the sum method, the rescaled values were summed. Last, the merged lists were reordered in decreasing order of the derived value.

$$a_{\text{rescaled}} = \frac{a}{\text{chem\_max} - \text{chem\_min}} \tag{5}$$

Next, the Receiver Operating Characteristic Area Under Curve (ROC AUC) was calculated for each ranked list [21]. If AUC equals one, all known ligands of a set are ranked before the decoy molecules. In contrast, if the value is zero all the decoy molecules are ranked before the known ligands. A value of 0.5 means random order where

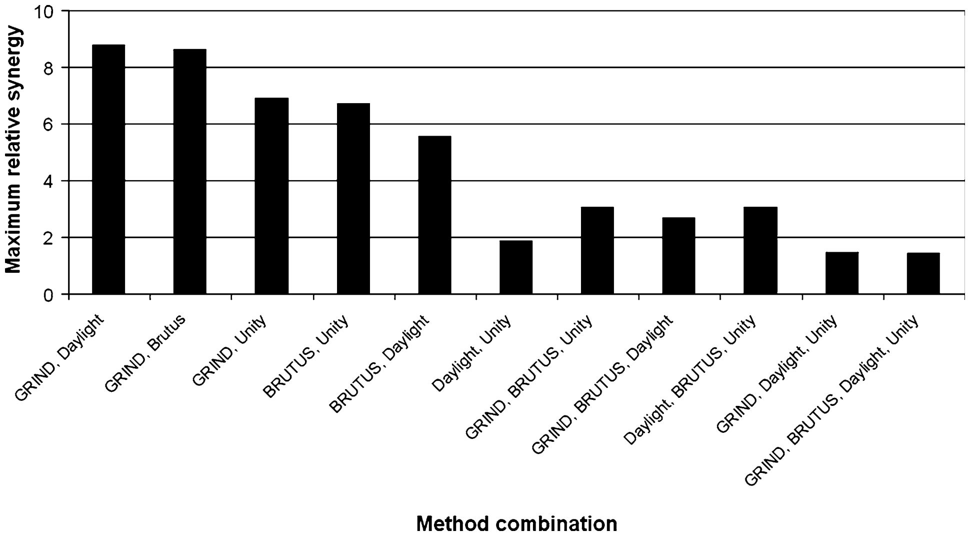

Fig. 5 Maximum relative synergies for different combinations of methods. The highest gains in synergy are gained when two methods are combined. Exception to this is pair Daylight/Unity where the relative synergy is modest. This is due to the similarity of the methods. Adding a third or fourth method has clearly a smaller impact on the synergy

there is no enrichment of known ligands. Also enrichment of known ligands at top 1, 2, 5, and 10% of the ranked lists were calculated.

Results

Distribution of molecule pairs

Figure 1a–d visualize the distribution of molecule pairs as function of biological similarity and four measures of chemical similarity. With all the metrics the distribution is skewed towards high biological similarity with increasing chemical similarity. This is most pronounced with Daylight, Unity and GRIND (Fig. 1b–d). The low level of skewness with BRUTUS (Fig. 1a) is due to the small sample size ($N = 3{,}018{,}315$). This leads to only relatively few cases where both the biological and chemical similarity would be high. Overall, the skewness of the distributions is a sign of positive link between biological and chemical similarity.

Chemical versus biological similarity with a single method

Chemical similarity methods are compared here for their ability to enrich biologically correlating compound pairs. Figure 2 illustrates how the share of biologically similar compound pairs (true positive rate) grows as we get closer to the most chemically similar pairs.

The two methods based on structure fingerprints (Unity and Daylight) outperform those based on molecular interaction fields (Brutus and GRIND). One possible explanation for this is that the biologically similar compound pairs share a common structure scaffold. Unity and Daylight fingerprints are suited better in identifying such pairs. Brutus and GRIND rely on the three dimensional structures, this requires the two molecules to have the right

Table 1 Median ROC AUC values and enrichment factors across all the 2,805 template molecules from the DUD data set

Method	Median enrichment, top 1%	Median enrichment, top 2%	Median enrichment, top 5%	Median enrichment, top 10%	Median ROC AUC
GRIND	10.02	6.74	4.01	2.72	0.65
Brutus	10.38	7.08	4.00	2.58	0.62
MAX	11.67	7.55	4.50	3.00	0.67
SUM	12.80	8.54	4.81	3.11	0.66
Bio_r04	13.54	9.26	4.94	3.18	0.67
Bio_r05	13.54	9.26	4.91	3.15	0.67
Bio_r06	13.54	9.26	4.91	3.15	0.66
Bio_r07	13.54	9.21	4.91	3.15	0.66
Bio_r08	13.49	9.21	4.91	3.15	0.66
Bio_r09	13.40	9.21	4.90	3.15	0.66
Top bio	13.54	9.26	4.94	3.18	0.67

Independent of the size of the ranked list considered, all data fusion methods produce improved median enrichment over individual methods. The performance of biological data fusion (Bio_r04–Bio_r09) is superior to MAX and SUM rules when only the top end of the list is considered

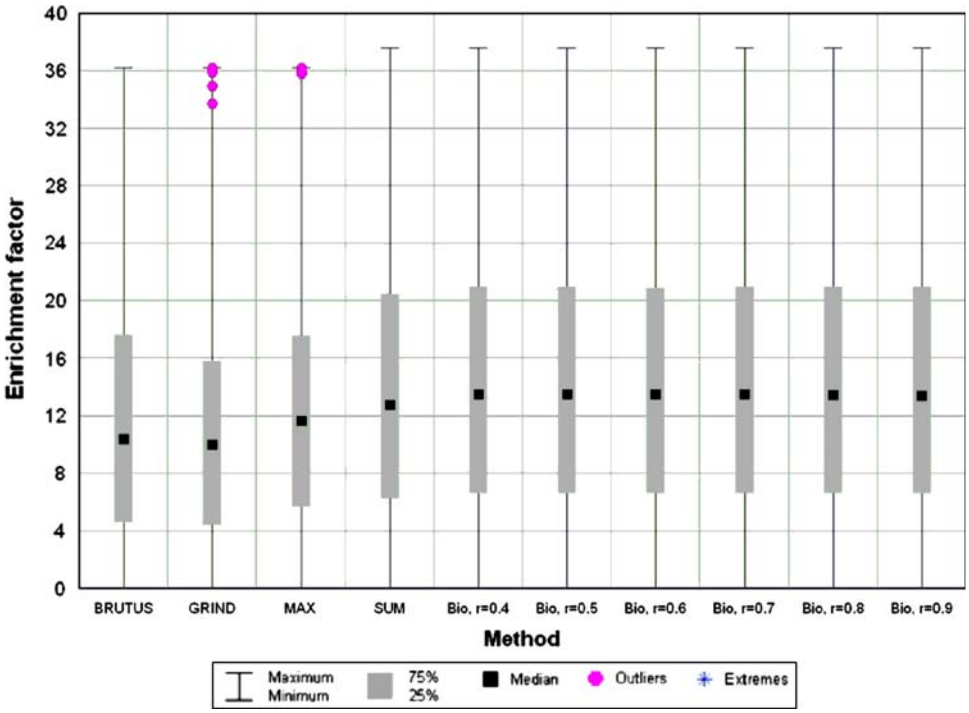

Fig. 6 Distribution of enrichment factors for the 2,805 ranking performed with external validation. Here only the top 1% of molecules are considered for each list. All data fusion approaches have higher median enrichment factors compared to either Brutus or GRIND used in isolation. Data fusion based on biological similarity estimation (Bio) has the highest median enrichment factors. The threshold for biological similarity has very limited effect to the results

conformation when compared. If this is not the case, very similar compounds might be classified as dissimilar. Indeed, 18.2% of the biologically similar compound pairs (threshold 0.8) share the same scaffold. This is in stark contrast with actives in the MUV dataset (http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html) where only 0.6% of pairs of actives share a scaffold.

It should also be noted that none of the methods reaches a true positive rate of one. This could be due to the general nature of the cytotoxicity measurement, the same endpoint can be achieved by multiple cellular response mechanisms. Another explanation is that even a perfect score with an individual method doesn't mean that the two molecules are identical. Either or both of the molecules might contain chemical features that the method doesn't recognize and which have an effect on the compound's activity. This problem can be partially alleviated by using more than one method as is shown.

Chemical similarity as function of biological similarity

The alternative way to look at the data is to plot the share of chemically similar compound pairs as a function of biological similarity (Fig. 3). For each of the four methods, the pairs with a similarity score in the top 1% were deemed as chemically similar. This threshold corresponds to a Brutus total score 2.591, Daylight Tanimoto score 0.492, GRIND score 0.920 and Unity Tanimoto score 0.50.

Just like Daylight and Unity fingerprint methods were able to enrich the biologically similar pairs better than Brutus and GRIND could (Fig. 2), ranking the pairs by biological similarity enrich the chemically similar compounds better when the chemical similarity is defined by either Daylight or Unity (Fig. 3).

Combining methods

Data fusion in virtual screening means combining results from several screening methods. This is known to give superior results compared to using a single method [7]. The approach is common practice in structure-based virtual screening where it is called consensus scoring. Since we have used four ligand-based methods in this paper, it is interesting to see what happens when we combine them, particularly the combination of a structure-based and a MIF-based method would appear reasonable.

Figure 4 visualizes the estimated biological similarity when Brutus and Unity similarities are combined. The biological similarity can be read from an intersection of Brutus total score and Unity Tanimoto similarity. This in effect is the share of biologically similar pairs over all pairs that have chemical similarity values equal to or greater than to the given Brutus and Unity scores. The plot has three contour lines corresponding to biological similarity values of 0.03, 0.06 and 0.09. The curved parts of the lines correspond to combinations of Brutus and Unity scores where synergy is observed.

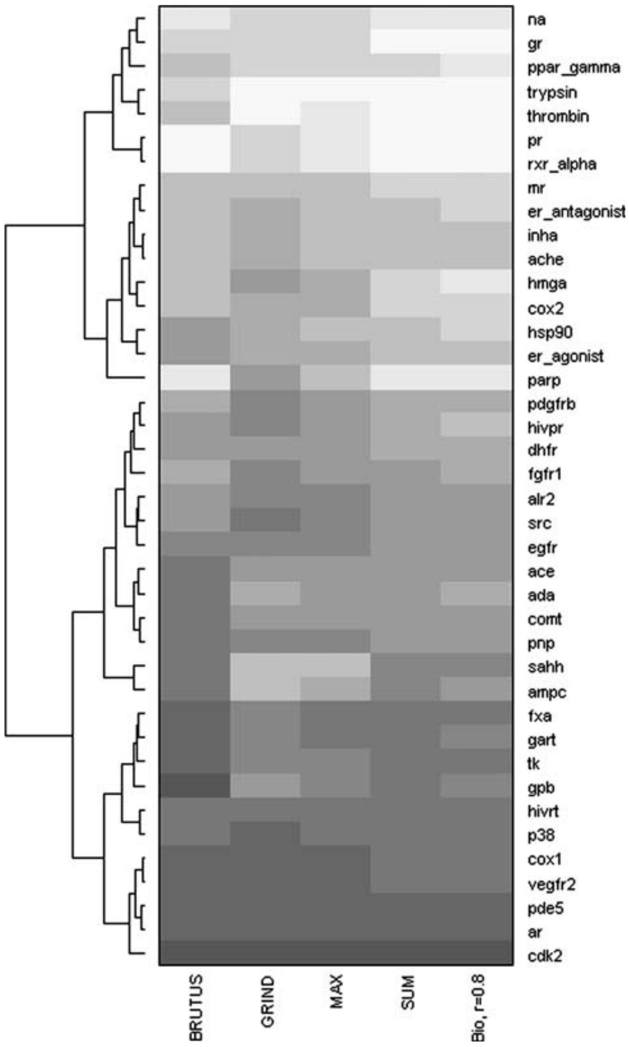

Fig. 7 The average enrichment factor at top 1% of each target class in DUD. The threshold for biological similarity used here for Bio ranking is 0.8. The target classes (*rows*) can be divided in four classes based on their enrichment factor profiles across the methods (*columns*). For detailed discussion of the groups can be found in the text. Whiter the cell in the heatmap, the higher the enrichment value. Target classes (*rows*) were clustered with hierarchical clustering employing Euclidian distance with complete linkage

An example of synergy is seen at the intersection of Brutus total score 2.40 and Unity Tanimoto similarity 0.49. Here the biological similarity estimate is $F(\text{brutus} = 2.40, \text{unity} = 0.49|r = 0.80) = 0.030445$. Taking only the Brutus score in account gives us $F(\text{brutus} = 2.40, \text{unity} = 0.00|r = 0.80) = 0.002106$ while considering only Unity score gives $F(\text{brutus} = 0.00, \text{unity} = 0.49|r = 0.80) = 0.009321$. Therefore, the relative synergy at this point is 3.26628 and the absolute synergy is 0.021124.

Figure 5 gives the maximum relative synergies for different combinations of methods when the threshold for biological similarity is set to 0.8. The highest values are for combinations of two methods. One major exception among these is the pair Unity–Daylight where the maximum relative synergy is only 1.88. This reflects the fact that both fingerprints are based on the same idea, i.e., iteration of molecule bond paths. Therefore they measure largely the same properties of the molecules. Maximum relative synergy drops again when a third or a fourth method is added in the combination. Obviously two methods already measure so diverse chemical properties that an additional method doesn't add anything dramatically new. These observations are similar also for the maximum absolute synergy and with different thresholds of biological similarity.

Directory of useful decoys (DUD)

Given chemical similarity of two molecules calculated with one or more methods, we are now able to estimate their biological similarity using the results above. Since we observed positive synergy when combining chemical similarity calculation methods, these results could also be applied to virtual screening. As described in Sect. "Materials and methods, chemical similarities of DUD known ligands were calculated with all the other molecules in the target set of the known ligand. Enrichment factors and ROC AUC values were calculated for each list derived this way.

Table 1 contains the median ROC AUC values and enrichment factors for the individual methods and their different combinations. The results are across all the individual 2,805 known ligand templates used. It is evident that when only the very top compounds are considered, we see improved results even for a large data set as this one. The median enrichment factors at top 1 percent are 10.02 and 10.38 for GRIND and Brutus, respectively. For the different data fusion methods, the value ranges from 11.67 to 13.54 meaning that we get synergy from combining GRIND and Brutus lists. Distributions of top 1% EF values are visualized in Fig. 6.

When one considers larger portions of the result lists, the synergy drops. For ROC curves, which take the whole list into account, the median AUC hardly differs between the two individual methods and the data fusion methods (Table 1). It is noteworthy that ranking based on biological similarity gives higher median enrichment factors than the MAX and SUM methods. The difference is largest when the top one percent of the lists is considered. The threshold for biological correlation hardly makes a difference.

The heatmap in Fig. 7 further divides the enrichment factors for the top 1% by the 40 target classes used. Here only one example of biological similarity ranking is shown (threshold 0.8) for clarity. The target classes are clustered across the vertical axis in roughly four groups.

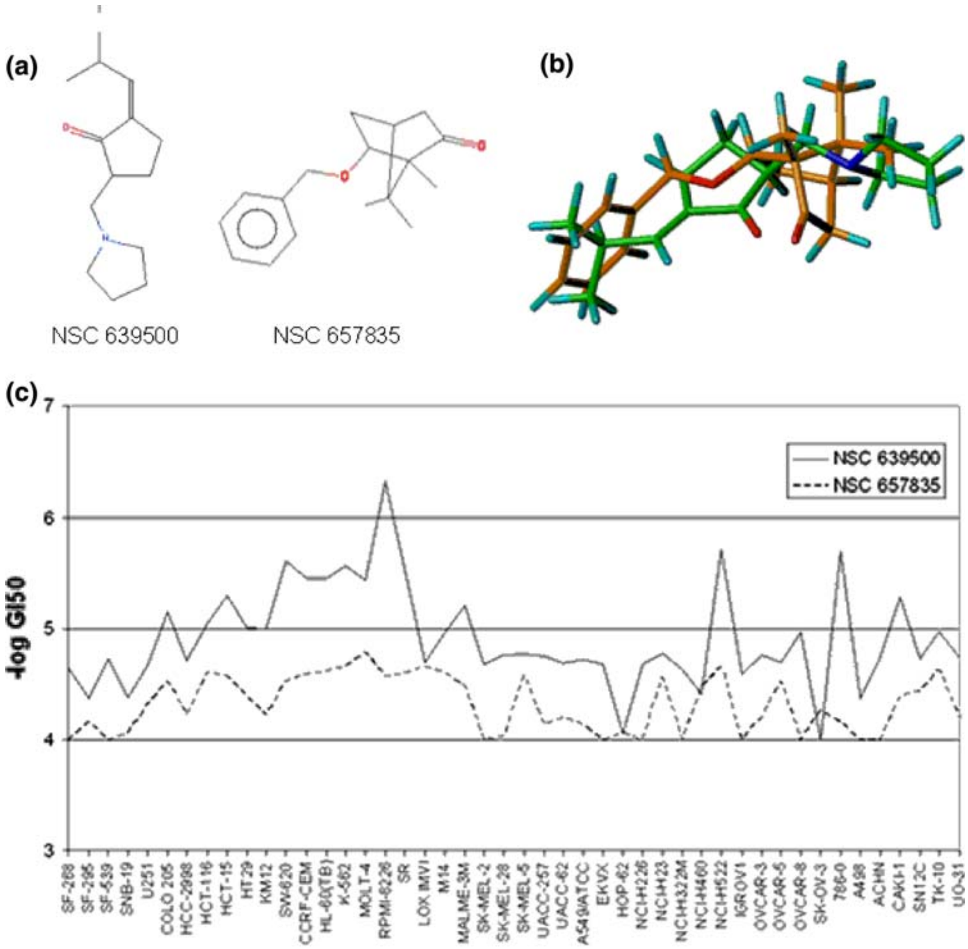

Fig. 8 **a** A molecule pair with high Brutus and GRIND but low Daylight and Unity similarity scores (Brutus = 2.960, GRIND = 0.915, Daylight = 0.227 and Unity = 0.25). **b** NSC 639500 (*green carbon atoms*) overlaid with NSC 657835 (*orange carbon atoms*). The good shape complementarity and similar orientation of the carbonyl groups explain the high Brutus score of the two molecules. **c** Cytotoxicity profiles of the two molecules across 46 cancer cell lines. The biological similarity is 0.546

The first group from the top, from Neuraminidase (na) to Retinoic X receptor alpha (rxr_alpha), contains targets for which either GRIND or Brutus suit well giving high average EFs. Combining the hit lists either with the SUM rule or with the biological ranking results in high EF values as well. Bio ranking is especially successful for ppar_-gamma giving an average EF of 22.3. For Brutus and GRIND, the average enrichment factors are 16.0 and 18.3, respectively.

The second group from the top (Mineralcorticoid Receptor, mr to Poly(ADP-ribose) polymerase, parp) contains target classes for which both Brutus and GRIND EFs are somewhat lower than for the first group (with the exception of parp for which the Brutus EF is a formidable 22.3). SUM and Bio ranking both are synergetic in all target groups except for parp where they have both have the same EF as Brutus alone. MAX rule's performance is worse in this group with even considerable negative synergy with certain targets (hmga and parp).

The average EFs get lower in the third and fourth groups of the heatmap. Within these groups very little to no synergy is gained with any of the fusion methods. In the third group, target classes sahh and ampc stand out as only the MAX rule is able to retain the enrichment given by GRIND descriptors.

In summary, if either GRIND or Brutus is able to enrich known ligands, the SUM rule and especially biological ranking are able to do so as well and even improve the enrichment.

Scaffold hopping

The quantity of active molecules (hit rate) one gets isn't the only objective in a virtual screening campaign. Also the scaffold diversity of the hits is an important criterion for project success. Having active compounds from several different scaffold classes is advantageous for a drug development project. If one compound fails to have a suitable bioavailability and safety profile, it is then possible to turn to other hits from different scaffold classes with possibly better properties. Also intellectual property issues play a role if one wants to develop a competitor to an already established drug [22].

Since MIF-based methods do not take the structure of the molecule into account explicitly, they are the obvious choice for scaffold hopping. To find cases of scaffold

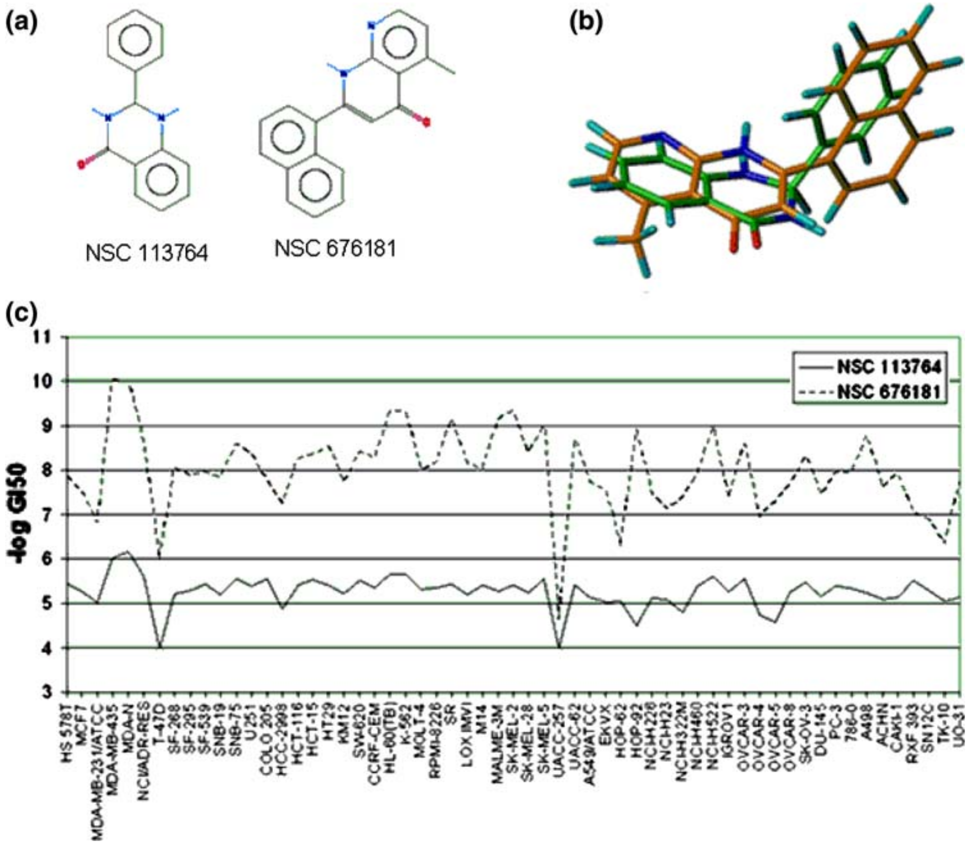

Fig. 9 **a** Two molecules with very similar scaffold structures that were missed by both Daylight and Unity fingerprints but scores similar by both GRIND and Brutus (Daylight = 0.155, Unity = 0.34, Brutus = 2.963, GRIND = 0.904). **b** Brutus overlay of the two molecules (NSC 113764 with *green carbons*, NSC 676181 with *golden carbons*). **c** Cytotoxicity profiles of the two compounds across 59 cancer cell lines. The biological similarity of the compounds is 0.743

hopping, we visually inspected the biologically correlating molecule pairs from the NCI dataset which were ranked similar by both GRIND and Brutus but dissimilar by both Daylight and Unity fingerprints. It was also interesting to see the cases where MIF-based methods failed to link biologically correlating compound pairs.

There were 23 molecule pairs which were biologically similar (GI50 profile correlation $\geq$ 0.50) and which had high scores with both MIF-based methods (GRIND score $\geq$ 0.90, Brutus total score $\geq$ 2.8) but the fingerprints methods didn't identify them as similar (Daylight $\leq$ 0.40, Unity $\leq$ 0.40).

These pairs could be divided in two groups. In the first are true scaffold hopping cases where the scaffold structures are clearly different. An example of this is the pair NSC 639500 and NSC 657835 (Fig. 8). The two molecules have a good steric overlaps plus they both have a hydrogen bond acceptor (carbonyl) which are overlaid and point to the same direction (Fig. 8b).

The second group includes compound pairs such as NSC 113764 and NSC 676181 where the scaffolds are very similar to the human eye (Fig. 9). The difference that causes the fingerprints to fail is the substitution of a nitrogen in NSC 113764 to a carbon in NSC 676181. Brutus can neatly overlay the molecules with a good score in spite of the differences (Fig. 9b).

Next we filtered the compound pairs with high biological similarity ($\geq$0.50), high Daylight and Unity scores (both$\geq$0.70) and low scores with Brutus and GRIND (Brutus $\leq$ 2.2, GRIND $\leq$ 0.850). This produced a list of 58 pairs.

Most of these pairs are cases where the other molecule is a substructure of the other. Naturally, in this case the larger of the two molecules has most of the same fingerprint features on as the smaller compound leading to a very high Tanimoto score. However, Brutus and GRIND scores are penalized heavily by the difference in sizes of the two molecules. An example of this is the pair NSC 302991 (Camptothecin) and NSC 681640 shown in Fig. 10. The former molecule is a substructure of the latter meaning that the fingerprints have no problem in matching them. The difference in size causes both Brutus and GRIND to miss the similarity. With Brutus, the overlay of the solution with the highest total score is a complete failure not even able to overlay the Camptothecin subgroup. However, the solution with the highest steric sub score is able to overlay the subgroups. Unfortunately, the other sub scores are low owing to the already mentioned size difference.

Second type of pairs is exemplified by pair NSC 642596 and NSC 657304 (Fig. 11). These compounds are very similar with the only differences being the two hydroxyl groups in the latter molecule and the nitrogen being

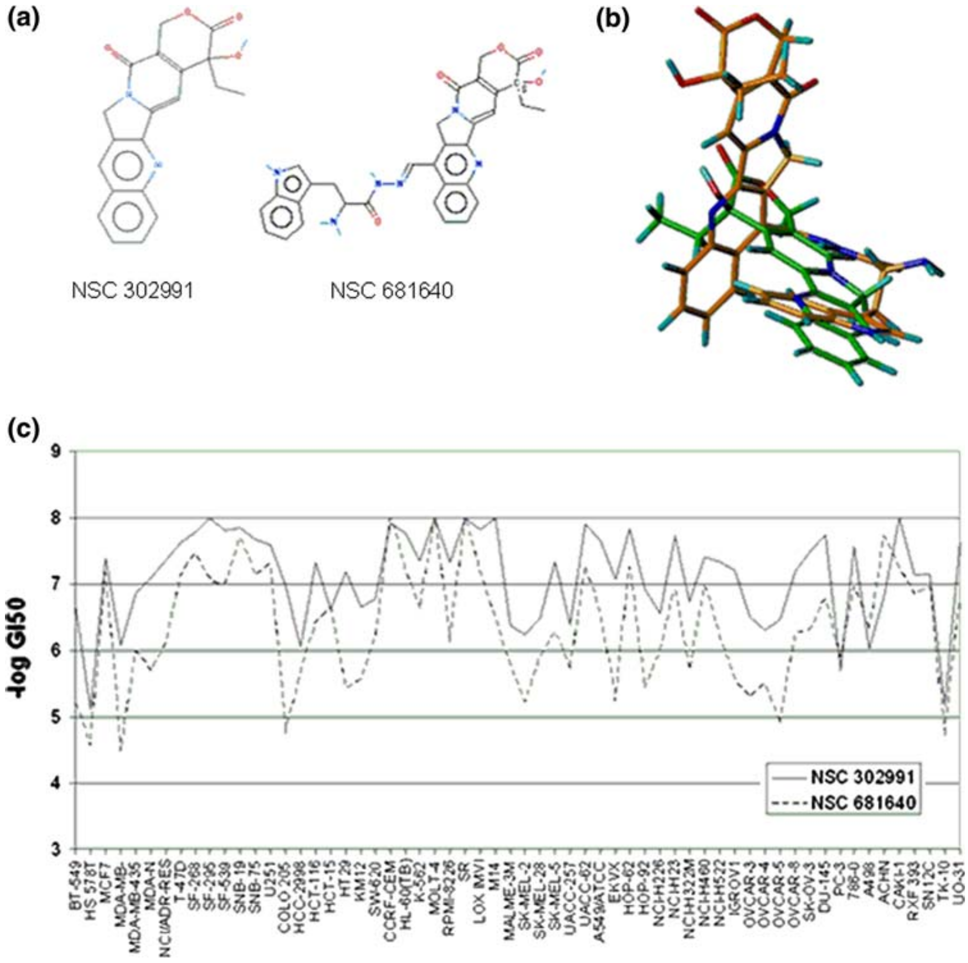

Fig. 10 **a** Camptothecin (NSC 302991) and its analogue compounds NSC 681640. Only the fingerprint methods were able to match these compounds as similar while both GRIND and Brutus failed to do so. Chemical similarity scores were: Brutus 1.923, GRIND 0.845, Unity 0.81 and Daylight 0.999. **b** Brutus overlay with the highest total score. NSC 302991 has *green carbon atoms* and NSC 681640 *orange carbon atoms*. Clearly the superposition is a failure. The superposition with the highest steric sub score was however able to correctly overlay the Camptothecin subgroups of the molecules. **c** Cytotoxicity profiles of NSC 302991 and NSC 681640 across 60 cancer cell lines. The biological similarity of the molecules is 0.765

charged in the former. Still both Brutus and GRIND give poor similarity scores. Brutus is able to overlay the molecules well but the above mentioned differences keep the total score low. Obviously, having similar shape is enough for activity in this case.

Conclusions

The results presented here are useful to a wide variety of applications where similarity of small molecules is required, most notably in virtual screening. Combining results from several methods leads to an improved hit rate as shown with the DUD dataset. Ranking molecules based on their predicted biological similarity gives superior synergy compared than can be obtained with the SUM/MAX rules. In addition to improved retrieval rate, the use of predicted biological similarity as criterion gives us a quantitative measure to decide which molecules to pick for further testing.

The data fusion has its limits however. Only combinations of two or three methods give us improvement over individual methods or combinations of their subsets (Fig. 5). The more different the logic behind two methods is the greater the synergy for combining the methods. This is perfectly logical as different methods measure different characteristics for the molecules and combining them gives us the good properties of both. Limiting the number of methods in use to a small number also leads to savings in software licensing costs.

Although BRUTUS and GRIND performed worse than the two fingerprint methods, their use shouldn't be ignored if novel chemotypes are sought as exemplified in Fig. 8. The choice of the method set depends on the objective of the work. If compounds built around the same scaffold structures are sought for lead optimization, a combination of fingerprint methods should be considered. Gaining synergy with more than one method can however be an issue as shown with the case of Daylight/UNITY fingerprints here. On the other hand, if scaffold hopping is desired to escape problems with legal and/or toxicology aspects, a combination of methods based on comparison of shape and interaction fields is recommended.

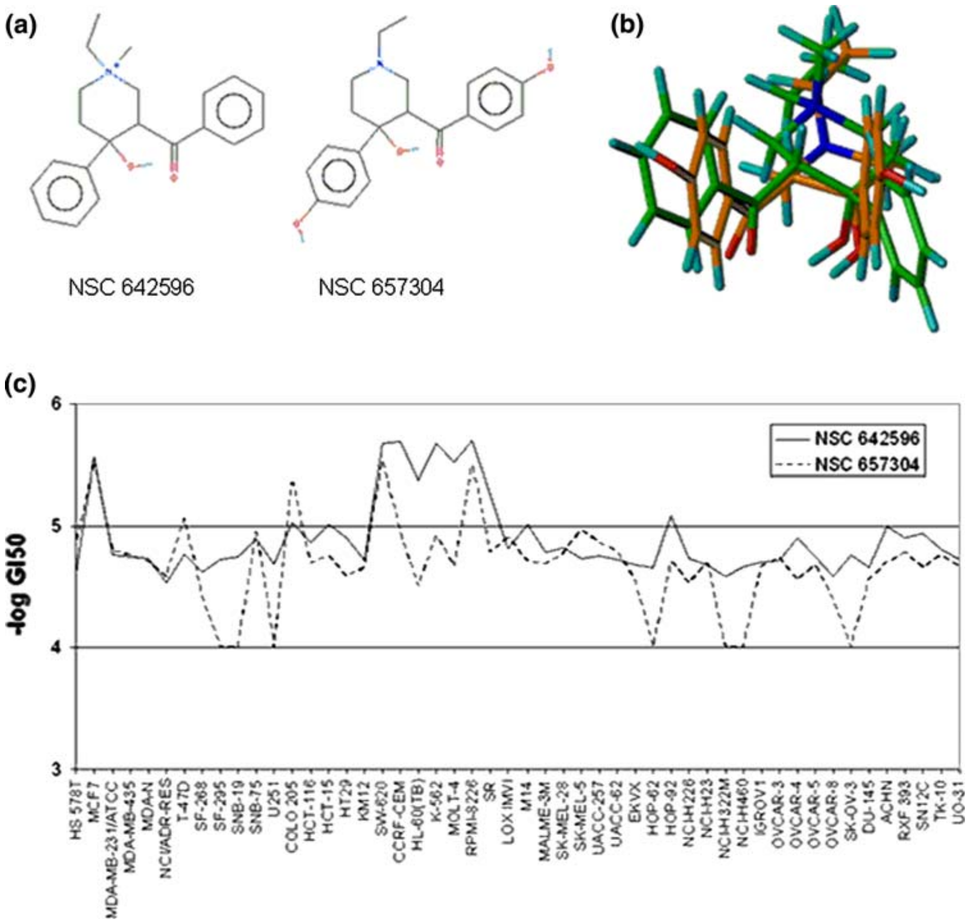

Fig. 11 **a** Two molecules which have similar size and scaffold. These two have high Daylight and Unity similarities but low scores with both GRIND and Brutus. (Brutus = 2.199, GRIND = 0.831, Daylight = 0.95 and Unity = 0.89). **b** Brutus overlay of the two compounds. The superposition is reasonable but structure differences in the two molecules cause the Brutus score to drop. Carbon atoms of NSC 642596 and NSC 657304 are colored *green* and *orange*, respectively. **c** Cytotoxicity profiles of the molecules across 50 cancer cell lines. Biological similarity of the molecules is 0.584

Recently, another article with a similar approach and conclusions was published by Muchmore et al. [8]. The authors had calculated chemical similarities with 10 different methods one of which was the Daylight fingerprints. In addition to the set of methods used, another major difference with our work was their definition of biological similarity which was based on a set of IC50 values measured against 23 protein targets for a compound library of more than 66,000.

Muchmore et al. had also found the synergy described in our work and compared it with synergies given by the MIN/SUM rules. They had also found out that the combinations of methods giving the most synergy are also that are based on different approaches. Like also our results show, they also had found little added benefit from combining different fingerprint based methods. Like with us, they had also validated their results with an external dataset of ligands.

Acknowledgments This study was supported by the Marie Curie Canceromics (MEXT-CT-2003-2728) grant from the EU, Cancer Organizations of Finland, Sigrid Juselius Foundation, and Academy of Finland (Centre of Excellence funding no. 213502). We also wish to thank Finnish IT Center for Science (CSC) for the computation resources used in this work. Anders Wallqvist is credited for sending the Daylight Tanimoto similarities and the biological similarities of the molecule pairs.

References

1. Martin YC, Kofron JL, Traphagen LM (2002) J Med Chem 45:4350
2. Willett P (2006) Drug Discov Today 11:1046
3. Goodford PJ (1985) J Med Chem 28:849
4. Pastor M, Cruciani G, McLay I et al (2000) J Med Chem 43:3233
5. Tervo AJ, Ronkko T, Nyronen TH et al (2005) J Med Chem 48:4076
6. Ronkko T, Tervo AJ, Parkkinen J et al (2006) J Comput Aided Mol Des 20:227
7. Willett P (2006) QSAR Comb Sci 25:1143
8. Muchmore SW, Debe DA, Metz JT et al (2008) J Chem Inf Model 48:941
9. Wallqvist A, Huang R, Thanki N et al (2006) J Chem Inf Model 46:430
10. Wang H, Klinginsmith J, Dong X et al (2007) J Chem Inf Model 47:2063
11. Covell DG, Wallqvist A, Huang R et al (2005) Proteins 59:403
12. Rabow AA, Shoemaker RH, Sausville EA et al (2002) J Med Chem 45:818
13. Wallqvist A, Huang R, Covell DG (2007) J Chem Inf Model 47:1414

14. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789
15. Benedetti P, Mannhold R, Cruciani G et al (2002) J Med Chem 45:1577
16. Cianchetta G, Singleton RW, Zhang M et al (2005) J Med Chem 48:2927
17. Cratteri P, Romanelli MN, Cruciani G et al (2004) J Comput Aided Mol Des 18:361
18. Gasteiger JR, Sadowski CJ (1990) Tetrahedron Comp Method 3:537
19. Fontaine F, Pastor M, Sanz F (2004) J Med Chem 47:2805
20. Whittle M, Gillet VJ, Willett P et al (2006) J Chem Inf Model 46:2206
21. Truchon JF, Bayly CI (2007) J Chem Inf Model 47:488
22. Brown N, Jacoby E (2006) Mini Rev Med Chem 6:1217