Enhancing the Effectiveness of Fingerprint-Based Virtual Screening: Use of Turbo Similarity Searching and of Fragment Frequencies of Occurrence

Shereena M. Arif, Jérôme Hert, John D. Holliday, Nurul Malim, and Peter Willett

Krebs Institute for Biomolecular Research and Department of Information Studies,
University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, United Kingdom
{p.willett@sheffield.ac.uk}

Abstract. Binary fingerprints encoding the presence of 2D fragment substructures in molecules are extensively used for similarity-based virtual screening in the agrochemical and pharmaceutical industries. This paper describes two techniques for enhancing the effectiveness of screening: the use of a second-level search based on the nearest neighbours of the initial reference structure; and the use of weighted fingerprints encoding the frequency of occurrence, rather than just the mere presence, of substructures. Experiments using several databases for which both structural and bioactivity data are available demonstrate the effectiveness of these two approaches.

Keywords: Chemoinformatics, Fingerprint, Fragment substructure, Similarity measure, Similarity searching, Turbo similarity searching, Virtual screening, Weighting scheme.

1 Introduction

Virtual screening, the ranking of molecules in order of probability of biological activity, plays an increasingly important role in the discovery of novel bioactive molecules in the agrochemical and pharmaceutical industries [1, 2]. There are many ways in which this can be achieved: here, we discuss the use of *similarity searching* for this purpose [3, 4]. Given a molecule that exhibits some biological activity of interest (the *reference structure*) and a database of molecules that have not previously been tested for that activity, a similarity search computes a measure of structural similarity between the reference structure and each of the database structures in turn. The database is then ranked in decreasing order of the computed similarities, and the top-ranked, nearest-neighbour molecules passed on for further consideration as having the greatest *a priori* probabilities of bioactivity. The most common similarity measure involves the use of 2D fingerprints and the Tanimoto coefficient, where a *fingerprint* is a binary vector encoding the presence or absence in a molecule of small substructural fragments [4]. Fingerprint-based similarity is clearly simple in concept but has proved to be very effective in operation [5-9].

Hert *et al.* have described an extension of similarity searching, *turbo similarity searching* (subsequently referred to here as TSS) [10]. The similar property principle states that molecules that are structurally similar are likely to exhibit similar bioactivities and properties [11, 12]: thus, the nearest neighbours of a bioactive reference structure are also expected to possess that particular bioactivity. Recent studies have demonstrated the increased effectiveness of searching that can be obtained if not one but multiple bioactive reference structures are available, using an approach called *group fusion* [9, 13, 14]. Here, each reference structure in turn is used for a similarity search, and then the resulting rankings combined to give a single consensus ranking [15]. TSS makes the assumption that the nearest neighbours of a reference structure are not just *likely* to be active (as suggested by the similar property principle) but actually *are* active; they can thus be used as the multiple reference structures required for the implementation of group fusion [10]. The user of a TSS system needs to do nothing more than is required for conventional similarity searching, i.e., the input of a bioactive reference structure; however, the final, combined search output is expected to yield a better level of enrichment than a conventional similarity search (hereafter SS) based on just the original reference structure. Hert *et al*. found that TSS yielded favorable results and they hence suggested that the approach provides a simple way of enhancing the effectiveness of current systems for virtual screening [10]. The original TSS experiments used the *MDL Drug Data Report* (MDDR) database with the molecules represented by one particular type of fingerprint (specifically the Pipeline Pilot ECFP_4 fingerprints). In the first part of the present paper, we consider the effectiveness of TSS when used with other databases and other types of fingerprint to determine the generality of TSS for virtual screening.

Fingerprints for similarity searching are normally binary, with each element of the fingerprint denoting the presence (one) or the absence (zero) of a particular substructural fragment in a molecule. Alternatively, it is possible to assign weights to fragments so that a fragment with a high weight that is common to both a reference structure and a database structure makes a greater contribution to the computed similarity than will a common fragment with a lesser weight. In the absence of the extensive training data needed for machine learning approaches to fragment weighting [16], one source of information that can be used for weighting fragments is the number of times that a fragment occurs in an individual molecule [17]. Several previous studies have suggested that occurrence-based fingerprints (i.e., weighted fingerprints that encode how often a substructure occurs in a molecule) can give better screening than incidence-based fingerprints (i.e., conventional binary fingerprints that encode merely the presence or absence of a substructure). However, the results to date have been far from consistent, with the experiments often involving only small datasets and with no attempt to explain the observed levels of performance: the study reported in the second part of this paper was carried out to address these limitations.

2 Experimental Details

2.1 Databases

Several databases have been used in our experiments. The largest number of experiments used the MDDR dataset of 102,514 molecules and eleven bioactivity classes first described by Hert *et al*. [13]. This file of molecules was also screened for ten activity classes (dataset MDDR-HET) that had been chosen to be as structurally diverse as possible, which provides a tougher test of a screening method's scaffold-hopping abilities [18]. Further experiments used: a dataset of 138,127 molecules and 14 activity classes taken from the *World of Molecular Bioactivity* database (WOMBAT, available from Sunset Molecular Discovery LLC); and a dataset of 41,192 molecules with 393 confirmed actives from the NCI database, which contains molecules tested in the US government's anti-AIDS programme.

2.2 Fingerprints

There are two main classes of fingerprint. The *dictionary-based* approach involves a pre-defined list of fragments: a molecule is checked for the presence of each of the fragments in the dictionary, and a bit set (or not set) when a fragment is present (or absent). The *molecule-based* approach involves hashing algorithms that allocate multiple fragments to each bit-position: a note is made of all fragments of a specific type (e.g., a chain of four connected non-hydrogen atoms) occurring in a molecule, and then each fragment is hashed to set one or more bits in the fingerprint.

The TSS experiments used the following fingerprints (which are described in detail by Gardiner *et al.* [19]): Pipeline Pilot ECFP_4 and FCFP_4 fingerprints (1024 bits, available from Accelrys Software Inc.); Tripos Unity fingerprints (988 bits, available from Tripos Inc.); BCI fingerprints (1052 bits, available from Digital Chemistry Ltd.); Daylight fingerprints (2048 bits, available from Daylight Chemical Information Systems Inc.); and MDL keys (166 bits, available from Symyx Technologies Inc.). Of these, the BCI and MDL fingerprints are dictionary-based, the Daylight and Pipeline Pilot fingerprints are molecule-based (using linear chains and circular substructures, respectively), and the Unity fingerprints are based on both approaches, thus encompassing both the main classes of fingerprint that are currently available.

The weighting experiments in the second part of the paper used the following fingerprints: Tripos holograms (which employ hashed fragments analogous to those used in the Tripos Unity fingerprints, with a fingerprint containing 997 elements); Pipeline Pilot ECFC_4 fingerprints (the occurrence version of the ECFP_4 fingerprints, with a fingerprint containing 1024 elements); and Sunset keys (available from Sunset Molecular Discovery LLC), for which the 559-element fingerprints are rather more generic in character than the other two types of descriptor studied here, as they combine chemical substructure recognition with topologically-relevant pharmacophore patterns based on atom-pairs.

2.3 Weighting Schemes

In the weighting experiments, each of the molecular representations (holograms, ECFC_4 or Sunset) was considered as a vector, X, where the i-th element, x_i, denotes the weight that the i-th fragment has in that molecule. If the i-th fragment occurs f_i times in a molecule ($f_i \geq 0$) then five weighting schemes (W1-W5) were considered. W1 and W2 are the raw incidence and occurrence data, i.e.,

$$\text{W1: } x_i = 1 \text{ (for } f_i > 0\text{); W2: } x_i = f_i .$$

W3 and W4 are two common standardizations in multivariate statistics:

$$\text{W3: } x_i = \ln(f_i)\text{; W4: } x_i = \sqrt{f_i} .$$

W5 involves a further standardization that has proved helpful in weighting studies in text retrieval [20]:

$$\text{W5: } x_i = 0.5 + 0.5\frac{f_i}{\max\{f_i\}} ,$$

where $\max\{f_i\}$ is the frequency of occurrence of the most frequently occurring fragment in a molecule.

3 Turbo Similarity Searching

The results of the TSS searches are presented in Table 1. The measure of retrieval effectiveness used here, and also for the weighting experiments in the next section, is the *recall*, i.e., the fraction of the active molecules retrieved at some cut-off point in the ranking. Our experiments involved a cut-off of 5%, so that, e.g., a recall of 20% of the actives would correspond to a four-fold enrichment of the output as compared with random screening of the database. The recall values in Table 1 are the mean percentage of actives retrieved in the top-5%, averaged over all reference structures for each activity class and then over all activity classes for each database.

The searches of the MDDR, MDDR-HET and NCI datasets used each active molecule in turn as the reference structure. The results for these datasets are shown in Tables 1a-1c, where SS denotes a conventional similarity search and where TSS-*x* denotes a turbo similarity search based on the original reference structure combined with its *x* nearest neighbours. Results are presented for *x*= 10, 20, 50 and 100, with the best SS and TSS performance marked as bold-faced and shaded. When the MDDR classes are used (Table 1a) there is often a noticeable increase in the recall of the search as more nearest neighbours are included in a TSS, with the best searches using 50-100 nearest neighbours. However, SS is superior to TSS for the MDL fingerprints, and there is little difference in performance for the Unity fingerprints. The ECFP_4 fingerprint gives the best results, both in the initial SS and in the degree of enhancement when TSS is used: for this fingerprint, the maximum TSS recall corresponds to an increase of ca. 15% of the SS recall, a significant finding since this enhancement is achieved without any additional effort on the part of the user carrying out the similarity search.

Table 1. Similarity (SS) and turbo similarity (TSS) searches of (a) MDDR, (b) MDDR-HET, (c) NCI and (d) WOMBAT datasets

1a

Fingerprint	SS	TSS-10	TSS-20	TSS-50	TSS-100
BCI	32.8	33.8	34.2	34.7	34.9
Daylight	31.5	32.4	32.6	33.1	32.8
ECFP_4	**39.2**	41.9	42.9	44.5	**45.1**
FCFP_4	36.1	37.9	38.9	40.1	40.8
MDL	30.2	27.9	28.0	28.1	28.2
Unity	30.2	30.8	30.9	31.0	31.1

1b

Fingerprint	SS	TSS-10	TSS-20	TSS-50	TSS-100
BCI	20.7	20.9	20.6	20.2	19.6
Daylight	18.3	18.0	17.4	16.7	16.4
ECFP_4	**20.9**	22.3	**22.5**	**22.5**	22.0
FCFP_4	20.2	21.1	21.1	20.7	20.1
MDL	20.0	20.0	19.5	18.9	18.3
Unity	16.6	15.8	15.2	14.1	13.8

1c

Fingerprint	SS	TSS-10	TSS-20	TSS-50	TSS-100
BCI	**12.1**	12.3	12.3	12.5	**12.8**
Daylight	10.4	10.5	10.4	10.2	10.0
ECFP_4	10.5	10.3	10.3	10.4	10.7
FCFP_4	10.8	10.9	11.1	11.1	11.1
MDL	11.9	11.9	12.0	12.1	12.3
Unity	11.5	11.5	11.6	11.8	11.7

1d

Fingerprint	SS	TSS-10	TSS-20	TSS-50	TSS-100
BCI	39.0	39.6	39.8	40.0	40.0
Daylight	35.1	35.9	36.0	35.6	36.2
ECFP_4	**47.2**	48.6	49.5	50.6	**51.9**
FCFP_4	42.2	43.0	43.9	44.7	45.1
MDL	36.6	37.1	37.1	37.2	36.9
Unity	36.8	37.3	37.8	37.5	37.4

A very different pattern of behaviour is observed with the MDDR-HET results presented in Table 1b. The degree of enhancement for this more challenging screening task is much less notable, even for the ECFP_4 fingerprint, and for most of the fingerprints there would appear to be little or no advantage in using TSS. Similar comments apply to the searches of the NCI dataset shown in Table 1c.

In the WOMBAT experiments, ten molecules were chosen at random from each activity class to be the reference structures for searching. The results of these searches are detailed in Table 1d, from which one can draw similar conclusions as from Table 1a: the initial SS recall is high but there is still a substantial increase in the effectiveness of the TSS searches for the ECFP_4 and (to a lesser extent) the FCFP_4 fingerprints; however, TSS provides only limited benefits with the other types of fingerprint.

Table 2. Similarity (SS) and turbo similarity (TSS-SSA) searches of (a) MDDR-HET and (b) NCI datasets

2a

Fingerprint	SS	TSS-10	TSS-20	TSS-50	TSS-100
BCI	20.7	27.1	27.1	26.0	24.8
Daylight	18.3	25.0	23.3	21.7	21.0
ECFP_4	**20.9**	21.5	28.5	**28.8**	27.9
FCFP_4	20.2	18.3	24.0	25.9	25.5
MDL	20.2	26.5	25.5	24.2	23.4
Unity	16.6	25.1	23.4	21.1	19.8

2b

Fingerprint	SS	TSS-10	TSS-20	TSS-50	TSS-100
BCI	**12.1**	12.9	11.9	11.2	11.5
Daylight	10.4	10.7	9.8	9.2	9.5
ECFP_4	10.5	**14.5**	11.8	10.4	10.4
FCFP_4	10.8	13.3	11.9	10.9	11.0
MDL	11.9	10.9	10.7	10.7	11.0
Unity	11.5	10.4	9.9	9.8	10.1

The results in Table 1 hence suggest that TSS can provide a simple way of improving the effectiveness of similarity searching for at least some types of fingerprints if the active molecules are not too structurally diverse. If they are diverse, as is the case with the MDDR-HET or NCI datasets, then Hert *et al*. suggest an alternative form of TSS – referred to here as TSS-SSA – in which the nearest neighbours from the basic SS search are processed using a machine-learning technique, rather than group fusion as discussed thus far [18].

Machine learning involves analysing a training set containing known active and inactive molecules and then developing a decision rule to rank the remaining test-set molecules in order of decreasing probability of activity. Hert *et al*. suggested that the nearest neighbours of the known reference structure could form the training-set's actives with the remainder of the dataset forming the training-set inactives [18]. The decision rule is based on the technique known as substructural analysis (an early form of naïve Bayesian classifier). Substructural analysis (hereafter SSA) computes a weight for each bit in a fingerprint describing the corresponding fragment's propensity to occur in active or in inactive molecules [21]. The weighting scheme used was the *R2* weight, which has the form

$$R2 = \log\left(\frac{A_j/N_A}{I_j/N_I}\right).$$

Here, A_j and I_j are the numbers of active and inactive training-set molecules with bit j set, and N_A and N_I are the numbers of active and inactive training-set molecules [22]. A molecule's overall score is the sum of the *R2* weights for its constituent fragments, and the molecules in a dataset are ranked in decreasing order of the sum of scores. The results of using TSS-SSA are shown in Table 2, where it will be seen that the use of SSA, rather than of group fusion, in the second-stage search has brought about substantial increases in screening performance with all fingerprints for MDDR-HET;

with NCI, substantial performance increases were obtained only with ECFP_4 and FCFP_4 in the TSS-10 searches.

Taken together, the results in Tables 1 and 2 suggest that TSS can bring about substantial enhancements in virtual-screening performance in some cases, especially when the highly effective ECFP_4 fingerprint is used.

4 Use of Fragment Occurrence Data

4.1 Screening Performance

Previous studies of weighted similarity searching have considered just the incidence (W1) and occurrence (W2) weighting schemes: here, we considered all similarity measures involving either of these, and those where both the reference structure and the database structures were weighted using W3, W4 or W5. In the following, a similarity measure M*ab* denotes a measure with weight *a* and *b* applied to the fingerprints of the database structures and of the reference structure, respectively.

Searches were carried out using each of the 19 resulting similarity measures on both the MDDR and WOMBAT datasets, using ten different reference structures for each of the associated activity classes and using holograms, ECFC_4 and Sunset fingerprints. The recall values for these searches are shown in Table 3: the recall here is the mean number of actives retrieved in the top-5%, averaged over the ten reference structures for each class and then over all classes for each database. The scheme with the best mean recall in each column again has the value bold-faced and shaded.

It is possible to assess the consistency of the results using Kendall's *W* test of statistical significance, which is used to evaluate the level of agreement between *k* different sets of ranked judgments of the same set of *N* different objects [23]. Here, we have considered each of the fingerprint/dataset combinations as a judge ranking the different similarity measures in order of decreasing effectiveness (as measured by the recall values), i.e., k=6 and N=19. Converting the values in Table 3 to ranks, we obtain a value for *W* of 0.57, which is significant at the 0.01 level of statistical significance using a modified χ^2 test with N-1 degrees of freedom. Since a significant level of agreement has been achieved, the best overall ranking of the *N* objects is the objects' mean ranks when averaged over the *k* judges [23]. This gives the following ranking of the similarity measures:

M44 > M14 > M33=M55 > M11=M12=M51 > M22 > M31 > M42 > M41 > M15
> M52 > M13 > M24 > M32 > M23 > M21 > M25.

Thus M44 and M14 (both involving W4, the square root of the raw frequencies of occurrence) are at the top of the rankings; M11, M33, M55, M51 and M22 all do well; and M32, M21, M23, M24 and M25 perform very poorly.

The work hence suggests that the inclusion of occurrence information can increase the effectiveness of current similarity searching systems, which predominantly use binary fingerprints. Of the various weighting schemes we have chosen, our results indicate the general effectiveness of the W4 scheme, which seeks to lessen the contribution made by the most frequently occurring fragments within a molecule.

Table 3. Similarity searches of the MDDR and WOMBAT databases using different weighting schemes.

Weight	MDDR			WOMBAT		
	Holograms	ECFC_4	Sunset	Holograms	ECFC_4	Sunset
M11	120.8	211.9	162.0	118.9	188.2	157.2
M12	105.7	**227.2**	152.8	105.6	193.4	153.3
M13	145.3	95.2	143.6	143.6	85.1	137.1
M14	114.6	219.4	164.7	114.7	191.1	**165.2**
M15	141.5	183.3	135.0	140.3	163.7	131.7
M21	65.3	126.4	16.5	65.0	116.0	10.5
M22	**187.2**	185.8	127.0	152.5	165.8	139.3
M23	103.2	59.1	24.1	91.7	40.7	15.5
M24	132.2	142.8	32.2	120.0	133.7	24.5
M25	52.4	76.2	16.6	47.3	66.8	9.6
M31	123.5	197.6	**165.3**	115.5	154.3	154.9
M32	103.0	171.0	87.4	100.4	122.8	74.1
M33	178.8	166.7	151.8	**156.1**	158.9	159.7
M41	140.0	215.0	92.5	134.7	186.7	90.4
M42	146.0	213.7	95.6	137.0	172.2	84.0
M44	170.7	223.5	159.1	153.5	192.6	162.3
M51	93.3	226.8	157.8	95.5	**196.0**	160.4
M52	103.1	222.5	130.2	101.9	193.7	132.4
M55	130.7	208.3	161.8	127.1	188.8	157.7

Table 4. Mean values of the non-zero elements of each type of weighted fingerprint for the MDDR and WOMBAT fingerprints

Mean value	MDDR			WOMBAT		
	Holograms	ECFC_4	Sunset	Holograms	ECFC_4	Sunset
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

4.2 Analysis of Similarity Measures

We can draw two further conclusions from our results: that symmetric similarity measures (i.e., measures M*ab* where *a*=*b*) tend to do better than asymmetric measures (i.e., where *a*≠*b*); and that many of the measures involving W2 perform very badly. These conclusions may be rationalized by considering the interactions that occur when two weighting schemes *a* and *b* are combined to form a measure M*ab* and when the resulting measure is used to compute the Tanimoto similarity coefficient.

The basic form of the Tanimoto coefficient for molecules *X* and *Y* is

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

where the summations are over the non-zero elements in each fingerprint. If a molecule is matched with itself and if a symmetric measure is used, then $x_i=y_i$ for all i and the Tanimoto coefficient has the value of unity, which is the upper-bound value for this coefficient. However, the upper-bound may be less than unity if an asymmetric measure is used, as we now demonstrate. Assume that all fragments in a molecule occur equifrequently, and are thus assigned the same weight, W_{NZ}, which is the mean value of the non-zero elements in a molecule's fingerprint when that molecule is weighted using some particular weighting scheme. Then the self-similarity for a molecule X using the measure M*ab*, with weights $W_{NZ}(a)$ and $W_{NZ}(b)$, is

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

Values for W_{NZ} using each of the schemes W1-W5 for the two datasets are shown in Table 4, and these can be used to compute the similarities S_{XX}. For example, if using the MDDR holograms and the W1 and W2 weights, then the values of W_{NZ} from the table are 1.00 and 2.45, respectively: this gives an upper-bound of 0.54 to the self-similarity of a molecule in the W1 representation with itself in the W2 representation (i.e., M12). This value can be compared with the corresponding M12 upper-bounds for MDDR Sunset (0.26) and MDDR ECFCP_4 (0.78), demonstrating the wide range of upper-bound values for the same similarity measure that is obtained using the different fingerprints. Analogous upper-bounds can be computed using the data in Table 4 for all of the other measures M*ab*: these computations show that combinations of the form M2*b* have low upper-bounds for all three types of fingerprint. Thus, if there is large discrepancy in the weights computed using the two weighting schemes involved in the chosen similarity measure then there will be a much smaller range of possible similarity values than if the weights are of comparable magnitude. If only a limited range of values is available to the coefficient, then the ranking will be less discriminating resulting in the poor (and in some cases very poor) screening performance that is demonstrated in Table 3 for some combinations of similarity measure and representation, e.g., WOMBAT Sunset M21 and M25.

The similarity analysis above is grossly simplified in that it considers self-similarities (rather than the similarities between a reference structure and a database structure) and it considers only upper-bound values (which are likely to differ from the largest similarities that are actually obtained during a similarity search). Even so, more detailed examination demonstrates the general correctness of the analysis above, with the similarity behavior observed here mirroring that obtained in searches of entire databases (rather than in self-similarity calculations) using actual (rather than upper-bound) similarities: this more detailed work will be reported shortly. We hence conclude that the upper-bound value for the Tanimoto coefficient depends on the natures of the weighting schemes *a* and *b*: if *a*=*b* then the upper-bound will be unity; however, if this is not the case and the corresponding weights differ substantially, then the upper-bound can be markedly less than unity. This implies a reduction (and in some cases, a severe reduction) in the discriminatory power of the resulting similarity measure when it is used for virtual screening.

5 Conclusions

Similarity-based approaches are widely used for virtual screening. Conventional similarity searching involves using a binary fingerprint describing a bioactive reference structure to rank a chemical database in order of decreasing probability of activity. In this paper, we have described two ways in which the conventional approach can be enhanced: turbo similarity searching based on identifying and then exploiting the reference structure's nearest neighbours; and taking account of fragments' frequencies of occurrence in molecules.

The search results in Tables 1 and 2 show that turbo similarity searching based on a consensus approach called group fusion can provide substantial enhancements in screening performance if the normal similarity search provides a good starting point, i.e., if the similar property principle holds and if the actives are well clustered using the chosen structure representation and similarity measure. This was particularly the case in the searches based on the ECFP_4 fingerprint; indeed, this would appear to be the representation of choice for similarity-based virtual screening using binary fingerprints.

The search results in Table 3 show that fingerprint representations encoding the occurrence-frequencies of fragment substructures can perform much better than conventional binary fingerprints in similarity-based screening, especially using symmetric similarity measures that include the W4 square-root weight; that said, some other combinations of weights can perform very badly. An upper-bound analysis provides a rationalization of the observed variations in performance, this demonstrating the subtle interactions that may occur between the representation and the weighting scheme when a chemical similarity measure is created.

Current work on similarity-based virtual screening includes considering alternative consensus rules for the implementation of the group fusion stage of TSS, and the use of different similarity coefficients for weighted fingerprint searching.

Acknowledgements. We thank the following: Kristian Birchall for assistance with the WOMBAT data; the Government of Malaysia, and the Novartis Institutes for Biomedical Research for funding; and Accelrys Software Inc., Daylight Chemical Information Systems Inc., Digital Chemistry Limited, the Royal Society, SciTegic Inc., Sunset Molecular Discovery LLC, Symyx Technologies Inc., Tripos Inc. and the Wolfson Foundation for data, software and laboratory support.

References

1. Stahura, F.L., Bajorath, J.: Virtual Screening Methods That Complement High-Throughput Screening. Combin. Chem. High-Through. Screening 7, 259-269 (2004)

2. Alvarez, J., Shoichet, B. (eds.): Virtual Screening in Drug Discovery. CRC Press, Boca Raton (2005)
3. Eckert, H., Bajorath, J.: Molecular Similarity Analysis in Virtual Screening: Foundations, Limitation and Novel Approaches. Drug Discov. Today 12, 225-233 (2007)
4. Willett, P.: Similarity Methods in Chemoinformatics. Ann. Rev. Inform. Sci. Technol. 43, 3-71 (2009)
5. Sheridan, R.P., Kearsley, S.K.: Why Do We Need So Many Chemical Similarity Search Methods? Drug Discov. Today 7, 903-911 (2002)
6. Nikolova, N., Jaworska, J.: Approaches to Measure Chemical Similarity - a Review. QSAR Combin. Sci. 22, 1006-1026 (2003)
7. Maldonado, A.G., Doucet, J.P., Petitjean, M., Fan, B.-T.: Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. Mol. Diversity 10, 39-79 (2006)
8. Glen, R.C., Adams, S.E.: Similarity Metrics and Descriptor Spaces - Which Combinations to Choose? QSAR Combin. Sci. 25, 1133-1142 (2006)
9. Sheridan, R.P.: Chemical Similarity Searches: When Is Complexity Justified? Expert Opin. Drug Discov. 2, 423-430 (2007)
10. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A.: Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbour Information. J. Med. Chem. 48, 7049-7054 (2005)
11. Johnson, M.A., Maggiora, G.M. (eds.): Concepts and Applications of Molecular Similarity. John Wiley, New York (1990)
12. Martin, Y.C., Kofron, J.L., Traphagen, L.M.: Do Structurally Similar Molecules Have Similar Biological Activities? J. Med. Chem. 45, 4350-4358 (2002)
13. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A.: Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. J. Chem. Inf. Comput. Sci. 44, 1177-1185 (2004)
14. Whittle, M., Gillet, V.J., Willett, P., Alex, A., Loesel, J.: Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. J. Chem. Inf. Comput. Sci. 44, 1840-1848 (2004)
15. Willett, P.: Data Fusion in Ligand-Based Virtual Screening. QSAR Combin. Sci. 25, 1143-1152 (2006)
16. Goldman, B.B., Walters, W.P.: Machine Learning in Computational Chemistry. Ann. Report. Comput. Chem. 2, 127-140 (2006)
17. Willett, P., Winterman, V.: A Comparison of Some Measures of Inter-Molecular Structural Similarity. Quant. Struct.-Activ. Relat. 5, 18-25 (1986)
18. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A.: New Methods for Ligand-Based Virtual Screening: Use of Data-Fusion and Machine-Learning Techniques to Enhance the Effectiveness of Similarity Searching. J. Chem. Inf. Model. 46, 462-470 (2006)
19. Gardiner, E.J., Gillet, V.J., Haranczyk, M., Hert, J., Holliday, J.D., Malim, N., Patel, Y., Willett, P.: Turbo Similarity Searching: Effect of Fingerprint and

Dataset on Virtual-Screening Performance. Stat. Anal. Data Mining, in press (2009)

20. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Inf. Proc. Manag. 24, 513-523 (1988)
21. Cramer, R.D., Redl, G., Berkoff, C.E.: Substructural Analysis. A Novel Approach to the Problem of Drug Design. J. Med. Chem. 17, 533-535 (1974)
22. Ormerod, A., Willett, P., Bawden, D.: Comparison of Fragment Weighting Schemes for Substructural Analysis. Quant. Struct.-Activ. Relat. 8, 115-129 (1989)
23. Siegel, S., Castellan, N.J.: Nonparametric Statistics for the Behavioural Sciences. McGraw-Hill, New York (1988)