J Comput Aided Mol Des (2008) 22:111–118
DOI 10.1007/s10822-008-9178-7

FTree query construction for virtual screening: a statistical analysis

Christof Gerlach · Howard Broughton · Andrea Zaliani

Received: 31 July 2007 / Accepted: 10 January 2008 / Published online: 24 January 2008

Abstract FTrees (FT) is a known chemoinformatic tool able to condense molecular descriptions into a graph object and to search for actives in large databases using graph similarity. The query graph is classically derived from a known active molecule, or a set of actives, for which a similar compound has to be found. Recently, FT similarity has been extended to fragment space, widening its capabilities. If a user were able to build a knowledge-based FT query from information other than a known active structure, the similarity search could be combined with other, normally separate, fields like de-novo design or pharmacophore searches. With this aim in mind, we performed a comprehensive analysis of several databases in terms of FT description and provide a basic statistical analysis of the FT spaces so far at hand. Vendors' catalogue collections and MDDR as a source of potential or known "actives", respectively, have been used. With the results reported herein, a set of ranges, mean values and standard deviations for several query parameters are presented in order to set a reference guide for the users. Applications on how to use this information in FT query building are also provided, using a newly built 3D-pharmacophore from 57 5HT-1F agonists and a published one which was used for virtual screening for tRNA-guanine transglycosylase (TGT) inhibitors.

Keywords Feature trees · MDDR · ZINC · Query building · Database profiling · FTrees

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9178-7) contains supplementary material, which is available to authorized users.

C. Gerlach · A. Zaliani
Eli Lilly & Co. Research Laboratories, Essener Bogen 7, 22419 Hamburg, Germany

Present Address:
C. Gerlach
Bayer-Schering Pharma AG, Müllerstrasse 178, 13342 Berlin, Germany

H. Broughton
Discovery Chemistry Research and Technologies, Lilly Research Laboratories, Centro de Investigacion Lilly, Avenida de la Industria 30, 28108 Alcobendas, Madrid, Spain

A. Zaliani (✉)
Zentrum für Bioinformatik, Universität Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany
e-mail: zaliani@zbh.uni-hamburg.de

Introduction

In recent years, several chemoinformatic tools have provided convincing results in effectively retrieving actives from a random collection of molecules [1–7]. Using either fingerprinting methods coupled with similarity indexes, like Tanimoto [8], or taking advantage of 3D-search algorithms, chemoinformaticians working in drug discovery have several tools to perform one of their major tasks. FTrees (FT) is, among of these tools, quite a distinct method, as it uses graphs to describe molecules and a specific metric to compare them [9–11]. The feature tree descriptor differs in that it is a non-linear representation of the molecule. Instead of employing a bit string or vector, the molecule is represented by a tree structure. Thus, also the similarity index used by FTree differs in that it makes use of matching before to be calculated [9] as the trees are dependent from molecular size while a general fingerprint with definite length is not.

Using graphs to reduce molecular complexity has been demonstrated to be as effective as any other fingerprinting

method [12–14] and, in recent years, several fingerprinting methods have appeared in literature, which make use of graph features [2, 3, 5, 15]. The success of such methods is not only due to their efficacy in extracting important information about the atom connections but also, more importantly, to their ability to maintain pharmacophoric information and, at the same time, to assure sufficient fuzziness to allow similarity metrics to rank as "similar" what seem to be, at first sight, different structural motifs. Thus, graph-based chemoinformatic approaches have raised a lot of interest for their scaffold-hopping potentials, due to this property [2].

Standard protocols in virtual screening (VS) require that users, after having identified one or more actives, use them to rank a database of molecules [16]. This can be done by finding, at the top of the ranked list, compounds with the highest probabilities to be "similar" to the query structures, thus enhancing the probability of finding other actives. This has been the classical way FT has been used so far, and still will be used. However, we were interested in following a parallel but lateral approach.

If a user does not have a classical query (i.e. an active structure), but general information which could be eventually translated into a query graph, like, for instance, diverse molecules which suggest sparse pharmacophoric information, it might be still useful to perform a FT query based on these info. Actually the user cannot take these information and directly build a FT query, as he/she needs to know how to build query graphs. We thought that this task can be performed by hand, provided that users know which are average values of the most relevant characteristics defining a common drug-like feature tree query. One can, for instance, know that an active has a certain substructure or a set of pharmacophoric points at a certain distance in space or in bonds, or that must have a particular shape given a certain conformation from NMR and/or X-ray experimental data [17, 18]. In these cases what can be done to search for similar compounds? Can we build a FT query from general substructure and/or from a set of pharmacophoric points and their distances? MTree approach can be seen as a pioneering effort in this direction [19]. There, a set of known active compounds were converted into a sort of consensus tree, called MTree-model, for further similarity searches by means of FT. In principle, we are not here aiming to a more performant similarity method. On the contrary, we aimed to broaden the applicability of FT searches by taking advantage of the features of known FT Spaces. Along this lines we addressed a generalized statistical analysis of FT-Spaces known.

Molecule databases have been investigated either coming from pure synthetic collections or from annotated drug-like database like MDL Drug Data Report (MDDR) [20]. A couple of examples on how to build query graph from the reported statistical description are finally given.

Materials and methods

Data sets

The datasets within this study were obtained from two different sources, namely the ZINC-database [21] and MDDR [20]. From ZINC release 2006 the druglike (2066906 molecules), leadlike (643959 molecules) and fragment (56995 molecules) subsets were downloaded as mol2-files [22]. A subset of MDDR (156169 molecules) with a maximal molecular weight of 800 Da was compiled. This cutoff was chosen to avoid polypeptides but to maintain larger organic annotated drugs. ZINC subsets were used with the provided molecular weight cutoffs of 500 Da for druglike, 350 Da for leadlike and 250 Da for fraglike, following the definition given for each subset from the ZINC homepage [22]. File conversion for MDDR from sdf to mol2 was carried out by means of Corina [23]. For comparison of FTrees vs. molecular weight and molecular VdW-Volume the appropriate values were calculated using the descriptors "vdw_vol" and "weight" as implemented in MOE 08.2006 [24]. FT conversion for each molecule in the four different datasets was performed using the program FTrees v.1.5.1 [25]. The conversion was carried out using standard settings included in the default config.dat file and the corresponding static data files.

FTree statistics

FTree statistics were derived at three resolutions. At first, all properties of the whole FTree were extracted and analyzed. In a second step, the characteristics for each node were considered. Finally, the bond-properties of each FTree were described. For the value-extraction a PYTHON-script was used [26]. Statistical analysis was performed using the scientific module SciPy [27].

Query building from a newly developed 3D-pharmacopore model

A set of 57 compounds actives as 5HT-1F [28] agonists were used to build a pharmacophoric model within PHASE [29, 30] as implemented in Maestro v.7.5 [31]. Of the 60 tested molecules present in the paper, three smaller compounds (no. 14a, 28 and 29 in the original numeration) were excluded (see supplementary materials). We wanted to explore only a qualitative pharmacophore common to

actives and we did not want to execute a complete 3D-QSAR study selecting a pharmacophore also quantitatively predictive. Each molecule was evaluated in its protonation state at pH = 7.0 and its conformational space was sampled using the OPLS2005 force field and GB/SA solvation model with a max of 100 conformers within 10 Kcal/mol from minima or an RMSD deviation of 2.0 Å or more from each other. No activity threshold was given and the pharmacophoric recognition was performed using default PHASE rules. Extraction of best pharmacophore model allowed matching of 51 actives with 5 pharmacophoric points (one Acceptor, one Donor, one positively charged and two aromatic rings lying in the same plane) as depicted in Fig. 1.

The suggested pharmacophore was translated manually into a FTree query with the following procedure which will be used for the second example as well. Starting from the xyz-coordinates of the pharmacophoric points and their properties assigned by PHASE, we simply joined the points directly one another trying (1) not to build cycles and (2) to use the shortest bond and their minimal amount. It must be here noted that we performed a 100% translation of pure pharmacophoric nature assigned by PHASE to FT nodes. While this operation can be debatable, due to the mixed nature of pharmacophoric content of FT-nodes, we found it consistent with the overall practical spirit of the approach. After this initial step, the refinement of the FTree query with adding extra-nodes, deleting existing ones or changing the node properties was guided by the derived statistical values obtained by our analysis (see Results and Discussion). Additional nodes which were clearly not proposed by PHASE pharmacophore, but which could be added without qualitatively disturb the overall hypothesis, were then manually introduced in accordance with average distances and properties derived from the statistical analysis. All the four manually constructed FTrees for this example were used as a query for a FTree search against the 57 5HT-1F compounds described by Filla et al. [28]. For each query the similarity value at level x [32] was extracted and the average similarity value was calculated. The FTree search was carried out by means of FTrees v.1.5.1 using standard settings included in the default config.dat and the corresponding static data files.

Query building from a structure-based derived pharmacophore

A structure based derived pharmacophore published by Brenk et al. [33] served as a starting point to generate different FT queries. This hypothesis was used as a pharmacophore filter step within a VS protocol to detect TGT inhibitors.

As in the case of the 5HT-1F example, each pharmacophore point was extracted by using the xyz coordinates and the physicochemical properties as annotated by Brenk et al. [33]. The conversion was also guided by the derived statistical values (see Results and Discussion).

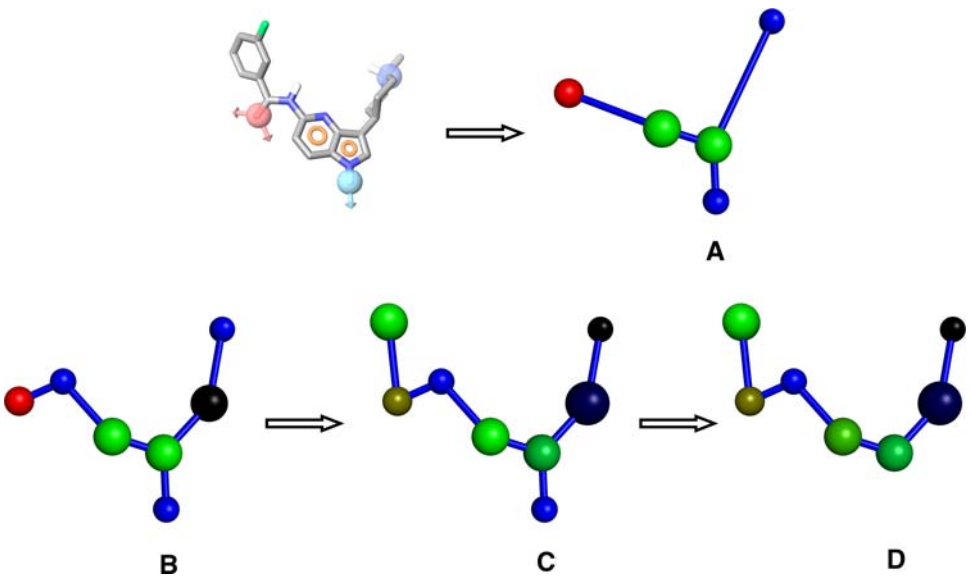

Fig. 1 Molecule 3 h (in the original numeration of [28]) with derived pharmacophore by means of PHASE. Aromatic centers are shown as orange torus, H-bond-acceptor center is colored in red, H-bond donor centers are shown in blue and cyan. (**a**) The manual constructed feature tree derived from the suggested PHASE pharmacophore. (**b**) Extended FTrees guided by the derived statistics: The amount of feature nodes is increased, considering the derived statistics. Incorporating additional nodes are guided by the average distances between two feature nodes. (**c**, **d**) Donor functionalities are included in a ring node by raising the donor nature of the ring component

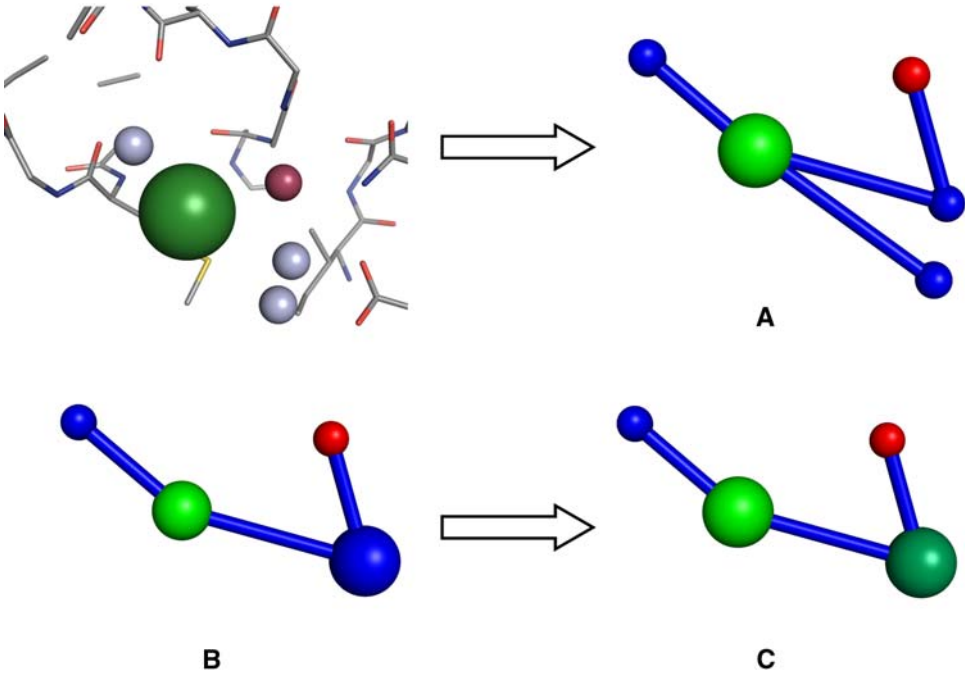

Fig. 2 Structure based pharmacophore surrounded by aminoacids of the binding pocket published by Brenk et al. [33] (**a**) Derived FTree through direct translation (**b**) Considering the closed distances between the two donor centers the nodes are merged to one ring node with ring properties. (**c**) Adding further hydrophobic properties to the node describes the node in its natural way

Three different FT queries (Fig. 2) were constructed to perform a FT search against the retrieved hits with biological activity (compounds 3–16 [33]) against TGT published by Brenk et al. Afterwards the corresponding similarity values were determined and the average similarity value was calculated. Also here the FT search was carried out by means of FTrees v.1.5.1 using standard settings included in the default config.dat and the corresponding static data files.

Results and discussion

Statistical analyses

The results of the statistical analyses of the four datasets are reported in Tables 1 and 2, respectively. The first table collects data referring to FT graphs, i.e. is directly linked to the parent molecule structure. In Table 2, instead, information are presented regarding the features contained in each node. So, for instance, column 3 (Mean Volume) is referred to FT graph in Table 1, while to nodes in Table 2. We also reported standard deviation as a sort of confidence value for the user and not implying that their distribution is unimodal.

Data and trends in Table 1 are, in general, known and parallel to those presented in previous work [34, 35]. This is also due to the fact the several descriptors tracked are not reciprocally independent. However, in FT space, some detailed information are buried within the concept of graph node. As expected [34], the mean molecular weight (MW) is higher for MDDR than for all ZINC subsets, due to the limits set for each of them. Also the mean molecular volume is in the range of what can be found for a ligand with crystallographic evidence: it should be below 400 A^3 as reported by Laurie et al. [36]. However, only molecules coming from leadlike and fraglike selection of ZINC and only parts of druglike and MDDR collection are within this boundary. This finding ensured us that the overall molecule set we chose was good enough to represent a set of possible ligands and that their average description paralleled what is already known such as the "Rule of five" [37], "Rule of three" [38] and leadlike definitions [35]. Moreover, we identified new useful average values, which we successfully used to select suitable query trees to be tested in search experiments against actives, and to provide further insights on how different commercial and bioactive collections look like under FTree description.

So, for instance, the average number of recognized nodes within each molecular graph is 6–11 and, according to graph theory, the number of bonds connecting them has to be one less. Interestingly, the number of rings is decisively restricted between 1.6 and 3.4 with the highest value shown by the MDDR. This peculiar aspect might be explained either by the amount of combinatorial libraries contained in vendors' catalogues, which tend to overemphasize a scaffold or substructure, or by the higher MW. The slightly higher number of branches present in MDDR can speculatively be interpreted as a result of optimized sidechain decoration on known active scaffolds. This argument can be reinforced by looking at the average numbers of FT-branches found in the three different vendor

Table 1 Statistical analyses of main characteristics of features trees derived from four molecular databases

Source	MolWeight[a]	Std Dev	MolVol(A^3)[b]	Std Dev	NofNodes[c]	Std Dev	NofisRing[d]	Std Dev
Druglike	341.80	59.78	439.90	81.84	9.23	2.84	2.75	0.96
Fraglike	207.73	31.34	276.84	53.92	5.83	2.29	1.61	0.77
Leadlike	293.06	44.94	377.78	65.40	8.22	2.60	2.32	0.86
MDDR	465.08	262.51	521.61	278.90	11.13	10.43	3.44	1.51

Source	NofBranches[e]	Std Dev	Bond distance (A)[f]	Std Dev	isRING_isRING[g]	Std Dev	isRING_isnotRING[h]	Std Dev	isnotRING_isnotRING[i]	Std Dev
Druglike	0.96	0.87	2.31	0.79	3.07	0.88	2.91	0.39	1.64	0.34
Fraglike	0.56	0.70	2.22	0.74	2.92	0.90	2.84	0.33	1.64	0.36
Leadlike	0.80	0.80	2.27	0.76	3.02	0.88	2.89	0.35	1.63	0.32
MDDR	1.03	1.71	2.51	0.81	3.12	0.91	3.02	0.38	1.73	0.35

[a] Molecular Weight of the parent molecule; [b] molecular Volume of the parent molecule; [c] number of Nodes per FTree; [d] amount of FTree-Nodes which are rings; [e] amount of Branches in the FTree; [f] distance between two FTree nodes; [g] distance between two FTree nodes which are rings; [h] distance between one FTree node with ring and one without ring functionalities; [i] distance between two FTree nodes which are not rings

Table 2 Statistical analyses of main characteristics of features per FT-nodes derived from four molecular databases

Source	Mean-Size[a]	Std Dev	Node-Vol(A^3)[b]	Std Dev	Mean-Radius[c]	Std Dev	Mean-NoFeature[d]	Std Dev	Mean-isRing[e]	Std Dev
Druglike	2.57	2.18	35.80	28.34	1.91	0.52	4.95	4.46	0.30	0.47
Fraglike	2.45	2.15	36.25	28.64	1.92	0.52	4.63	4.46	0.29	0.48
Leadlike	2.49	2.12	34.65	27.42	1.89	0.51	4.89	4.43	0.29	0.46
MDDR	2.94	2.65	32.15	28.42	1.84	0.50	3.13	1.07	0.32	0.50

Source	Mean-Nof-Neighbours[f]	Std Dev	Mean-Nof-NeighboursIsRing[g]	Std Dev	Mean-Nof-NeighboursIsNotRing[h]	Std Dev	Mean_Nof_hydrophobic[i]	Std Dev	Mean-ratio Hydrophobic_NofFeatures[j]	Std Dev
Druglike	1.78	0.67	0.55	0.60	1.23	0.83	2.80	4.37	0.33	0.42
Fraglike	1.66	0.71	0.47	0.56	1.18	0.87	2.32	4.16	0.30	0.42
Leadlike	1.76	0.67	0.52	0.58	1.24	0.84	2.58	4.2	0.31	0.41
MDDR	1.82	0.63	0.54	0.65	1.27	0.84	1.46	2.28	0.72	0.39

[a] Mean size of a feature node; [b] mean volume of a feature node; [c] mean radius of a feature node; [d] mean number of features given by the chemical description; [e] amount of feature nodes which is annotated as ring per FTree; [f] mean number of features connected to the considered node; [g] mean number of features which are rings and connected to the considered node; [h] mean number of features which are rings and connected to the considered node; [i] mean number of hydrophobic properties per feature node; [j] mean ratio between hydrophobic features and all features

sets The branch average value goes down from 1.03 to 0.96 to 0.80 to 0.56 for corresponding MDDR, druglike, leadlike and fraglike database, respectively. This represents more than a 10% loss for each vendor set going from arguably more (complex) to less branched (simple), and was somehow expected even if not in such size This is, in our opinion, one of the most relevant differences among the sets we found.

Moreover, in FT, the number of features per node is dependent on the amount of pharmacophoric relevant atoms, while the volumes are dependent on all atoms (comprised hydrogens). We suspect the difference we saw within databases for these values may be due to higher number or higher diverse heterocycles arguably present in MDDR. Vendor catalogues are notoriously biased towards easy-accessible building blocks and not towards heterocyclic systems with probable higher synthetic complexity. This might also be the reason why the ZINC fragment set showed highest mean node volume together with the mean smallest node size. The lack of pharmacophoric atoms within nodes worked in this direction, reducing mean size of nodes while raising their mean volume. For instance, it is known that in FT space, higher halogen content (Cl, Br and I) is a cause for higher MW (and volume) while number of features per node remains unchanged. Only fluorine is considered potential hydrogen-bond acceptor and raises the number of features per node while leaving rather unaltered its average volume. This behavior is

clearly dependent on the mean number of *pure* features per node, which is consistently lower for the annotated drug collection See, for instance, the number of pure hydrophobic features per node in Table 2. It is another dramatic difference between known active molecules and commercial ones.

We also tracked whether differentiation exists between number of nodes and feature types contained in them, against node and feature type contained *around* them (e.g. Nof-Neighbours, Nof-Neighbours-Is-Ring and Nof-Neighbours-Is-not-Ring in Table 2).

Here, the total amount of neighboring features is preferentially *ring* in vendors' catalogues while being *not-ring* for MDDR (see Table 2 bottom set of columns). Sizes of FT-features are important, as they collect more or less atoms depending on the pharmacophoric content of the node [11].

Using these statistical results, one can speculate that fragmenting MDDR, for instance, could lead to more "druggable" fragments than those obtained shredding vendors' catalogues. Certainly, the possibility of searching in fragment space raised interest as to which is the best source of "druggable" fragments [39]. Here we showed that, indeed, there are source-dependent differences in both levels of description: at graph and at graph nodes. It is clear, on the other side, that fragment constitution depends heavily not only on the set of parent molecules but also on the process of shredding them. Here we refer only to differences inherent to parent molecules.

Also interesting we found that the average amount of neighbors per node varies with the sources and, not surprisingly, the lowest amount can be found in a fragment-like vendor catalogue, while MDDR showed the highest number of neighbors. This fact can also suggest that research strategies towards certain biological targets, and the associated subtype selectivity quest, can strongly influence the topology of the scaffold or chemotype used to find the desired activities, as suggested by Vieth et al. [34]. We just saw that the distribution of hydrophobic features shows impressive differentiation between vendors' collections and MDDR. This is not surprising *per se* as chemical syntheses and separation of the products in an industrial environment can be easier with more lipophilic compounds. However, MDDR showed the highest percentage of lipophilic character per node (Table 2) which may indicate that lipophilicity in annotated drugs is "spread" all over the nodes with less *pure* (100%) hydrophobic nodes. This mirrors the above discussion on the mean synthetic feasibility of annotated drug being inferior to vendors' molecules. Having more "pure" hydrophobic nodes in their FTrees, ZINC molecules showed thus a lower ratio of hydrophobes' percentage per feature.

Interestingly, Hessler et al. [19] showed that embedding any 2D molecular graphs in a bi-dimensional hexagonal pattern where each atoms, or group thereof, overlay hexagons, has a valid impact in FT search. This is the visualization of M-Tree approach above mentioned. We now found that an hexagonal pattern has a solid statistical root. The hexagonal scheme implicitly suggests that a distance between two nodes should be around 2.38 Å. In fact, taking an hexagon layout having 1.375 Å long edges (the mean length of a normal C[sp3]–C[sp3] and C[sp2]–C[sp2] bond), the resulting distance among hexagon centers is approx 2.38 Å. This is exactly what we found with an average inter-node distance of 2.35 Å.

Applications of statistical information to FT Searches

Finally, to put these findings into practice, we used them to support the construction of an FT query for two test cases We wanted to test whether a FT query built from a set of 3D information (like a pharmacophore or a protein cavity) generates similarity scores in the same ballpark of a classical FT query built from real molecules. We focused both on a newly generated simple 5-point pharmacophore derived from 5HT-1F data set [28] and from a structure-based derived pharmacophore model published by Brenk et al. [33].

In the first case, we operated refining initial 5-point query, well below the average number of nodes (five vs. an average of 6–11). This way, we included some nodes among the five suggested ones from PHASE. Doing so, we did not reduce the original distances between the nodes suggested by PHASE, as they laid externally. See, for instance, between positive node and the first aromatic node in Fig. 1a,b where distance was 5.57Å, quite double than the statistical significant one found by our analysis (2.3–2.5 Å). So we generated four different FT queries (Fig. 1) inserting convenient nodes at the convenient distances. It is also clear that the pharmacophoric nature of added node in this particular case was dictated by reference active compounds, while in reality users could not know anything about it. However, the limited amount of nodes may allow the generation of several geometrically similar graphs with extra-nodes of different nature. For instance, PHASE was not able to generate 3D-pharmacophores aligning the extra aromatic ring attached with carboxamides or sulphonamides to the indole central moiety. We thought that this happened either for the limited conformational sampling used by PHASE or because we did not force the program to deliver a quantitative model through the pharmacophore (see Materials and Methods). The SAR tables clearly showed the importance of an extra-aromatic group in that position, so we added an extra aromatic (green) node to

Table 3 Calculated averages, Std Dev and median of the similarity value at level x [32] against 57 5HT-1F actives for the four manual constructed feature trees and for the three manual constructed feature trees derived from a structure based pharmacophore for TGT inhibitors

	5HT-1F				TGT		
	Feature tree query A	Feature tree query B	Feature tree query C	Feature tree query D	Feature tree query A	Feature tree query B	Feature tree query C
Average Similarity	0.6530	0.7481	0.8056	0.8316	0.6515	0.7172	0.7891
Standard Deviation	0.0480	0.0276	0.0470	0.0475	0.0645	0.0721	0.0735
Median	0.6630	0.7450	0.8193	0.8444	0.6405	0.7011	0.7824

query 1B to produce 1C. We also tried to improve the similarity by removing the hydrogen-donor node attached to indole ring and raising the donor nature of the aromatic node (see 1C query to 1D). FT queries produced and their corresponding retrieving results against the 57 actives are reported in Table 3. As expected, average similarity increased from 0.6530 to 0.8316 going from a low-defined out-of-average FT query (like a pure pharmacophore) to a more consistent 8-node FT query. These similarity values cannot be compared with those coming from a classical FT search as the query is far less detailed in this case. However, FT search starting from a poorly detailed query tree can be optimized and results pushed to more meaningful levels of chemical similarity.

In the second case (Fig. 2), a published structure-based pharmacophore served as a starting point. Also here, the FT query construction was guided by the distance between each node. Like in the first case, the pure conversion into a FT query did not lead to an optimal query. By the connection of the structure-based derived points (Fig. 2a) the average similarity is 0.6515. Furthermore, we merged the two donor points to one ring system, as suggested by higher distance average (Fig. 2b). This modification resulted into slight improvement to an averaged similarity value of 0.7172. The incorporation of aromatic properties to this ring function led to an increase of the average similarity value to 0.7891 (Fig. 2c). Even if in this latter case the final queries are well below the average in terms of number of nodes we were satisfied by the average similarity values found starting from a so limited structure-based pharmacophore.

We noticed from both cases, that at least three parameters have a real impact upon FT query building. The first is the minimal number of nodes, which has to be within an average of at least 7-8 nodes. This does not imply that limited number of informative nodes can be effective, though, as shown in Fig. 2. Secondly, the inter-node distances are relevant, and usually they are not respected by an abstract 3D-pharmacophore. Here, the correction is more complicated as infinite graphs can be generated interpolating two points quite far each other. In the present tutorial case, though, the distance was such that only one extra node could be interposed, but, in principle, there might be a degeneracy problem in the type of path one could use to connect two points. Finally, the inner nature of a node can be really decisive, as shown in Table 3. One can find better similarity eliminating a 100% donor node on the azaindole ring, which was a precise part of the initial pharmacophore, and incorporating it in the five member ring node, raising the donor nature of this node (query D from query C in Fig. 1). This characteristic of nodes makes FT query building from a pharmacophore more difficult than previously expected, but it might open new tuning possibilities as all pharmacophores are usually generated as pure points (e.g. 100% charged or not charged; 100% acceptor or donor). They are defined this way as they are derived by rules but, in theory, they bring any weight (as partial charges, for instance). Better results incorporating external charges in FT searches have been already reported [3].

Conclusions

We have performed a statistical analysis of FT description of four different databases. The aim was not only to show differences in FT space among databases, but mainly to have an overview of the average values and thresholds of FT description in order to help building FT queries.

Some of these thresholds are new, other already directly and indirectly known. Among the former those which were followed to build the ZINC subsets, such as the “Rule of five” [38], “Rule of three” [39] and leadlike definitions [35], can now be directly translated in FT space parameters. Among those not yet known, we found a definitive good complementarity between our findings and the MTree approach and, moreover, interesting node properties which definitively separate MDDR from commercial databases.

Finally, guided by these statistical values, practical manual adaptations of a pharmacophore query generated by PHASE and a published pharmacophore from Brenk et al. have been shown to become a feature tree for FT similarity searches. This manual adaptation enables derivation of FT queries not only from known active

molecules, but also from different sorts of information. We provided here some thresholds values for main FT descriptors which can be used to fine-tune FT query building and thus similarity searches.

Looking to the future, with respect to combinatorial FT fragment space searches, the upper and lower boundaries of feature tree properties can be incorporated using the statistical analysis performed here.

Acknowledgements We are in debt to Joerg Degen and Prof. Matthias Rarey (ZBH, Hamburg) for their suggestions and to the reviewers for their constructing criticism.

References

1. Willett P (2006) Drug Discov Today 11:1046
2. Stiefl N, Zaliani A (2006) J Chem Inf Model 46:87
3. Stiefl N, Watson IA, Baumann K, Zaliani A (2006) J Chem Inf Model 46:208
4. Fechner U, Paetz J, Schneider G (2005) QSAR Comb Sci 24(8):961
5. Gillet VJ, Willett P, Bradshaw J (2003) J Chem Inf Comput Sci 43:338
6. Bajorath J (2002) Nat Rev Drug Discov 1:882
7. Bajorath J (2002) Drug Discov Today 7:1035
8. Tanimoto TT (1957) IBM Internal Report
9. Rarey M, Stahl M (2001) J Comput Aided Mol Des 15(6): 497–520
10. Rarey M, Dixon JS (1998) J Comput Aided Mol Des 12:471
11. Bender A, Mussa HY, Gill GS, Glen RC (2004) J Med Chem 47:6569
12. Evers A, Hessler G, Matter H, Klabunde T (2005) J Med Chem 48:5448
13. Bissantz C, Schalon C, Guba W, Stahl M (2005) Proteins 61:938
14. Schneider G, Schneider P, Renner S (2006) QSAR Comb Sci 25:1162
15. Renner S, Schneider G (2006) ChemMedChem 1:181
16. Klebe G (2006) Drug Discov Today 11:580
17. Hartshorn MJ, Murray CA, Cleasby A, Frederickson M, Tickle IJ, Jhoti H (2005) J Med Chem 48:403
18. Hajduk PJ, Huth JR, Fesik SW (2005) J Med Chem 48:2518
19. Hessler G, Zimmermann M, Matter H, Evers A, Naumann T, Lengauer T, Rarey M (2005) J Med Chem 48:6575
20. The MDL Drug Data Report is available from MDL Information Systems Inc., San Leandro CA
21. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 45:177
22. http://blaster.docking.org/zinc/subset1 (accessed Dec 2007)
23. Gasteiger J, Rudolph C, Sadowski J (1992) Tetrahedron Comput Method 3:537
24. http://www.chemcomp.com (accessed Dec 2007)
25. FTrees 1.5.1, http://www.biosolveit.de/FTrees (accessed Dec 2007) available from BioSolveIT, St. Augustin, Germany
26. http://www.python.org (accessed Dec 2007)
27. http://www.scipy.org (accessed Dec 2007)
28. Filla SA, Mathes BM, Johnson KW, Phebus LA, Cohen ML, Nelson DL, Zgombick JM, Erickson JA, Schenck KW, Wainscott DB, Branchek TA, Schaus JM (2003) J Med Chem 46:3060
29. Dixon SL, Smondyrev AM, Rao SN (2006) Chem Biol Drug Des 67:370
30. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) J Comput Aided Mol Des 20:647
31. http://www.schroedinger.com (accessed Dec 2007)
32. http://www.biosolveit.de/FTrees/download/ftrees_ug.pdf (accessed Dec 2007)
33. Brenk R, Naerum L, Gradler U, Gerber HD, Garcia GA, Reuter K, Stubbs MT, Klebe G (2003) J Med Chem 46:1133
34. Vieth M, Siegel MG, Higgs RE, Watson IA, Robertson DH, Savin KA, Durst GL, Hipskind PA (2004) J Med Chem 47:224
35. Teague SJ, Davis AM, Leeson PD, Oprea T (1999) Angew Chem Int Ed 38:3743
36. Laurie AT, Jackson RM (2005) Bioinformatics 21:1908
37. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Adv Drug Deliv Rev 46:3
38. Carr RA, Congreve M, Murray CW, Rees DC (2005) Drug Discov Today 10:987
39. Rarey M, Stahl M (2001) J Comput Aided Mol Des 15:497