BIOINFORMATICS

Vol. 23 ECCB 2006, pages e104–e109
doi:10.1093/bioinformatics/btl292

Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family

Rafael J. Najmanovich[1,2,*], Abdellah Allali-Hassani[2], Richard J. Morris[1,†], Ludmila Dombrovsky[2], Patricia W. Pan[3], Masoud Vedadi[2], Alexander N. Plotnikov[2,4], Aled Edwards[2,5], Cheryl Arrowsmith[2,5] and Janet M. Thornton[1]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK, [2]Structural Genomics Consortium, [3]Department of Medical Biophysics, [4]Physiology Department and [5]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada M5G 1L6

ABSTRACT

Motivation: In the present work we combine computational analysis and experimental data to explore the extent to which binding site similarities between members of the human cytosolic sulfotransferase family correlate with small-molecule binding profiles. Conversely, from a small-molecule point of view, we explore the extent to which structural similarities between small molecules correlate to protein binding profiles.

Results: The comparison of binding site structural similarities and small-molecule binding profiles shows that proteins with similar small-molecule binding profiles tend to have a higher degree of binding site similarity but the latter is not sufficient to predict small-molecule binding patterns, highlighting the difficulty of predicting small-molecule binding patterns from sequence or structure. Likewise, from a small-molecule perspective, small molecules with similar protein binding profiles tend to be topologically similar but topological similarity is not sufficient to predict their protein binding patterns. These observations have important consequences for function prediction and drug design.

Contact: rafael.najmanovich@ebi.ac.uk

1 INTRODUCTION

A large fraction of cellular and biochemical processes involve the interaction between proteins (particularly enzymes) and small molecules (molecules other than nucleic acids and polypeptides). At a molecular level, these interactions are determined by the interplay of many factors including entropic effects and various molecular forces between interacting groups present in the protein and small molecule (Israelachvili, 1992; Leckband and Israelachvili, 2001). The term molecular recognition is used here as an amalgam of all these factors.

A detailed analysis of the enthalpic and entropic effects involved in molecular recognition requires extensive molecular dynamics simulations using more realistic force fields than those presently available (Becker *et al*., 2001). In their absence, the analysis of similarities can improve our understanding of molecular recognition by highlighting those structural elements that have the largest effect on the interactions between small molecules and proteins, thus providing a focal point for more detailed studies.

The extent to which similarities between non-homologous proteins, in terms of their overall sequence and structure, are related to similarities between their cognate ligands (substrates as well as cofactors) has been studied recently (Mitchell, 2001; Nobeli *et al*., 2005). Nobeli *et al*. (2005) report that a correlation between protein and ligand similarity can only be clearly established for very similar proteins. However, from an evolutionary point of view, while a pair of protein sequences might have diverged enough to be classified as non-homologous, the conservation of similarity between their cognate ligands constitutes evidence of homology between them.

A growing body of experimental evidence (Copley, 2003; O'Brien and Herschlag, 1999; Shears, 2004) suggests that catalytic promiscuity is integral to the function of various proteins. In recent years, the notion of binding promiscuity has also received support from computational docking studies (Koehler and Villar, 2000; Macchiarulo *et al*., 2004) where it has been observed that the complex formed between a given protein and its cognate ligand rarely is the most stable.

We use experimental data on the thermostability effect due to the binding of a small molecule to a protein as a measure of the strength of binding between the small molecule and the protein. The same data can be seen as (1) a set of small-molecule binding profiles to the proteins studied and (2) a set of protein binding profiles for each small molecule. In principle, unlike Nobeli *et al*. (2005), the set of ligands is not connected evolutionarily to any particular protein. Consequently, the binding patterns observed are not the result of any evolutionary optimization process.

The analysis of similarities in small-molecule binding profiles between members of a given family offers the possibility of comparing binding site similarities to small-molecule binding profile similarities within the evolutionary framework (i.e. overall sequence evolution) of the given family. Likewise, the comparison of small-molecule topological similarities and protein binding profiles can shed light in the relationship between structural similarity and molecular recognition from the point of view of small molecules.

The family of human cytosolic sulfotransferases (SULT) is involved in drug metabolism, detoxification and hormone regulation. The enzymatic reaction involves the transfer of a sulfuryl group from a donor molecule to a hydroxyl group in the acceptor

*To whom correspondence should be addressed.
†Present address: John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK

Table 1. Protein structures used in this study

Name[a]	PDB code[b]		Cleft Size[c]	Ligands[d]
1A1	1ls6A[e]	1.90	194 (51)	A3P, pNT
1A3	2a3rA[e]	2.60	177 (49)	A3P, LPD
1B1	1xv1A[f]	2.10	181 (51)	A3P
1C1	1zheA[f]	2.22	168 (43)	PAP
1C2	2ad1A[f]	2.00	67 (21)	—
2A1	1efhA[e]	2.40	275 (72)	A3P
4A1	1zd1B[f]	2.24	106 (32)	Unknown[g]

[a]The names of SULTs follow a scheme related to their pairwise sequence identity.
[b]Three-letter PDB codes are followed by the one-letter chain code of the subunit used in the study followed by the resolution in Ångstrons.
[c]Number of atoms in cleft (number of residues contributing atoms to the cleft).
[d]Relevant ligands present in the binding site. A3P (adenosine-3′,5′-diphosphate), pNT (p-nitrophenol), LDP (L-dopamine), PCQ (3,5,3′,5′-tetrachloro-biphenyl-4,4′-diol), NHE (*N*-cyclohexyltaurine), PLO (pregnenolone) and PAP (3′-phosphate-adenosine-5′-diphosphate).
[e]1ls6: (Gamage *et al*., 2003), 2a3r (Lu *et al*., 2005) and 1efh (Pedersen *et al*., 2000).
[f]Structures purified and solved by the Structural Genomics Consortium (SGC), to be published.
[g]The crystal contains an unknown *Escherichia coli* metabolite sequestered during expression.

molecule. In general the sulfuryl donor is 3′-phosphoadenosine-5′-phosphosulfate (PAPS). Substrate promiscuity is integral to the function of SULTs, particularly in detoxification and drug metabolism where the addition of the sulfuryl group increases the water solubility of the substrate thus facilitating its excretion. SULTs belong to the Rossmann fold group of proteins within the α/β class composed of a five-strand β-sheet flanked by a series of α-helices. The cofactor and substrate binding site lie on one side of the β-sheet with several loops forming a binding site 'trapdoor' that interacts with both cofactor and substrate. Trapdoor loop residues have been shown to be responsible for substrate specificity (Coughtrie, 2002). Various residues in the sheet interact with the substrate and some are essential for catalytic activity (Chapman *et al*., 2004; Glatt and Meinl, 2004).

2 METHODS

In the present study, we analyze a subset of selected members of the human cytosolic sulfotransferase family. Table 1 gives some information about the structures used.

2.1 Detection of clefts

Clefts are detected using the Surfnet algorithm (Laskowski, 1995). In short, a cleft is composed of a set of overlapping spheres. Each sphere is defined by a pair of atoms in the protein such that the radius of the sphere, placed at the mid-point of the line segment connecting the two atoms, lies between an upper and a lower bound radii threshold values without overlapping the van der Waals radius sphere of any protein atom. We use values of 1.5 and 4.0 Å for the lower and upper bound thresholds, respectively. In what follows, a cleft is defined as the atoms used to create the cleft spheres as well as their respective residue's C_α atom. The Surfnet algorithm defines many clefts of varying sizes for a given protein. While in general it might be difficult to detect the biologically relevant cleft (i.e. the cleft encompassing the binding site) (Glaser *et al*., 2006), in the cases analyzed in the present work, the biologically relevant cleft is known in advance. No information from bound ligands was used to restrict the size of the Surfnet cleft or to define residues as being part of the binding site.

2.2 Binding site sequence similarity

We created a multiple sequence alignment of the sequences of the proteins used in this study using HMMer (Eddy, 1998) and the pfam (Bateman *et al*., 2004) Sulfotransferase_1 (PF00685.15) hidden Markov model. Sequence similarity is measured using the Tanimoto coefficient (Gasteiger and Engel, 2003). The residues defining the cleft are mapped onto the sequence alignment and used to define a local Tanimoto score of sequence identity (LTS_{seq}):

$$LTS_{seq} = \frac{N_{comm}}{N_{Total} - N_{comm}}, \quad (1)$$

where N_{comm} represents the number of columns in the alignment where both sequences under comparison contain cleft residues of the same type and N_{Total} represents the sum of the total number of cleft residues in both proteins. The use of LTS_{seq} is a natural choice, since there is no straightforward way to define a sequence overlap to be used as normalization factor when comparing binding site residues, as these can be spread widely in the primary sequences of the proteins under comparison. For global sequence comparison, a global Tanimoto score of sequence identity (GTS_{seq}) is defined using all columns in the alignment (except those containing gaps for both sequences).

2.3 Binding site structural similarity

Given the sets of atoms defining the clefts under comparison, the question that needs to be answered is what is the largest subset of atoms in both clefts in direct correspondence with each other geometrically as well as chemically. This is a combinatorial optimization problem where, in principle, each possible set of atom correspondences might be a solution and the largest such set is the global solution. Graph theory offers a means to solve this problem via the detection of the maximal (largest) clique in an association graph. Further details on graph theory can be found elsewhere (Gross and Yellen, 2004). In the present work, we use the standard algorithm of Bron and Kerbosh (1973) for the detection of cliques.

Depending on the number of atoms being compared, the size of the association graph might make it practically unfeasible to detect the largest clique when considering all non-hydrogen binding site atoms. In order to overcome this difficulty we perform the graph matching in two stages.

In the first stage, an initial superimposition is performed via the detection of the largest clique in an association graph constructed using only C_α atoms of identical residues in the two clefts. A maximum distance difference of 2.0 Å is used to create edges in the association graph, imposing an upper bound of the same magnitude in the coordinates root mean square distance (RMSD) of corresponding C_α atoms.

Once the largest C_α clique is obtained its transformation matrix and translation vector are used to superimpose all atoms in the two clefts using the least square method of Arun *et al*. (1987) based on the singular value decomposition of the coordinates variance-covariance matrix.

In the second graph matching stage, all non-hydrogen atoms are used. Association graph nodes are created with the requirement that two atoms, one from each cleft, be of the same atom type as well as that their spatial distance be within 1.5 Å. This spatial distance constraint is used to decrease the size of the association graph and is the reason why the initial superimposition is performed. In the present work, we use eight atoms type classes (Sobolev *et al*., 1996, 1999) comprising the following classes: hydrophilic, acceptor, donor, hydrophobic, aromatic, neutral, neutral-donor and neutral-acceptor. Similar to the first stage, a maximum distance difference in defining association graph edges is used. This second threshold is set to 1.5 Å and again defines an upper bound of that magnitude in the RMSD between corresponding non-hydrogen cleft atoms.

Table 2. Aggregation temperature difference values, ΔT_{agg} (°C)

Small molecule[a]	Protein[b] 1C1	1C2	1B1	4A1	1A1	1A3	2A1	Σ[e]
01. DBHD[c]	10.8	0	11.4	6	5.4	3.5	8.4	*6*
02. DBHM[d]	9.3	0	7.4	3	3.6	0	4.9	*5*
03. Pyridoxal 5-phosphate	8.7	4.5	5.4	6	4.3	6.6	9	*7*
04. Resveratrol	6.6	0	8.6	NT	0	0	3.8	*3*
05. AMPPNP	4.5	5.2	4.5	0	4.4	5.6	3.4	*6*
06. PAP	4.5	6.4	7.8	0	6.7	6.6	7.4	*6*
07. Quercetin dihydrate	NT	4.6	8.4	10	6.5	NT	9.1	*5*
08. R-(-)-Apomorphine	0	0	12.6	NT	0	0	0	*1*
09. 4-Aminophenol	0	0	3.3	0	0	0	0	*1*
10. p-Cresol	0	0	0	7	0	0	0	*1*
11. (±)-Epinephrine	0	0	4.6	16	0	0	0	*2*
12. Isoprenaline	0	0	5	NT	0	0	0	*1*
13. (-)-Nor epinephrine	0	0	4.5	2	0	0	0	*2*
14. 2-Hydroxy estradiol	0	0	0	NT	0	0	3	*1*
15. Penta-chlorophenol	0	0	3.7	0	0	0	0	*1*
16. Lithocholic acid	0	0	0	0	0	0	4.4	*1*
17. Dehydroiso andro-steron 3-sulfate	0	0	0	0	0	0	2.8	*1*
Total[f]	*6*	*4*	*13*	*7*	*6*	*4*	*10*	

[a]Each molecule is referred in the text according by its numeric code. For example, H_01 refers to DBHD.
[b]Larger ΔT_{agg} values correspond to increased thermostability. NT: No T_{agg} observed, possible due to the T_{agg} in the presence of the molecule being beyond the measurement capability of the instrument.
[c]DBHD: (3,5-dibromo-4-hydroxy-benzoic acid (6,8-chloro-4-oxo-4H-chromen-3-ylmethylene)-hydrazide).
[d]DBHM: (3,5-dibromo-4-hydroxy-benzoic acid (6-chloro-4-oxo-4H-chromen-3-ylmethylene)-hydrazide).
[e]Number of different proteins binding a given small molecule (ignoring NT values).
[f]Number of different small molecules binding a given protein (ignoring NT values).

We define a Local Tanimoto Score of structural similarity (LTS_{3D}) to measure local structural similarity is as follows:

$$LTS_{3D} = \frac{V_{num}}{N_{Total} - V_{num}}, \quad (2)$$

where V_{num} represents the size of the largest clique (number of similar atoms in either first or second stage) and N_{Total} is the sum of the total number of atoms in each cleft being compared. The same measure is used to calculate local structural similarity considering only C_α atoms ($LTS_{3D\text{-}C\alpha}$) or all non-hydrogen atoms ($LTS_{3D\text{-}all}$). The normalization N_{Total} describes the total number of C_α atoms or the total number of non-hydrogen atoms in both clefts, respectively.

2.4 Small-molecule structural similarity

The similarity between small molecules was calculated as a Tanimoto coefficient using topological hashed fingerprints with linear paths of six bonds using ChemAxon's Jchem package (http://www.chemaxon.com).

2.5 Small-molecule binding profiles

A focused small-molecule library comprising 90 compounds including known substrates, products, inhibitors and other molecules with a high degree of similarity to the above was screened as to their effect in increasing the thermostability of each of the proteins tested. The assays were performed using the StarGazer technology. This apparatus measures protein aggregation via light scattering as a function of temperature within the 27–80°C range. For a given protein, an aggregation temperature (T_{agg}) can be extrapolated from the observed inflection point in a curve fitted onto the measured light scattering intensities. Aggregation temperatures are related to the melting temperature (T_m) of the protein and thus serve as a measure of the thermostability of the protein. A shift is observed in T_{agg} in the presence of saturating concentrations of a small molecule proportional to the ligand binding equilibrium constant due to the coupling of the protein denaturation reaction and the ligand binding equilibria (Shrake and Ross, 1990, 1992). In total, 17 molecules provide an increased thermostability effect ($\Delta T_{agg} > 2$°C) to any of the proteins tested (Table 2).

Members of the SULT family vary considerably in the number of ligands they bind (from 4 to 13). Likewise, the number of proteins that bind a given small molecule in the set also varies (1–7). An expanded and more in-depth description of the experimental work, including that involved in the purification and structural determination of various SULTs used here will be published elsewhere.

2.6 Clustering quality

The cophenetic matrix is a matrix where each element H_{ij} describes the height at which clustering subjects (i,j) are first merged together. The correlation coefficient between the cophenetic matrix and the dissimilarity matrix used to create the dendrogram, ρ, is a measure of the extent to which the dendrogram reflects the structure within the data (Everitt, 2001). This correlation value is used to select the clustering algorithm that produces the dendrogram that best describes the structure within data. All dendrograms in the present work were created (using the above criteria) with the average linkage method.

3 RESULTS

3.1 Binding site sequence versus structure

Given that the normalization factors N_{Total} used in Equations (1) and (2) (for $LTS_{3D\text{-}C\alpha}$) are identical, whenever LTS_{seq} and $LTS_{3D\text{-}C\alpha}$ are equal for a given pair of clefts, all identical binding site residues in the aligned sequences are found via structural comparison. In Figure 1 we show a plot of LTS_{seq} against $LTS_{3D\text{-}C\alpha}$ where each

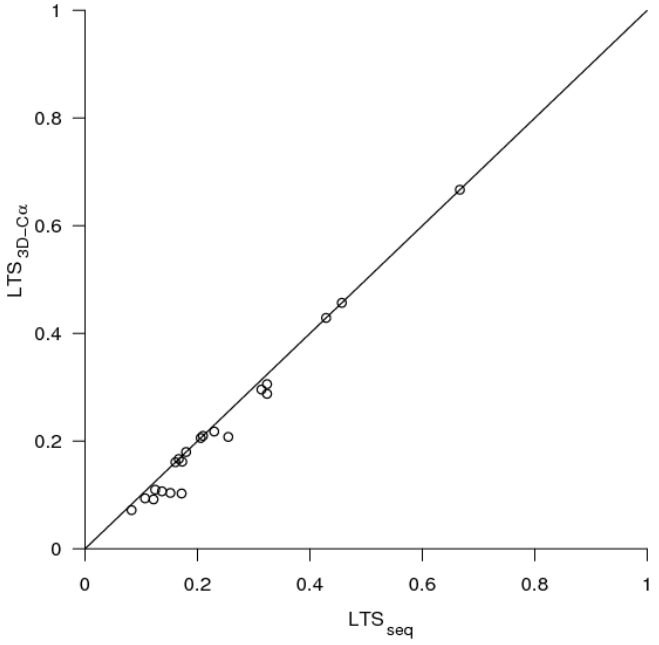

Fig. 1. Local sequence similarity (LTS_{seq}) against local C_α structural similarity (LTS_{3DCa}).

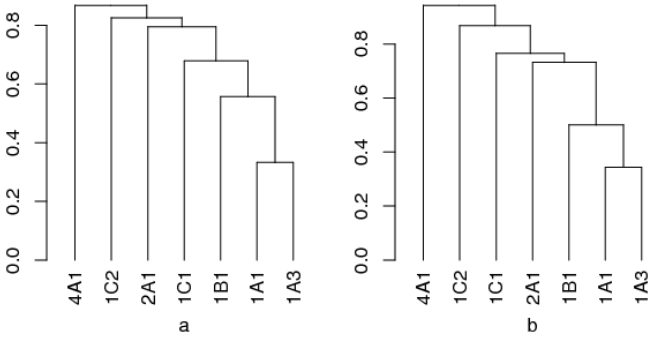

Fig. 2. Comparison of dendrogram of the SULT proteins based on binding site similarities. **(a)** Binding site sequence similarities (LTS_{seq}), $\rho = 0.977$. **(b)** Binding site all-atom structural similarities ($LTS_{3D\text{-}All}$), $\rho = 0.992$.

point represents one pairwise comparison. In this respect, it is important to note that the use of a carefully curated hidden Markov model to align the sequences is far more stringent than the use of pairwise sequence alignment methods, thus adding more significance to the success in locating equivalent residues through the pairwise structural comparison. The observations that do not fall on the diagonal involve primarily SULT4A1 and are due to local backbone movements (loop movements).

Figure 2 shows the dendrograms of the clustering of the SULT proteins in terms of their binding site sequence similarity (LTS_{seq}) or all-atom binding site structural similarity ($LTS_{3D\text{-}All}$). The dendrograms are very similar but that of $LTS_{3D\text{-}All}$ better represents the structure within the data ($\rho = 0.992$) than LTS_{seq} ($\rho = 0.977$). Table 3 contains a summary of the correlation between the various similarity measures of members of the SULT protein family.

Table 3. Correlation coefficients between SULT similarity measures

	LTS_{seq}	$LTS_{3D\text{-}C\alpha}$	$LTS_{3D\text{-}All}$	SmBP
GTS_{seq}	0.829	0.826	0.710	0.570
LTS_{seq}		0.992	0.938	0.386
$LTS_{3D\text{-}C\alpha}$			0.966	0.406
$LTS_{3D\text{-}All}$				0.334

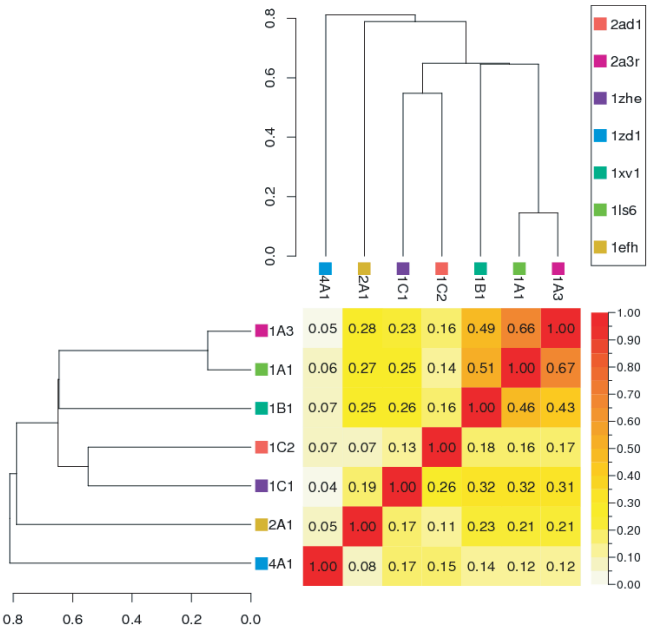

Fig. 3. Overall sequence compared with binding site sequence and structural similarity. Top and left dendrograms represent overall sequence identity (GTS_{seq}, $\rho = 0.997$), the upper half matrix (above diagonal elements) heatmap represents binding site structural similarities ($LTS_{3D\text{-}All}$) and the lower half matrix (below diagonal elements) represents binding site sequence similarity (LTS_{seq}).

Another question that can be inspected through the analysis of LTS_{seq} and $LTS_{3D\text{-}All}$ is that of the evolution of the binding site vis-à-vis that of the overall sequence within the family. In Figure 3 we present a clustering of the SULT protein sequences according to their pairwise global sequence similarity coefficients (GTS_{seq}).

The same plot shows the pairwise LTS_{seq} similarity coefficients in the lower half matrix (below diagonal elements) as a heatmap. The correlation between GTS_{seq} and LTS_{seq} or $LTS_{3D\text{-}All}$ is 0.829 and 0.710, respectively (Table 3), suggesting that binding site similarities are not well represented by overall sequence relationships. The average values of GTS_{seq}, LTS_{seq} and $LTS_{3D\text{-}All}$ are 0.30, 0.24 and 0.21, respectively, showing that binding site sequences are less conserved than overall protein sequences within the SULT family as binding site diversity brings about the diversity required for the specific function of the various members of the family.

3.2 All-atom comparisons and the effect flexibility

Side chain and backbone movements can have a strong influence on the atomic correspondences found through structural comparisons.

While a more relaxed choice of parameters used for the structural comparison can somehow account for the effects of flexibility, a balance must be struck between the increased computational resources required (time and memory) to solve the graph matching problem and the actual resources available.

The comparison of specific $LTS_{3D\text{-}All}$ and LTS_{seq} values (symmetric elements with respect to the diagonal in Fig. 3) does not show a clear tendency; in some cases, $LTS_{3D\text{-}All}$ is larger, in others, it is smaller. One major difference between $LTS_{3D\text{-}All}$ and LTS_{seq} is seen in the pairwise comparisons involving SULT4A1, where $LTS_{3D\text{-}All}$ values are much smaller than LTS_{seq}. The reason for this difference can be rationalized by the fact that the majority of the comparisons involving SULT4A1 shows large values of LTS_{seq}-$LTS_{3D\text{-}C\alpha}$ (data not shown), thus suggesting that the lower structural similarities between SULT4A1 and the rest of the dataset are due to the effects of backbone flexibility.

Furthermore, inspection of Figure 3 reveals that SULT2A1 is more similar in terms of binding site sequence and structure to the members of the SULT1 subfamily studied here (1A1, 1A3, 1C1, 1C2 and 1B1) than its classification in terms of overall sequence would suggest. This observation is still valid when all members of the SULT1 and SULT2 subfamilies are included in the analysis (data not shown).

3.3 Binding site structural similarities vis-à-vis small-molecule binding profiles

In the present analysis we make the necessary assumption that the thermostability effects observed experimentally (Table 2) are due to the interaction of the small molecules with the binding site and not other sites on the surface of the protein or regions that only become available as partially folded protein states become populated as a function of temperature.

The set of 17 small-molecule ΔT_{agg} values for a given protein can be thought of as a combined measurement of the effect that particular qualities of the protein's binding site has in its interactions with the set of small molecules. It is therefore interesting to compare the dendrogram obtained from the clustering of these ΔT_{agg} small-molecule binding profiles (SmBP, columns in Table 2) to the binding site sequence and structural similarities. Ideally, one would like to compare experimental and calculated ΔG values but neither are available.

In Figure 4 we present a dendrogram based on SmBP similarities and compare it with a heatmap matrix where the upper half of the matrix (above diagonal elements) represents binding site structural similarities ($LTS_{3D\text{-}All}$) and the lower half of the matrix (below diagonal elements) represents binding site sequence similarities (LTS_{seq}). The correlation coefficients between the SmBP distance matrix and LTS_{seq} or $LTS_{3D\text{-}All}$ are 0.386 and 0.334, respectively. Proteins with very similar small-molecule binding profiles show high levels of binding site similarity (1A3-1C2). It is interesting to note that while 1B1 shows a high degree of binding site similarity to both 1A3 and 1C2, its small-molecule binding profile is quite different from that of 1A3 and 1C2, pointing to the possibility that minor binding site differences may actually have a major impact in its interaction with different ligands. Furthermore, the small-molecule binding profile of 2A1 supports the observation made in the previous section that 2A1 is more similar to the SULT1 subfamily than its global sequence similarity would suggest.

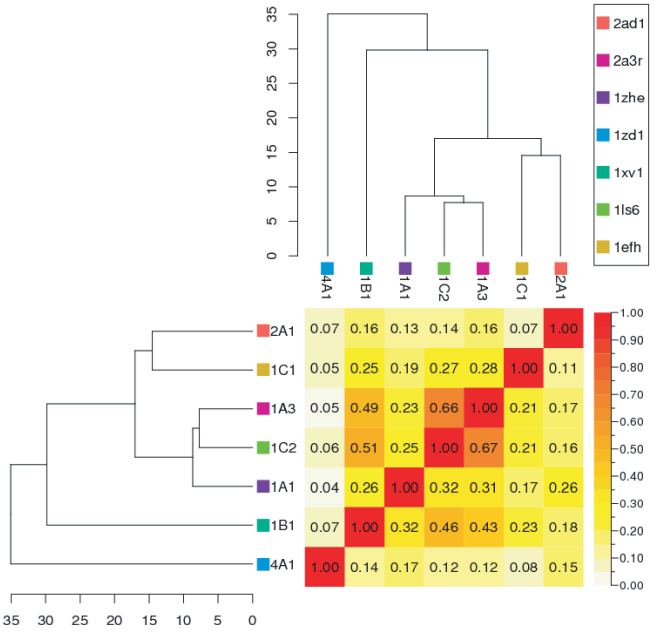

Fig. 4. Small-molecule binding profiles (top and left dendrograms, $\rho = 0.967$) compared with binding site sequence similarity (below diagonal matrix elements) and binding site structural similarity (above diagonal matrix elements).

3.4 Small-molecule structural similarities vis-à-vis protein binding profiles

The same ΔT_{agg} values presented in Table 2 can be analyzed from a small-molecule perspective. That is, for a given ligand one can utilize its pattern of binding to the set of proteins as a combined measure of the effect that particular characteristics of the small molecule affect its interactions with the set of proteins. Such protein binding profiles (PBP, rows in Table 2) can be used to cluster the set of small molecules to compare the resulting dendrogram to the matrix of pairwise ligand topological fingerprint (TFP) similarities (Fig. 5).

From a small-molecule point of view, it is again difficult to rationalize the binding profiles in terms of small-molecule structural similarities (0.253 correlation). Several but not all small molecules with similar binding profiles tend to show high degree of structural similarity.

4 CONCLUSIONS

The method for the detection of local structural similarities and the Tanimoto coefficients of sequence and structural similarity developed here are able in conjunction to detect binding site sequence and structural similarities between members of the human cytosolic sulfotransferase family.

Binding site sequence and structural comparisons uncovered similarities between 2A1 and members of the SULT1 subfamily not seen through overall sequence comparisons.

Proteins with similar small-molecule binding profiles show binding site sequence and structural similarities but the opposite is not true, namely, binding site similarity is not sufficient to predict the pattern of binding of different ligands to the given protein. From a computational point of view, this fact suggests that the accurate

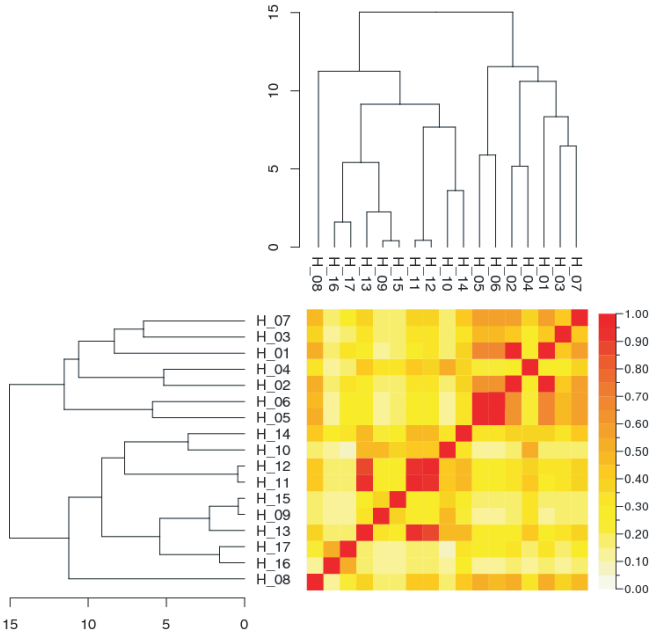

Fig. 5. Clustering of small molecules according to their protein binding profiles (PBP, $\rho = 0.775$) compared with pairwise topological hashed fingerprint similarity (TFP). In this case, symmetric elements with respect to the diagonal are identical.

prediction of cognate ligands from structure might not be possible until a better understanding of the process of molecular recognition is reached. From a pharmaceutical point of view, it is assuring that even similar binding sites show significant differences in the way they bind the same molecule as this diminishes the chances that a drug developed for a specific protein will have the same effect on related proteins.

Conversely, from a small-molecule perspective, a similar situation occurs. Namely, small molecules that bind with a similar pattern to a series of proteins tend to be topologically similar but topological similarity in itself is not sufficient to predict binding patterns. This observation suggests that we do not have a proper metric with which to gauge the effect that small topological differences can have in the binding pattern of a small molecule. While even related small molecules may bind to proteins in different orientations or with otherwise different interacting parts of their scaffolds, this deficiency of topological similarity as a metric is significant given its common usage in the selection of representative molecules for the creation of libraries in drug design, potentially leading to the oversight of important molecules.

Ultimately of course, we need accurate methods to calculate ΔG values associated to the binding of small molecules to proteins, but in their absence, simple similarity-based scores provide clues that can help cluster small molecules and protein according to their binding profiles.

ACKNOWLEDGEMENTS

The SGC is a public private charitable partnership that receives funds from Canada Foundation for Innovation, Canadian Institutes for Health Research, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, the Ontario Innovation Trust, the Ontario Research and Development Challenge Fund and the Wellcome Trust.

REFERENCES

Arun,K.S. *et al.* (1987) Least-squares fitting of 2 3-D point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, **9**, 699–700.

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Becker,O., MacKerell,E.,Jr, Roux,B. and Watanabe,M. (2001) *Computational Biochemistry and Biophysics*. CRC Press, New York & Basel.

Bron,C. and Kerbosch,J. (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.

Chapman,E. *et al.* (2004) Sulfotransferases: structure, mechanism, biological activity, inhibition, and synthetic utility. *Angew. Chem. Int. Ed. Engl.*, **43**, 3526–3548.

Copley,S.D. (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.*, **7**, 265–272.

Coughtrie,M.W. (2002) Sulfation through the looking glass—recent advances in sulfotransferase research for the curious. *Pharmacogenomics J.*, **2**, 297–308.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–766.

Everitt,B. (2001) Cluster analysis of subjects, hierarchical methods. In Armitage,P. and Colton,T. (eds), *Encyclopaedia of Biostatistics*. John Wiley & Sons, inc, Chichester New York, Weinheim, Brisbane, Singapore, Toronto.

Gamage,N.U. *et al.* (2003) Structure of a human carcinogen-converting enzyme, SULT1A1. Structural and kinetic implications of substrate inhibition. *J. Biol. Chem.*, **278**, 7655–7662.

Gasteiger,J. and Engel,T. (2003) *Chemoinformatics: A Textbook*. Wiley-VCH Verlag GmbH & Co., KgaA, Weinheim.

Glaser,F. *et al.* (2006) A method for localizing ligand binding pockets in protein structures. *Proteins*, **62**, 479–488.

Glatt,H. and Meinl,W. (2004) Pharmacogenetics of soluble sulfotransferases (SULTs). *Naunyn Schmiedebergs Arch. Pharmacol.*, **369**, 55–68.

Gross,J.L. and Yellen,J. (2004) *Handbook of Graph Theory*. CRC Press, FL.

Israelachvili,J. (1992) *Intermolecular and Surface Forces*. 2nd edn. Academic Press, London.

Koehler,R.T. and Villar,H.O. (2000) Statistical relationships among docking scores for different protein binding sites. *J. Comput.-Aided Mol. Des.*, **14**, 23–37.

Laskowski,R.A. (1995) Surfnet—a program for visualizing molecular-surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.

Leckband,D. and Israelachvili,J. (2001) Intermolecular forces in biology. *Q. Rev. Biophys.*, **34**, 105–267.

Lu,J.H. *et al.* (2005) Crystal structure of human sulfotransferase SULT1A3 in complex with dopamine and 3′-phosphoadenosine 5′-phosphate. *Biochem. Biophys. Res. Commun.*, **335**, 417–423.

Macchiarulo,A. *et al.* (2004) Ligand selectivity and competition between enzymes *in silico*. *Nat. Biotechnol.*, **22**, 1039–1045.

Mitchell,J.B. (2001) The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. *J. Chem. Inf. Comput. Sci.*, **41**, 1617–1622.

Nobeli,I. *et al.* (2005) A ligand-centric analysis of the diversity and evolution of protein-ligand relationships in *E.coli*. *J. Mol. Biol.*, **347**, 415–436.

O'Brien,P.J. and Herschlag,D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.*, **6**, R91–R105.

Pedersen,L.C. *et al.* (2000) Crystal structure of SULT2A3, human hydroxysteroid sulfotransferase. *FEBS Lett.*, **475**, 61–64.

Shears,S.B. (2004) How versatile are inositol phosphate kinases? *Biochem. J*, **377**, 265–280.

Shrake,A. and Ross,P.D. (1990) Ligand-induced biphasic protein denaturation. *J. Biol. Chem.*, **265**, 5055–5059.

Shrake,A. and Ross,P.D. (1992) Origins and consequences of ligand-induced multiphasic thermal protein denaturation. *Biopolymers*, **32**, 925–940.

Sobolev,V. *et al.* (1996) Molecular docking using surface complementarity. *Proteins*, **25**, 120–129.

Sobolev,V. *et al.* (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.