J Comput Aided Mol Des (2009) 23:655–668
DOI 10.1007/s10822-009-9285-0

Analysis and use of fragment-occurrence data in similarity-based virtual screening

Shereena M. Arif · John D. Holliday · Peter Willett

Received: 9 February 2009 / Accepted: 19 May 2009 / Published online: 18 June 2009

Abstract Current systems for similarity-based virtual screening use similarity measures in which all the fragments in a fingerprint contribute equally to the calculation of structural similarity. This paper discusses the weighting of fragments on the basis of their frequencies of occurrence in molecules. Extensive experiments with sets of active molecules from the *MDL Drug Data Report* and the *World of Molecular Bioactivity* databases, using fingerprints encoding Tripos holograms, Pipeline Pilot ECFC_4 circular substructures and Sunset Molecular keys, demonstrate clearly that frequency-based screening is generally more effective than conventional, unweighted screening. The results suggest that standardising the raw occurrence frequencies by taking the square root of the frequencies will maximise the effectiveness of virtual screening. An upper-bound analysis shows the complex interactions that can take place between representations, weighting schemes and similarity coefficients when similarity measures are computed, and provides a rationalisation of the relative performance of the various weighting schemes.

Keywords Fingerprint · Fragment occurrences · Ligand-based virtual screening · Similarity searching · Substructural fragment · Tanimoto coefficient · Virtual screening · Weighting scheme

S. M. Arif · J. D. Holliday · P. Willett (✉)
Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK
e-mail: p.willett@sheffield.ac.uk

Introduction

There is much current interest in the use of virtual screening methods to enhance the cost-effectiveness of research programmes to discover novel drugs and agrochemicals [1–6]. A range of approaches, both structure-based and ligand-based, have been described in the literature and both play a key role in the lead-discovery stage of research programmes in the pharmaceutical and agrochemical industries. In this paper, we focus on ligand-based virtual screening, specifically on similarity searching using 2D fingerprint representations of molecular structure.

A 2D fingerprint is a vector that encodes the presence or absence of topological substructures (typically atom-, bond- or ring-centred fragments) in a molecule [7, 8], and many different types of fingerprint have been described in the literature [9, 10]. A fingerprint is clearly an extremely simple type of structural representation, but still contains sufficient information to enable effective similarity-based virtual screening to be carried out (see, e.g., [11–17]). Here, the similarity is computed between a reference structure of known biological activity and each of the structures in a database; the most similar structures—the nearest neighbours—are then prime candidates for biological screening. The similarity is computed using a similarity coefficient, normally the Tanimoto coefficient, which is based on the substructures common to the fingerprints representing the reference structure and the current database structure [18].

Fingerprints have traditionally been binary in character, encoding merely the presence (one) or absence (zero) of a 2D substructural fragment in a molecule, but there is no reason why this should necessarily be the case. Instead, it is possible to assign weights to fragments that describe their relative degree of importance in the molecules in which

they occur. Thus, a fragment with a high weight that occurred in both a reference structure and a database structure would make a greater contribution to the overall degree of similarity between those two molecules than would a fragment that was common to them but that had a lower weight. Note that weighting is very different in nature from standardisation, which is commonly used with real-valued data to ensure that all of the attributes comprising a molecular representation, e.g., different types of computed physicochemical property, are measured on the same scale [19]. For example, the well-known Z standardisation ensures that attributes such as logP, molecular weight, number of rotatable bonds etc. will all have means and standard deviations of zero and unity, respectively, and thus make comparable contributions to a similarity calculation. Weighting, conversely, seeks to increase the differences between the various components of a molecular representation that are all of the same type, e.g., the occurrence of fragments in a molecule.

The weighting of fragments on the basis of activity data has been extensively used in approaches to chemical machine learning [15, 20, 21], but this type of weighting cannot be used in chemical similarity searching, since the requisite data are not available; similar comments apply to approaches based on ligand-protein interactions, e.g., recent work on the weighting of substructural features using the FlexX scoring function [22] or the use of X-ray or NMR data [23]. In the typical similarity-searching context, conversely, very limited information is available for the computation of fragment weights, specifically the identity of one, or a few, active molecules. In one of the very first studies of similarity searching, Willett and Winterman discussed three types of weighting: weighting based on the number of times that a fragment occurred in an individual molecule; weighting based on the number of times that a fragment occurred in an entire database; and weighting based on the total number of fragments within a molecule [24]. In addition, Jorgensen et al. have discussed weighting based on the type of fragment (ring system, linker or side-chain) that is common to a pair of molecules that is being compared [25]. In this paper, we focus on the first of the types discussed by Willett and Winterman: specifically, we compare occurrence-based representations (i.e., weighted ones that encode how often a fragment substructure occurs in a molecule) with incidence-based representations (i.e., binary ones that encode merely the presence or absence of a fragment substructure).

Willett and Winterman reported detailed experiments in which simulated property prediction was carried out on 16 QSAR and QSPR datasets represented by lists of atom-centred or bond-centred fragments [24]. They concluded that occurrence-based representations were slightly, but significantly, superior to incidence-based representations; however, the experiments were on a very small scale with the datasets only containing 20–129 structures. Property prediction experiments using small QSAR and QSPR datasets were also reported by Olah et al. [26] and by Azencott et al. [27], both of whom again found that occurrence-based representations performed better than the corresponding incidence-based representations. A preference for occurrence-based representations was observed by Chen and Reynolds in simulated virtual screening experiments using the *NCI AIDS* and *MDL Drug Data Report* (MDDR) databases [28], although they noted that the highly specific fragment definitions that were employed (atom-pairs and atom-sequences) meant that there was often little difference between the two types of representation. The autocorrelation descriptors used by Fechner et al. contained normalised counts of two-point pharmacophores, these being defined by generalised atom-types that resulted in multiple occurrences in molecules [29]. Simulated virtual screening experiments on the small COBRA dataset showed that these occurrence-based representations were slightly better than the corresponding incidence-based versions, although the authors concluded that the latter could be used with little loss of performance. In like vein, Stiefl et al. discussed the use of two-point topological pharmacophore points derived from reduced graphs, and found little difference between weighted and binary forms in searches of the MDDR database [30]. Brown and Martin used binary and occurrence-weighted MACCS keys for cluster-based property prediction and found that the latter were slightly superior [31]; Ewing et al. discussed searches of small, in-house GPCR files using MACCS keys and three-point topological pharmacophores in both weighted and binary forms, and found that the occurrence-based representations were consistently superior [32]; and Good et al. found that inclusion of frequency information enhanced the performance of four-point pharmacophores in 3D searches of a Factor Xa dataset [33].

The work to date hence suggests that fingerprints encoding the occurrences of substructural fragments may be able to give better screening performance than conventional, binary fingerprints. However: the results have been far from consistent and the performance differences often quite small; many of the previous studies were limited, either in terms of the numbers of molecules involved or in the extent to which the weighted and binary fingerprints differed; and there has been no attempt to explain the observed levels of performance. It hence seems appropriate to revisit the use of occurrence-based weighting schemes, especially given that non-binary fingerprints are becoming more common in modern chemoinformatics systems, and we here report a detailed study of the use of such schemes in similarity-based virtual screening.

Methods

Datasets

Our experiments have involved carrying out simulated virtual screening experiments, in which the similarity is computed between a reference structure of known biological activity and each of the molecules in a database. The molecules are ranked in decreasing similarity, a threshold is applied to retrieve some fixed number of the top-ranked molecules, and the activity (or otherwise) of these nearest neighbours noted as a measure of the effectiveness of the search. This is possible here since the databases used—the *MDL Drug Data Report* database (MDDR, from Symyx Technologies at http://www.mdli.com/products/knowledge/drug_data_report/index.jsp) and the *World Of Molecular Bioactivity* database (WOMBAT, from Sunset Molecular Discovery LLC at http://sunsetmolecular.com/products/?id=4)—both contain information about the activities of their constituent molecules. In the case of MDDR, the bioactivity data is qualitative: a molecule is noted as exhibiting a specific activity, and it is assumed to be inactive if that is not the case. In the case of WOMBAT, the original bioactivity data is quantitative: we have converted this to qualitative for our experiments by marking a molecule as inactive if the measured activity value is less than a threshold value, as described below.

Several activity classes were chosen for each of the two databases, as listed in Table 1. Table 1a lists the MDDR classes, which were selected in collaboration with the Novartis Institutes for BioMedical Research and which have been used in several previous studies of ligand-based virtual screening by both ourselves and others (e.g., [34–38]). Each row of the table contains an activity class, a short abbreviation of the name, the number of molecules belonging to the class, the number of distinct ring systems occurring in the set of active molecules for the class, and an indication of the class's diversity. The ring systems considered here (referred to as "active ring systems" in the following) are based on the atomic frameworks of Bemis and Murcko [39], as implemented in the Murko scaffolds routine in the Pipeline Pilot software. The diversity figures were obtained by matching each molecule with every other in its activity class, calculating similarities using the standard Tripos Unity 2D fingerprint and the Tanimoto coefficient, and then computing the mean intra-set similarity. The WOMBAT activity classes in Table 1b mirror closely the MDDR classes in Table 1a. We have chosen for each activity that species for which there is the largest number of molecules with a measured pIC50 $\geq$ 5.0; these molecules are marked as active for that class; molecules with pIC50 $<$ 5.0 for that species are removed from the dataset, as are all molecules with the chosen activity but tested in species other than the chosen one. In all there were 102,535 molecules in the MDDR dataset and 138,127 molecules in the WOMBAT dataset.

In each case, ten representative reference structures from an activity class were chosen for searching: the choices were made using a MaxMin diversity selection procedure, to ensure that the reference structures covered the full range of structural types within each activity class [40]. The numbers of actives retrieved in these similarity searches were then averaged over the ten reference structures, using cut-offs of the top-1% and the top-5% of the similarity rankings. We also noted the numbers of distinct Murcko scaffolds in the active molecules that were retrieved, rather than just the number of active molecules. As recommended by Good et al. [33], this was done to assess the effectiveness of the various weighting schemes for scaffold-hopping [41–43].

Structural representations

The MDDR and WOMBAT molecules were represented by fingerprints encoding three types of topological descriptors: Tripos molecular holograms (available from Tripos Inc. at http://www.tripos.com); Pipeline Pilot ECFC_4 circular substructures (available from Accelrys Software Inc. at http://www.accelrys.com); and Sunset Molecular Discovery LLC keys (available from http://www.sunsetmolecular.com). These have been chosen as exemplifying three very different approaches to the representation of molecular topologies that are all available in commercially available chemoinformatics software systems.

The Tripos holograms were originally developed for 2D QSAR applications and are vectors in which each element contains the number of times that a specific bit has been set by a superimposed-coding procedure [44, 45]. The fragments here are chains of atoms containing 4–7 atoms and ignoring connected hydrogens and stereochemistry, with each such chain hashed into a fixed-length vector containing 997 elements using three different hashing procedures. The Pipeline Pilot ECFC_4 fingerprints encode circular substructures describing a central atom and all atoms within a two-bond radius of it. These substructures are processed using a hashing scheme based on the Morgan algorithm for graph canonicalisation [46]; in our experiments, the resulting integer codes were again hashed to give a fixed-length fingerprint containing 1,024 elements, a procedure that we have found to be highly effective in similarity-based virtual screening experiments [10, 47]. The two types of descriptor hence differ in both the types of topological substructure that are encoded (linear or circular, for holograms and ECFC_4, respectively) and in the number of fingerprint-elements associated with each fragment (three or one, for holograms and ECFC_4, respectively); however, both types of substructure are highly specific in nature and

Table 1 Activity classes used in the virtual screening experiments, chosen from the (a) MDDR and (b) WOMBAT databases

	Active molecules	Active ring systems	Mean pairwise similarity
(a) Activity class (abbreviation)			
5HT3 antagonists (5HT3)	752	417	0.35
5HT1A agonists (5HT1)	827	450	0.34
5HT reuptake inhibitors (5HT)	359	181	0.35
D2 antagonists (D2)	395	258	0.35
Renin inhibitors (REN)	1,125	554	0.57
Angiotensin II AT1 antagonists (ANG)	943	464	0.40
Thrombin inhibitors (THR)	803	425	0.42
Substance P antagonists (SUBP)	1,246	586	0.40
HIV protease inhibitors (HIV)	750	461	0.45
Cyclooxygenase inhibitors (COX)	636	282	0.27
Protein kinase C inhibitors (PKC)	453	171	0.32
(b) Activity class (species)			
5HT3 antagonists (rat)	220	117	0.38
5HT1A antagonists (rat)	592	224	0.40
D2 antagonists (rat)	910	324	0.37
Renin inhibitors (human)	474	253	0.59
Angiotensin II AT1 antagonists (rat)	724	253	0.44
Thrombin inhibitors (human)	421	196	0.42
Substance P antagonists (human)	558	186	0.43
HIV protease inhibitors (human)	1,128	473	0.44
Cyclooxygenase inhibitors (human)	965	220	0.32
Protein kinase C inhibitors (rat)	142	31	0.57
Acetylcholine esterase inhibitors (human)	503	220	0.37
Factor Xa inhibitors (human)	842	328	0.39
Matrix metalloprotease inhibitors (human)	694	280	0.44
Phosphodiesterase inhibitors (human)	596	270	0.36

the resulting fingerprints hence provide a very precise description of molecular topology.

The Sunset keys are rather different in that they derive from two, more generic, approaches to the description of molecular topology. Specifically, they have been inspired by the MDL 320 keys [48] and the CATS (chemically advanced template search) concept [49]: they hence combine chemical substructure recognition (MDL-style) with topologically-relevant pharmacophore patterns based on atom-pairs (CATS-style), in an effort to bridge the gap between substructural and pharmacophore descriptors. The fingerprints are thus more general in nature than the two previous ones; they have been studied previously in an extended evaluation of descriptors for mapping chemistry–biology relationships, this validation involving over a thousand QSAR series, each containing 25 or more compounds and spanning 2 log units in activity, using automated multivariate statistics [26, 50]. The Sunset key-set contains 560 keys encoded by SMARTS: our experiments used 559 of these since one SMARTS (although correctly formed) could not be processed by Pipeline Pilot.

Weighting schemes

Each of the molecular representations (which will be described subsequently as hologram, ECFC_4 or Sunset) can be considered as a vector, X, where the i-th element, x_i, denotes the weight that the i-th fragment has in that molecule. Assume that this i-th fragment occurs f_i times in a molecule, where $f_i \geq 0$. Then we consider in this study the following five weighting schemes (W1–W5):

$$\mathrm{W1} : x_i = 1$$

This is the simplest, binary weight, encoding just the incidence (i.e., the presence or absence) of the i-th fragment, and was obtained by setting to unity all elements in X for which the corresponding fragment occurred one or more times. Alternatively, the occurrence of the i-th fragment is encoded by setting

$$\mathrm{W2} : x_i = f_i,$$

i.e., using the raw frequency data in the representation. W1 and W2 are the obvious weighting schemes, and the ones

that are normally meant when binary and weighted fingerprints are referred to in the literature. However, we have also considered three further ways in which the occurrence frequencies can be used. The first two are standard normalisations in data analysis, and involve taking either the natural logarithm,

$$\mathrm{W3}: x_i = \ln(f_i)$$

or the square root,

$$\mathrm{W4}: x_i = \sqrt{f_i}$$

of the frequency of occurrence. Given that the log of unity is zero, the use of W3 yields a fingerprint that focuses on the more-frequently occurring fragments in a molecule. The effect of the W4 weight is to lessen the contribution of the more generic fragments that can occur relatively frequently within molecules and that thus result in high values for W2 (as demonstrated by the values for W2 in the Mean value of non-zero elements" part of Table 2, as discussed further below). The final scheme, W5, is a normalised version of W2 in which the observed occurrence is expressed as a fraction of the largest value for f_i (i.e., the most frequently occurring fragment in that molecule), and the result further normalised to give a value between 0.5 and 1. This procedure hence takes molecular size (as approximated by numbers of fragment substructures) into account and has been found to be very successful in studies of index-term weighting in text retrieval [51, 52].

$$\mathrm{W5}: x_i = 0.5 + 0.5\frac{f_i}{\max\{f_i\}}.$$

The molecular characterisations of the MDDR and WOMBAT datasets resulting from the five weighting schemes are summarised in Table 2. Reading down from the top: the first row contains the total number of non-zero elements in all the fingerprints representing a dataset; the second row contains the mean number of non-zero elements when averaged over all of the fingerprints in the dataset (102535 for MDDR and 138127 for WOMBAT); the third row contains the fingerprint density, i.e., the mean number of non-zero elements divided by the number of elements in the entire fingerprint (997 for holograms, 1024 for ECFC_4 and 559 for Sunset); the next three rows contain the same three sets of data for W3 (where use of the logarithm has converted all of the $f_i = 1$ values to zero); and the final five rows contain the mean value of each non-zero element. Several features of the representations are evident from this table: the ECFC_4 fingerprints are much sparser than the hologram and Sunset fingerprints; the use of W3 results in a marked, and in some cases an extremely marked, reduction in the number of non-zero elements (since the log function converts all the unity-valued elements to zero); the non-binary Sunset elements have noticeably larger values than the corresponding hologram and ECFC_4 elements; and, as would be expected from their definitions, the mean non-zero element values for W2–W4 are greater than unity (the value for W1) whereas the value for W5 is less than unity.

Similarity coefficient

The similarity S_{XY} between two fragment vectors X and Y was computed in all cases using the full form of the Tanimoto coefficient [9].

$$S_{xy} = \frac{\sum x_i y_i}{\sum {x_i}^2 + \sum {y_i}^2 - \sum x_i y_i},$$

where the summations are over all of the elements in each fingerprint (i.e., 997, 1,024 or 559 elements for holograms, ECFC_4 and Sunset, respectively).

Table 2 Statistical data describing the MDDR and WOMBAT fingerprints

	MDDR			WOMBAT		
	Holograms	ECFC_4	Sunset	Holograms	ECFC_4	Sunset
Number of non-zero elements	35,010,818	5,375,756	20,454,197	44,537,885	6,950,009	26,853,131
Mean number of non-zero elements	341.44	52.43	199.48	322.44	50.32	194.41
Fingerprint density	0.34	0.05	0.36	0.32	0.05	0.34
Number of non-zero elements (W3)	19,810,705	1,553,981	14,311,637	24,844,403	2,100,292	18,758,456
Mean number of non-zero elements (W3)	193.21	15.15	139.57	179.87	15.21	135.81
Fingerprint density (W3)	0.19	0.01	0.25	0.18	0.01	0.24
Mean value of the non-zero elements						
W1	1.00	1.00	1.00	1.00	1.00	1.00
W2	2.45	1.70	4.57	2.46	1.76	4.46
W3	1.04	1.07	1.43	1.04	1.08	1.41
W4	1.44	1.22	1.86	1.44	1.24	1.84
W5	0.60	0.61	0.57	0.60	0.61	0.57

Results

Each of the five different weighting schemes can be applied to the reference structure and to each of the database structures, giving a total of 25 possible similarity measures for the searches using a given type of fingerprint. Our principal interest is in W1 and W2 (the incidence and occurrence representations) and we have hence considered all of the measures that involve either or both of these two schemes; we have also considered those measures where both the reference structure and the database structures are weighted using W3, W4 or W5. In what follows, we refer to each similarity measure by M*ab*, where *a* denotes the weight applied to the database structures' fingerprints and *b* the weight applied to the reference structure's fingerprint so that, e.g., M13 refers to the set of searches (ten searches for each of the chosen activity classes) in which the database structures are coded using W1 (conventional binary weighting) and the reference structures are coded using W3 (the natural logarithm of the occurrence frequencies).

Initial results

Our initial results are summarised in Table 3, which lists the averaged results (numbers of retrieved actives in the top-5% of a sorted ranking) for the searches of the eleven MDDR activity classes using the Tripos holograms. Each column lists the mean values for the searches for a particular activity class (as denoted by the abbreviated form of the class name from Table 1a), and the penultimate column on the right-hand side of the table is the mean value for that similarity measure when averaged over the eleven activity classes (the final column is discussed below). The weighting scheme with the best average recall in each column (i.e., the most effective screening when averaged over the ten active reference structures for that activity class) is strongly shaded and the recall value bold-faced; any scheme with an average recall within 5% of the value for the best weighting scheme is shown lightly shaded.

Visual inspection of the results in Table 3 suggests the following. First, symmetric measures where both the reference structure and the database structures are represented in the same way give good results: indeed, for all but W1, the most effective screening for a given weighting of the reference structure is often when that same weighting scheme is also used to weight the database structures. Second, the best searches are obtained with M22 (raw fragment occurrences for both reference structure and database structures), followed by M33 and then M44 (the logarithm and the square root, respectively, of these fragment occurrences).

The significance, if any, of the differences in performance was tested with Kendall's W test of statistical significance, which is used to evaluate the consistency of k different sets of ranked judgements of the same set of N different objects [53]. Here, we have considered each of the activity classes as a judge ranking the different similarity measures in order of decreasing effectiveness (as measured by the mean number of actives retrieved), i.e., $k = 11$ and $N = 19$. Converting the values in Table 3 to ranks, the

Table 3 Mean numbers of actives retrieved in the top-5% of the ranked database in searches for the eleven MDDR activity classes using Tripos holograms, where the activity classes are described by the abbreviations listed in Table 1a. In each case, the mean is averaged over searches for ten different reference structures. The right-hand columns give the mean numbers of actives averaged over the ten searches for each of the eleven activity classes, and the mean rank when the weights are ranked in decreasing order of numbers of actives retrieved. The weighting scheme with the best average recall in each column is bold-faced and strongly shaded; anything with an average recall within 5% of the value for the best weighting scheme is shown lightly shaded

Similarity measure	Activity class											Mean	
	5HT3	5HT1	5HT	D2	REN	ANG	THR	SUBP	HIV	COX	PKC	Actives	Rank
M11	107.7	83.6	33.8	29.6	421.0	231.2	89.7	119.3	118.5	29.2	64.9	120.8	11.3
M12	83.0	48.4	24.4	19.6	316.3	270.0	73.1	122.4	92.7	21.5	90.8	105.7	13.8
M13	132.8	137.3	**47.1**	**51.9**	518.6	201.6	97.3	194.1	111.0	51.3	55.7	143.3	6.7
M14	102.3	70.3	31.3	25.7	368.4	253.2	83.2	118.1	104.4	24.0	80.1	114.6	12.6
M15	129.9	125.1	42.2	47.9	510.9	209.2	110.9	161.5	120.1	45.8	53.2	141.5	7.4
M21	145.9	95.8	36.9	33.2	111.3	84.0	88.2	19.4	37.7	45.1	20.6	65.3	12.7
M22	150.9	117.4	36.2	46.8	**788.2**	321.5	115.4	**208.6**	**159.8**	54.9	59.9	**187.2**	3.9
M23	133.0	123.6	37.1	45.1	338.7	120.2	89.2	97.3	55.1	**58.2**	37.9	103.2	10.6
M24	**155.4**	127.9	36.0	47.6	448.3	203.6	112.8	133.7	88.3	54.2	46.3	132.2	7.9
M25	133.1	85.4	35.4	31.0	78.7	49.0	65.4	13.2	23.6	43.8	17.4	52.4	14.8
M31	93.0	71.5	27.0	27.4	412.4	246.0	85.9	164.9	124.3	31.3	74.7	123.5	11.4
M32	69.9	53.7	21.0	23.7	244.1	237.6	75.7	198.9	105.2	16.3	86.3	102.9	13.6
M33	134.7	**139.8**	41.6	51.2	726.6	294.2	**118.5**	201.4	141.6	53.7	63.7	**178.8**	3.5
M41	144.4	117.1	38.7	40.9	494.2	254.3	116.5	124.2	118.1	50.4	41.4	140.0	7.4
M42	120.6	84.5	32.8	28.8	459.3	**339.0**	92.6	201.2	139.1	33.2	74.4	146.0	8.4
M44	139.2	107.0	39.6	40.2	659.5	330.0	111.4	204.0	144.9	45.5	56.7	170.7	5.4
M51	87.2	52.8	26.2	18.4	273.4	220.6	74.9	81.3	87.3	21.1	82.5	93.2	15.0
M52	94.6	46.7	20.9	17.0	269.1	251.4	73.9	152.1	84.4	21.0	**102.6**	103.1	14.5
M55	112.9	88.6	33.6	33.3	413.5	274.0	92.8	166.1	129.6	31.6	61.9	130.7	9.1

computed value for W is 0.44. The significance of this value can be tested using the χ^2 distribution since for $N > 7$,

$$\chi^2 = k(N-1)W$$

with N-1 degrees of freedom. This yields a value of 86.3 for χ^2 with 18 degrees of freedom, which is significant at the 0.01 level of statistical significance. Given that a significant level of agreement has been achieved, Siegel and Castellan suggest that the best overall ranking of the N objects can be obtained using their mean ranks averaged over the k judges [53]. These mean ranks are listed in the final column of Table 3, where it will be seen that M22 and M33 are by some way the most effective of the two approaches. There is little to choose between these two measures: M33 is better in terms of the mean rank, but M22 in terms of the mean number of actives retrieved; M22 is better than M33 for six of the eleven activity classes (5HT3, REN, ANG, SUBP, HIV and COX), while M33 is better than M22 for the remainder (5HT1, 5HT, D2, THR and PKC). These two measures are also the most highly ranked, when medians, rather than means, are used to compute the average numbers of actives retrieved in the top-5% of the ranking.

Full results

Table 3 has been discussed in some detail to illustrate the data collected in this study and the analyses carried out. However, we have found [54] that it is always unwise to base conclusions as to the relative merits of different chemoinformatics techniques on a limited set of experiments, and results analogous to those in Table 3 were hence generated under the following conditions: MDDR or WOMBAT databases; hologram, ECFC_4 or Sunset fingerprints; analysis of the top-1% or the top-5% of the search rankings; mean numbers of active molecules or mean numbers of active ring systems. Taking these conditions together, our experiments have been carried out on a very large scale. Thus, just Table 3 on its own represents a total of 2,090 database similarity searches (ten reference structures for each of eleven activity classes searched using 19 weighting schemes), and the complete set of runs (as discussed below) comprised a total of 14,268 database searches (each evaluated in four different ways): we can thus have some confidence in the conclusions that we shall draw from the experiments.

The averaged results are presented in Tables 4 and 5 for the MDDR and WOMBAT databases, respectively: for example, the average results shown in the penultimate column on the right-hand side of Table 3 form the first column of results in Table 4b. We note here that every single column in Tables 4 and 5 gave a statistically significant ranking of the weighting schemes at the 0.01 level of statistical significance, when the data were analysed in the manner described above for Table 3.

Inspection of the results in Tables 4 and 5 shows that the two performance criteria—number of actives and number of active ring systems—tend to give analogous rankings of the various similarity measures; thus, if a measure performs well in terms of number of actives then it will also generally perform well in terms of the number of active scaffolds. In like manner, measures that perform well when the top-1% of the ranking is considered will generally perform well when the top-5% of the ranking is considered. Inspection of the shaded elements shows that the relative performance of the similarity measures involving Tripos holograms is rather different from that observed when the other two representations are used. In this respect, we note that the hologram representation involves a superimposed coding procedure that results in each element of the fingerprint encoding information about multiple linear substructures, and each linear substructure contributing to the occurrences in multiple fingerprint elements. This many-to-many mapping is very different from the one-to-one mapping represented by the Sunset fingerprints and the near one-to-one mapping represented by the ECFC_4 fingerprints (where the hashing to 1,024 elements introduces a very limited degree of fragment overlap [10]).

Two other observations can be made on the results in Tables 4 and 5. First, while ECFC_4 and Sunset are both based on one-to-one mappings, there are often noticeable differences in performance, with some of the Sunset measures (notably M21, M23, M24 and M25) resulting in very poor retrieval indeed. Second, as noted previously when discussing Table 3, there is a marked tendency for the five symmetric measures (i.e., those measures Mab for which $a = b$) to perform better than the 14 asymmetric measures (i.e., those where $a \neq b$). These two observations are discussed further in the next section.

We have carried out a series of analyses using the Kendall-W approach described previously for the data in Table 3. Thus, using the top-1% of the actives as the performance criterion, we can compute the degree of agreement between the six rankings (i.e., three types of fingerprint in each of the two databases) of the 19 similarity measures. The computed value for W is 0.65: this yields a value of 70.51 for χ^2 with 18 degrees of freedom, which is significant at the 0.01 level of statistical significance. The resulting ranking of the 19 measures is:

$$\begin{aligned} &\text{M44} > \text{M14} > \text{M55} > \text{M51} > \text{M11} > \text{M33} > \text{M22} \\ &= \text{M12} > \text{M41} > \text{M15} > \text{M31} = \text{M42} \\ &= \text{M52} > \text{M13} > \text{M24} > \text{M32} > \text{M21} > \text{M23} > \text{M25} \end{aligned}$$

Table 4 Average numbers of active molecules or numbers of active ring systems retrieved in the top-1% (a) or the top-5% (b) of searches of the MDDR database using holograms and ECFC_4 fingerprints. The weighting scheme with the best average recall in each column is bold-faced and strongly shaded; anything with an average recall within 5% of the value for the best weighting scheme is shown lightly shaded

Similarity measure	Holograms		ECFC_4		Sunset	
	Actives	Rings	Actives	Rings	Actives	Rings
M11	53.6	28.3	109.7	60.0	68.2	41.6
M12	42.5	22.9	118.7	**65.5**	61.5	37.9
M13	55.0	29.3	29.0	16.7	52.4	33.3
M14	52.4	28.0	114.9	62.9	70.7	42.7
M15	54.5	29.1	88.1	47.3	50.7	31.6
M21	23.5	12.2	50.7	25.2	2.7	1.7
M22	**89.1**	**45.4**	86.2	47.5	52.1	28.1
M23	28.3	14.6	13.6	7.7	4.2	2.5
M24	47.8	24.7	62.7	32.0	6.6	4.0
M25	15.8	8.6	25.0	13.6	2.5	1.5
M31	50.8	27.8	88.4	50.8	63.6	35.9
M32	33.9	19.6	55.0	35.1	12.2	10.0
M33	80.8	42.3	69.1	39.2	62.0	35.1
M41	56.7	29.4	109.3	58.5	26.2	14.7
M42	57.2	32.2	99.4	57.3	16.6	11.9
M44	74.8	40.7	114.6	62.1	65.0	37.6
M51	40.8	21.2	**119.9**	65.1	**72.3**	**43.5**
M52	36.6	20.1	115.6	62.2	45.7	29.8
M55	59.9	32.2	113.0	61.1	68.0	41.3

(a)

Similarity measure	Holograms		ECFC_4		Sunset	
	Actives	Rings	Actives	Rings	Actives	Rings
M11	120.8	63.6	211.9	114.4	162.0	95.2
M12	105.7	57.3	**227.2**	**124.5**	152.8	90.2
M13	145.3	77.7	95.2	54.4	143.6	84.5
M14	114.6	61.5	219.4	119.0	164.7	**98.0**
M15	141.5	75.8	183.3	99.2	135.0	78.5
M21	65.3	34.3	126.4	63.9	16.5	8.3
M22	**187.2**	**97.7**	185.8	98.9	127.0	67.9
M23	103.2	51.5	59.1	30.8	24.1	12.9
M24	132.2	67.6	142.8	73.0	32.2	17.4
M25	52.4	27.5	76.2	38.2	16.6	8.5
M31	123.5	68.3	197.6	109.1	**165.3**	91.9
M32	103.0	58.5	171.0	98.1	87.4	53.2
M33	178.8	93.2	166.7	91.8	151.8	83.8
M41	140.0	74.8	215.0	114.6	92.5	49.2
M42	146.0	79.4	213.7	118.1	95.6	56.9
M44	170.7	90.9	223.5	120.2	159.1	88.9
M51	93.3	50.6	226.8	121.9	157.8	93.1
M52	103.1	56.1	222.5	120.7	130.2	76.4
M55	130.7	71.0	208.3	112.4	161.8	94.6

(b)

Similar rankings are obtained using the other three performance criteria. Using the top-1% of the active scaffolds, the value for W is 0.65 (again significant) with the ranking:

$$\begin{aligned}&M44 > M14 > M55 > M11 > M51 > M12 > M33 > \\ &M22 > M41 > M15 > M42 > M52 > M13 = M31 > \\ &M24 > M32 > M21 > M23 > M25.\end{aligned}$$

Using the top-5% of the actives, the value for W is 0.57 (this yields a value for χ^2 of 61.11 which is again significant) with the ranking:

$$\begin{aligned}&M44 > M14 > M33 = M55 > M11 = M12 \\ &= M51 > M22 > M31 > M42 > M41 > M15 > M52 > \\ &M13 > M24 > M32 > M23 > M21 > M25.\end{aligned}$$

Using the top-5% of the active scaffolds, the value for W is 0.58 (this yields a value for χ^2 of 62.22 which is again significant) with the ranking:

$$\begin{aligned}&M44 > M14 > M12 > M11 = M55 > M51 > M33 > \\ &M42 > M31 > M22 > M15 > M52 > M41 > M13 > \\ &M24 = M32 > M21 = M23 > M25.\end{aligned}$$

Table 5 Average numbers of active molecules or numbers of active ring systems retrieved in the top-1% (a) or the top-5% (b) of searches of the WOMBAT database using holograms and ECFC_4 fingerprints. The weighting scheme with the best average recall in each column is boldfaced and strongly shaded; anything with an average recall within 5% of the value for the best weighting scheme is shown lightly shaded

Similarity measure	Holograms		ECFC_4		Sunset	
	Actives	Rings	Actives	Rings	Actives	Rings
M11	65.0	27.7	103.6	44.9	73.0	31.2
M12	50.8	21.6	**108.2**	**47.0**	66.9	28.8
M13	71.3	29.0	26.2	11.7	57.6	24.9
M14	61.5	25.9	105.8	45.8	**74.7**	**31.8**
M15	71.9	29.8	89.7	38.3	60.9	26.3
M21	28.3	11.3	50.0	20.2	1.5	0.5
M22	**82.8**	**34.8**	86.0	35.9	60.2	26.3
M23	31.6	12.4	9.0	3.8	2.0	0.6
M24	60.8	24.40	62.1	25.6	5.2	1.8
M25	17.1	7.1	26.4	9.4	1.2	0.3
M31	51.9	22.3	55.4	27.0	65.2	29.0
M32	37.9	14.7	25.0	14.0	7.5	4.3
M33	79.4	33.0	71.3	32.0	71.8	31.0
M41	73.3	31.1	100.8	42.7	31.1	11.9
M42	65.4	27.5	82.0	37.1	10.7	5.2
M44	81.7	34.6	103.0	43.9	73.7	31.4
M51	49.5	20.9	107.2	46.2	73.3	31.3
M52	45.4	19.2	104.7	44.8	51.7	23.2
M55	68.4	29.0	103.0	44.4	71.9	30.8

(a)

Similarity measure	Holograms		ECFC_4		Sunset	
	Actives	Rings	Actives	Rings	Actives	Rings
M11	118.9	52.9	188.2	81.8	157.2	69.5
M12	105.6	47.3	193.4	**84.7**	153.3	69.7
M13	143.6	59.6	85.1	37.8	137.1	60.0
M14	114.7	50.7	191.1	83.2	**165.2**	**74.6**
M15	140.3	59.0	163.7	70.6	131.7	57.3
M21	65.0	26.7	116.0	46.8	10.5	4.1
M22	152.5	63.3	165.8	68.4	139.3	58.8
M23	91.7	37.5	40.7	16.9	15.5	5.6
M24	120.0	49.2	133.7	54.3	24.5	9.1
M25	47.3	19.0	66.8	26.4	9.6	3.0
M31	115.5	51.1	154.3	69.7	154.9	66.8
M32	100.4	43.7	122.8	58.8	74.1	30.9
M33	**156.1**	64.6	158.9	67.6	159.7	67.8
M41	134.7	57.5	186.7	79.4	90.4	36.0
M42	137.0	60.4	172.2	75.7	84.0	40.4
M44	153.5	**64.8**	192.6	82.0	162.3	70.0
M51	95.5	43.0	**196.0**	84.6	160.4	72.1
M52	101.9	46.0	193.7	82.4	132.4	60.3
M55	127.1	55.7	188.8	80.7	157.7	69.3

(b)

Given the similarities noted previously, it is hardly surprising that all the criteria give broadly comparable results: M44 and then M14 are at the top of all rankings; M11, M33, M55, M51 and M22 all do well; M32, M21, M23, M24 and M25 perform very poorly.

As an alternative way of comparing the measures we have counted the number of times that an element is shaded in Tables 4 and 5. The ranking here is fairly similar, with the following measures being shaded at least once (just once in the cases of M31 and M41):

$$\begin{aligned} \text{M51} > \text{M14} = \text{M44} > \text{M11} > \text{M12} = \text{M22} = \text{M52} \\ = \text{M55} > \text{M33} > \text{M31} = \text{M41}. \end{aligned}$$

The main difference from the Kendall rankings is for M51, which is very frequently shaded for ECFC_4 and Sunset but never shaded for holograms; conversely, M22 is always shaded for holograms but never for ECFC_4 and Sunset. Thus, while we have focused here on the overall performance, there are marked differences when account is taken of the representation that is being used, i.e., the best

type of weighting scheme for one type of fingerprint may not be the best for another type of fingerprint.

Discussion

As noted above, the five symmetric measures generally give better results than the 14 asymmetric measures. If, for example, we consider the numbers of actives in the top-1% of the rankings as the performance criterion, then Table 6 lists the mean numbers of actives when averaged over the symmetric and over the asymmetric measures: it will be seen that the former markedly out-perform the latter for all combinations of dataset and representation. That said, there are many individual cases where an asymmetric measure gives excellent results, as exemplified by the multiple shadings in Tables 4 and 5 for M51.

An explanation for this behaviour can be obtained from a consideration of the interactions that occur when two weighting schemes a and b are combined to form a measure Mab and when the resulting combination is used in the computation of the Tanimoto coefficient. To simplify the following description, we shall initially consider the use of two weighting schemes when a molecule is compared with itself.

The basic form of the Tanimoto similarity coefficient between molecules X and Y is

$$S_{XY} = \frac{\sum x_i y_i}{\sum {x_i}^2 + \sum {y_i}^2 - \sum x_i y_i}.$$

When a molecule is matched with itself, i.e., when $X = Y$, all of the fragment substructures are identical. If a symmetric measure is used then $x_i = y_i$ for all i and the Tanimoto coefficient has the value

$$S_{XX} = \frac{\sum x_i x_i}{\sum {x_i}^2 + \sum {x_i}^2 - \sum x_i x_i} = 1.$$

The value of unity is the upper-bound value for this coefficient: the lower-bound for the coefficient is zero when, as here, only non-negative elements are involved (if this is not the case then the lower-bound is $-1/3$).

The upper-bound value will not, however, in general be unity if an asymmetric measure is used. To demonstrate this we make the (grossly) simplifying assumption that all of the fragments that are present in the molecule occur the same number of times, and are thus assigned the same weight; let this weight be W_{NZ}, the mean value of the non-zero elements in a fingerprint (as listed in the bottom part of Table 2). Then for a measure Mab with mean values $W_{NZ}(a)$ and $W_{NZ}(b)$, the similarity of a molecule with itself will be

$$S_{XX} = \frac{\sum W_{NZ}(a) W_{NZ}(b)}{\sum W_{NZ}(a)^2 + \sum W_{NZ}(b)^2 - \sum W_{NZ}(a) W_{NZ}(b)}$$

where the summations are over the non-zero elements in each fingerprint. The value of the resulting similarity can be calculated from the data in Table 2. Thus, if using the MDDR holograms and the W1 and W2 weights, then the values of W_{NZ} from Table 2 are 1.00 and 2.45, respectively; substituting these into the expression above, the upper-bound value of the Tanimoto coefficient for matching a molecule in the W1 representation with itself in the W2 representation is 0.54. The upper-bound for this combination of weights (M12) is still lower for MDDR Sunset (0.26) but rather higher for MDDR ECFC_4 (0.78): the range of computable similarity values hence varies significantly across these three types of fingerprint for this particular combination of weighting schemes. The very low value for Sunset arises from the large difference in the W_{NZ} values for W1 and W2: this difference is 3.57 (i.e., 4.57–1.00) for Sunset, as against 1.45 for holograms and just 0.70 for ECFC_4.

The data in Table 2 can be used to compute analogous upper-bound values for all the possible combinations of weighting schemes, and some of the resulting values are listed in Table 7 (all rounded to two decimal places; the three M31 values of unity are all slightly less than unity when three decimal places are used). It will be seen that the MDDR Sunset M21 value is one of the lowest of all the values in this table; indeed, the Sunset M21, M23, M24 and M25 values (both MDDR and WOMBAT) are the only combinations with upper-bounds lower than 0.40. Analogous behaviour is also observed for the holograms and ECFC_4: in both cases the M21, M23 and M25 values provide the six lowest values (and M24 provides the next two lowest values for the holograms). Thus, for all three types of fingerprint, combinations of the form M2b have low upper-bounds.

Table 6 Effect of using symmetric ($a = b$) or asymmetric ($a \neq b$) similarity measures (Mab). Each element of the table contains the mean number of actives in the top-1% of the ranking when averaged over the five symmetric and 14 asymmetric measures

Similarity measure Mab	MDDR			WOMBAT		
	Holograms	ECFC_4	Sunset	Holograms	ECFC_4	Sunset
$a = b$	71.7	98.5	63.1	75.5	93.4	70.1
$a \neq b$	42.6	77.9	34.8	51.2	67.8	36.4

Table 7 Computed upper-bound values (to two decimal places) of the Tanimoto coefficient for combinations of similarity measure, dataset and fingerprint

Similarity measure	MDDR			WOMBAT		
	Holograms	ECFC_4	Sunset	Holograms	ECFC_4	Sunset
M21	0.54	0.78	0.26	0.54	0.75	0.27
M31	1.00	1.00	0.89	1.00	0.99	0.89
M41	0.88	0.96	0.72	0.88	0.96	0.72
M51	0.79	0.80	0.76	0.79	0.80	0.76
M23	0.55	0.79	0.29	0.54	0.77	0.30
M24	0.63	0.84	0.32	0.62	0.82	0.33
M25	0.48	0.69	0.24	0.48	0.67	0.25

Thus far, we have made two simplifying assumptions. First, the matching of a molecule against itself: whereas in virtual screening, a single molecule, the reference structure, is matched against each of the database structures in turn (all of which are different from it). Second, a focus on the upper-bound values: these are the largest values that could possibly be obtained, and almost certainly different from the values that would be obtained in practice. We have hence taken the largest coefficient value for every search that was carried out, i.e., the value associated with the molecule that came at the top of the similarity ranking, and averaged these largest values across all the searches for all of the activity classes. The resulting mean values for each combination of similarity measure, dataset and fingerprint are shown in Table 8, where it will be seen that there is a fair measure of agreement between these observed largest values and the computed upper-bound values in Table 7. For example, considering the MDDR values (the WOMBAT ones are very similar): the lowest hologram values in Table 7 are for M21, M23 and M25 (0.54, 0.55 and 0.48) and this is also the case for Table 8 (0.40, 0.35 and 0.27). M31 is an outlier since the upper-bound values are high but the largest values less so, particularly for ECFC_4 where the observed value is really quite low (at 0.35). However, other factors may be involved here since the fingerprint densities are lowest for W3, especially for ECFC_4 where reference to Table 2 shows that less than 1% of the fingerprint elements contain non-zero values when this combination of parameters is employed.

The change in coefficient values that accompanies a change in the measure would not, in itself, be a problem if all the similarities changed by the same proportion, i.e., if the coefficient values were scaled linearly, since a ranking of the database structures would remain unchanged. This is not, however, the case, as is demonstrated by Fig. 1. Here, we summarise the similarity values obtained in a search of MDDR for one of the renin reference structures using ECFC_4 fingerprints and the M22 and M25 measures. Specifically, the database structures have been ranked in decreasing order of the M22 values, the mean similarity computed for each successive set of 1,000 structures using both the M22 and the M25 measure, and then the two sets of mean values plotted. Figure 1 demonstrates clearly that the changes in similarity are non-linear, with the reduction in the M25 values being proportionally greater at the top of the ranking than at the bottom. This behaviour arises from the constant contribution that is made to the denominator of the Tanimoto expression by the sum of the squared elements of the fingerprint describing the reference structure: this contribution is invariant across all of the structure in the database, whereas the values of the other two components of the denominator vary from one database structure to another.

This non-linear behaviour can bring about substantial reductions in the effectiveness of searching. The Similar Property Principle [11, 18, 55] would lead us to expect that the active molecules in an activity class are likely to be more similar to an active reference structure from that class

Table 8 Observed largest values (to two decimal places) of the Tanimoto coefficient for combinations of similarity measure, dataset and fingerprint. The results in each case are averaged over all of the similarity searches carried out using the specified combination of parameters

Similarity measure	MDDR			WOMBAT		
	Holograms	ECFC_4	Sunset	Holograms	ECFC_4	Sunset
M21	0.40	0.41	0.22	0.39	0.41	0.26
M31	0.50	0.35	0.57	0.51	0.36	0.57
M41	0.78	0.83	0.58	0.78	0.83	0.58
M51	0.78	0.78	0.75	0.77	0.79	0.75
M23	0.35	0.33	0.35	0.36	0.35	0.38
M24	0.66	0.67	0.53	0.65	0.66	0.54
M25	0.27	0.30	0.15	0.26	0.30	0.18

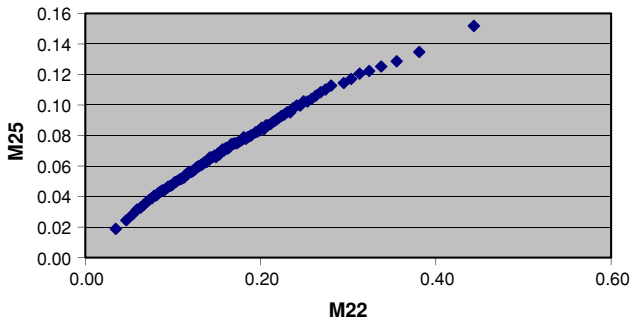

Fig. 1 Similarity between a renin reference structure and MDDR molecules using M22 and M25 similarity measures. The molecules were ranked in decreasing order of the M22 values, the mean similarity computed for each successive set of 1,000 structures using both M22 and M25, and then the two sets of mean values plotted

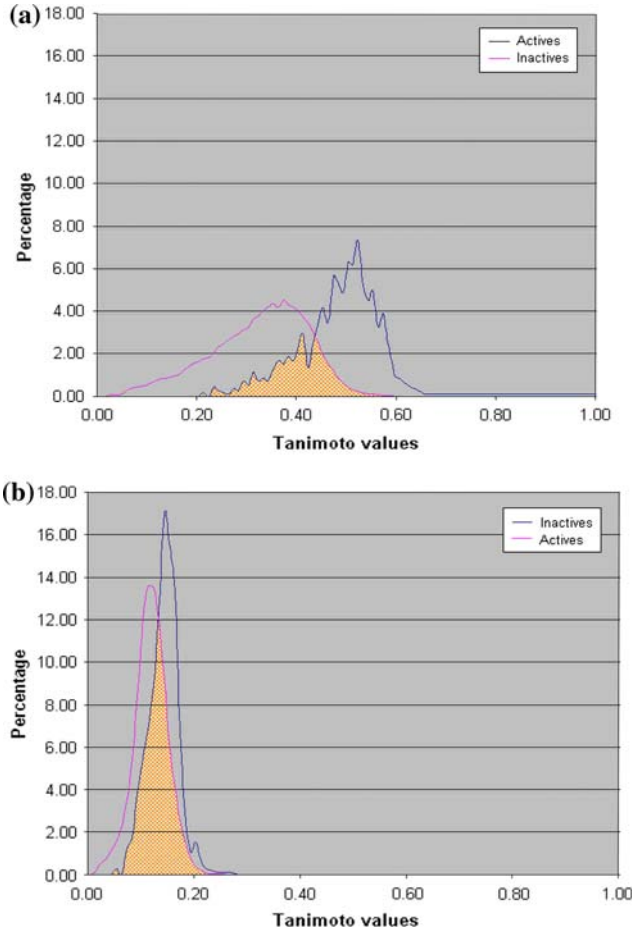

Fig. 2 Percentage frequency distributions for the similarity between a renin reference structure and the sets of active (*blue*) and inactive (*red*) MDDR molecules using **a** M22 and **b** M25 similarity measures

than are the inactive molecules (although there are, of course, many exceptions to this generalisation). Thus, if we plot the frequency distributions for the similarities between the reference structure and the set of active molecules, and the similarities between the reference structure and the set of inactive molecules (which we shall refer to as the Actives distribution and the Inactives distribution, respectively) then we would expect a plot such as that shown in Fig. 2a, which is based on the M22 measure. The figure shows the Actives (in blue) and the Inactives (in red) distributions for the MDDR search considered in Fig. 1. There is a clear separation of the two distributions, with the overlap (shown shaded) comprising 33.6% of each distribution. M22 is a symmetric measure with a consequent upper-bound similarity of unity, and there are large numbers of Actives similarities in the right-hand part of the distribution; indeed, the renin activity class is the most homogeneous of the MDDR classes (see Table 1a) and thus many of the Actives similarities are in excess of 0.80. Consider now Fig. 2b, which shows the Actives and Inactives distributions for the same search but using the asymmetric M25. The Inactives distribution has moved to the left with some "squeezing" of the distribution, but the Actives distribution has moved much sharply to the left, increasing the overlap to 65.6%. The lower-bound similarity of zero remains the same but the change in the upper-bound similarity for this measure disproportionally affects the Actives distribution since this has a much larger proportion of high similarity values: it is hence squeezed much more than is the Inactives distribution when the similarity measure is changed from M22 to M25 and the upper-bound decreased accordingly.

Our analysis would hence suggest that if there is large discrepancy in the weights computed using the two weighting schemes involved then screening effectiveness will be less than if the weights are comparable in magnitude. This is observed in practice, with the M2*b* searches in particular giving a consistently low level of screening, as reflected in the data in Tables 4 and 5, most obviously for some of the Sunset results where the differences in the weights, $W_{NZ}(a)$ and $W_{NZ}(b)$, are greatest and where the results in Tables 4 and 5 are very poor. This arises from the very generic nature of the Sunset fragments, and hence the relatively high frequencies with which individual fragments occur, not some limitation in the fragments themselves; indeed, any generic type of fragment descriptor would be expected to behave similarly in the weighting environment that is being investigated here.

Conclusions

In this paper, we have carried out a detailed study of the use of fragment occurrence data in similarity-based virtual screening. Experiments with fingerprint representations that encode not just the incidence but also the occurrence of topological fragment substructures demonstrate that the

inclusion of information regarding the frequencies with which fragments occur within a molecule will, in most cases, result in an increase in the effectiveness of screening when compared to comparable searches that use just incidence information. The extensive results presented here suggest that the best searches are obtained from fingerprints involving the square root of the raw fragment occurrence data, with an analysis of the observed variations in performance showing that the use of different weighting schemes for the reference structure and for the database structures can result in poor screening performance. Our results and discussion demonstrate, more clearly than in any previous study of which we are aware, the detailed interactions that can take place between representation, weighting scheme and similarity coefficient when a chemical similarity measure is created.

We have chosen to use the Tanimoto coefficient in our experiments given its known effectiveness and extensive use for binary, i.e., unweighted, similarity searching. However, there are many other coefficients that can be used for this purpose and some of these, such as the Forbes or Russell-Rao coefficients, may be superior to the Tanimoto coefficient for binary similarity searching in some circumstances [56, 57]: this may also be the case in a weighted searching environment, and we intend to explore this possibility in future work. We also intend to study the use of data fusion to combine rankings produced using different occurrence-based weighting schemes [58], and to explore the use of weighting schemes that take account of the frequency with which a fragment occurs in an entire database of molecules (rather than its frequency in a single molecule as here) [24].

Acknowledgments We thank Accelrys Software Inc., Sunset Molecular Discovery LLC, Symyx Technologies Inc. and Tripos Inc. for software and data, the Royal Society and the Wolfson Foundation for laboratory support, and the Government of Malaysia for funding.

References

1. Böhm H-J, Schneider G (eds) (2000) Virtual screening for bioactive molecules. Wiley-VCH, Weinheim
2. Klebe G (ed) (2000) Virtual screening: an alternative or complement to high throughput screening. Kluwer, Dordrecht
3. Bajorath J (2002) Nat Rev Drug Discov 1:882. doi:10.1038/nrd941
4. Lengauer T, Lemmen C, Rarey M, Zimmermann M (2004) Drug Discov Today 9:27. doi:10.1016/S1359-6446(04)02939-3
5. Oprea TI, Matter H (2004) Curr Opin Chem Biol 8:349. doi:10.1016/j.cbpa.2004.06.008
6. Alvarez J, Shoichet B (eds) (2005) Virtual screening in drug discovery. CRC Press, Boca Raton
7. Gasteiger J (ed) (2003) Handbook of chemoinformatics. Wiley-VCH, Weinheim
8. Leach AR, Gillet VJ (2007) An introduction to chemoinformatics, 2nd edn. Kluwer, Dordrecht
9. Willett P, Barnard JM, Downs GM (1998) J Chem Inf Comput Sci 38:983. doi:10.1021/ci9800211
10. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) Org Biomol Chem 2:3256. doi:10.1039/b409865j
11. Martin YC, Kofron JL, Traphagen LM (2002) J Med Chem 45:4350. doi:10.1021/jm020155c
12. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK (2004) J Chem Inf Comput Sci 44:1912. doi:10.1021/ci049782w
13. Godden JW, Stahura FL, Bajorath J (2005) J Chem Inf Comput Sci 45:1812. doi:10.1021/ci050276w
14. Willett P (2006) Drug Discov Today 11:1046. doi:10.1016/j.drudis.2006.10.005
15. Eckert H, Bajorath J (2007) Drug Discov Today 12:225. doi:10.1016/j.drudis.2007.01.011
16. Sheridan RP (2007) Expert Opin Drug Discov 2:423. doi:10.1517/17460441.2.4.423
17. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD (2007) J Chem Inf Model 47(4):1504. doi:10.1021/ci700052x
18. Willett P (2009) Ann Rev Inf Sci Technol 43:3
19. Chu C-W, Holliday JD, Willett P (2009) J Chem Inf Model 49:155. doi:10.1021/ci800224h
20. Ormerod A, Willett P, Bawden D (1989) Quant Struct-Activ Relat 8:115. doi:10.1002/qsar.19890080207
21. Goldman BB, Walters WP (2006) Ann Report Comput Chem 2:127
22. Crisman TJ, Sisay MT, Bajorath J (2008) J Chem Inf Model 48:1955. doi:10.1021/ci800229q
23. Stiefl N, Zaliani A (2006) J Chem Inf Model 46:587. doi:10.1021/ci050324c
24. Willett P, Winterman V (1986) Quant Struct-Activ Relat 5:18. doi:10.1002/qsar.19860050105
25. Jorgensen WL, Duffy EM (2002) Adv Drug Deliv Rev 54(3):355. doi:10.1016/S0169-409X(02)00008-X
26. Olah M, Bologa C, Oprea TI (2004) J Comput Aided Mol Des 18:437. doi:10.1007/s10822-004-4060-8
27. Azencott C-A, Ksikes A, Swamidass SJ, Chen JH, Ralaivola L, Baldi P (2007) J Chem Inf Model 47:965. doi:10.1021/ci600397p
28. Chen X, Reynolds CH (2002) J Chem Inf Comput Sci 42:1407. doi:10.1021/ci025531g
29. Fechner U, Paetz J, Schneider G (2005) QSAR Comb Sci 24:961. doi:10.1002/qsar.200530118
30. Stiefl N, Watson IA, Baumann K, Zaliani A (2006) J Chem Inf Model 46:208. doi:10.1021/ci050457y
31. Brown RD, Martin YC (1996) J Chem Inf Comput Sci 36:572. doi:10.1021/ci9501047
32. Ewing TJA, Baber JC, Feher F (2006) J Chem Inf Model 46:2423. doi:10.1021/ci060155b
33. Good AC, Cho SJ, Mason JS (2004) J Comput Aided Mol Des 18:523. doi:10.1007/s10822-004-4065-3
34. Bender A, Mussa HY, Glen RC, Reiling S (2004) J Chem Inf Comput Sci 44:1708. doi:10.1021/ci0498719
35. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2005) J Med Chem 48:7049. doi:10.1021/jm050316n
36. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2006) J Chem Inf Model 46:462. doi:10.1021/ci050348j
37. Grant JA, Haigh JA, Pickup BT, Nicholls A, Sayle RA (2006) J Chem Inf Model 46:1912. doi:10.1021/ci6002152
38. Fischer JR, Rarey M (2007) J Chem Inf Model 47:1341. doi:10.1021/ci700007b
39. Bemis GW, Murcko MA (1996) J Med Chem 39:2887. doi:10.1021/jm9602928

40. Snarey M, Terrett NK, Willett P, Wilton DJ (1997) J Mol Graph Model 15:372. doi:10.1016/S1093-3263(98)00008-4
41. Böhm H-J, Flohr A, Stahl M (2004) Drug Discov Today. Technology 1(3):217
42. Brown N, Jacoby E (2006) Mini Rev Med Chem 6:1217. doi:10.2174/138955706778742768
43. Schneider G, Schneider P, Renner S (2006) QSAR Comb Sci 25:1162. doi:10.1002/qsar.200610091
44. Tong W, Lowis DR, Perkins R, Chen Y, Welsh WJ, Goddette DW, Heritage TW, Sheehan DM (1998) J Chem Inf Comput Sci 38:669. doi:10.1021/ci980008g
45. Seel M, Turner DB, Willett P (1999) Quant Struct-Activ Relat 18:245. doi:10.1002/(SICI)1521-3838(199907)18:3<245::AID-QSAR245>3.0.CO;2-O
46. Hassan M, Brown RD, Varma-O'Brien S, Rogers D (2006) Mol Divers 10:283. doi:10.1007/s11030-006-9041-5
47. Gardiner EJ, Gillet VJ, Haranczyk M, Hert J, Holliday JD, Malim N, Patel Y, Willett P (2008) Statistical analysis and data mining (in press)
48. Durant JL, Leland BA, Henry DR, Nourse JG (2002) J Chem Inf Model 42:1273. doi:10.1021/ci010132r
49. Schneider G, Neidhart W, Giller T, Schmid G (1999) Angew Chem Int Ed 38:2894. doi:10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F
50. Bologa C, Allu TK, Olah M, Kappler MA, Oprea TI (2005) J Comput Aided Mol Des 19:625. doi:10.1007/s10822-005-9020-4
51. Salton G, Buckley C (1988) Inf Process Manage 24:513. doi:10.1016/0306-4573(88)90021-0
52. Salton G (1989) Automatic text processing. Addison-Wesley, Reading, MA
53. Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioural sciences, 2nd edn. McGraw-Hill, New York
54. Willett P (2004) Methods Mol Biol 275:51
55. Johnson MA, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. John Wiley, New York
56. Holliday JD, Hu C-Y, Willett P (2002) Comb Chem High Throughput Screen 5:155
57. Holliday JD, Salim N, Whittle M, Willett P (2003) J Chem Inf Comput Sci 43:819. doi:10.1021/ci034001x
58. Willett P (2006) QSAR Comb Sci 25:1143. doi:10.1002/qsar.200610084