Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches

Hanna Eckert and Jürgen Bajorath

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2. D-53113 Bonn, Germany

The success of ligand-based virtual-screening calculations is influenced highly by the nature of target-specific structure–activity relationships. This might pose severe constraints on the ability to recognize diverse structures with similar activity. Accordingly, the performance of similarity-based methods strongly depends on the class of compound that is studied, and approaches of different design and complexity often produce, overall, equally good (or bad) results. However, it is also found that there is often little overlap in the similarity relationships detected by different approaches, which rationalizes the need to develop alternative similarity methods. Among others, these include novel algorithms to navigate high-dimensional chemical spaces, train similarity calculations on specific compound classes, and detect remote similarity relationships.

Introduction

Essentially, every small molecule-based approach to either designing or identifying novel active compounds focuses on the exploration of ’molecular similarity’, albeit often from different points of view. Methods to analyze pharmacophores [1] or quantitative structure–activity relationships (QSARs) [2] focus on ‘local’ similarities when studying molecular determinants of biological activity, such as functional groups and their specific geometric arrangements and/or resulting chemical properties. By contrast, molecular-similarity analysis, as we understand it today, originates from the ‘similar property principle’ (SPP) [3] and employs a ‘global’ or ‘holistic’ molecular view. The appropriateness of such viewpoints is related directly to the nature of structure–activity relationships (SARs) that characterize biologically active molecules and present crucial determinants for the success of ligand-based virtual-screening (LBVS), irrespective of the methods used. Thus, fundamental considerations of molecular-similarity concepts are likely to be as important as the design of novel computational approaches. Accordingly, in this review we provide both insights into crucial aspects of molecular similarity and review some of the novel methodological developments in LBVS, including methods that explore high-dimensional chemical reference spaces or add activity class-specific training to similarity searching.

Molecular similarity and SARs

The SPP states that molecules that are similar overall should have similar biological activity [3]. Although this concept is intuitive and supported by many observations, medicinal chemists also know that small chemical changes in an active molecule can render it either nearly or completely inactive or increase its activity dramatically [4]. This situation provides the basis for lead-optimization efforts. Clearly, reasons for this apparent inconsistency must include fundamental differences in underlying SARs.

In a recent editorial [5], Gerry Maggiora, one of the pioneers of molecular similarity analysis, commented on the different nature of SARs in the light of limitations in the accuracy of QSAR models. He described molecular ‘activity landscapes’ as akin to either gently rolling hills or rugged canyons where the presence of ‘activity cliffs’ is likely to cause errors in QSAR modeling. Similarly, we can also rationalize SARs as either ‘continuous’ or ‘discontinuous’ in nature. In the presence of gently rolling hills, or continuous SARs, small changes in molecular structure will cause small effects on activity and the ‘biological activity radius’ will be populated by a spectrum of increasingly diverse structures of similar activity. Figure 1 shows an example of a continuous SAR in which the structural similarity of active compounds gradually ‘fades away’ when departing from known leads. Such SARs are consistent with the SPP and the holistic view of molecular similarity. This is in contrast to discontinuous SARs, where small

Corresponding author: Bajorath, J. (bajorath@bit.uni-bonn.de)

1359-6446/06/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2007.01.011

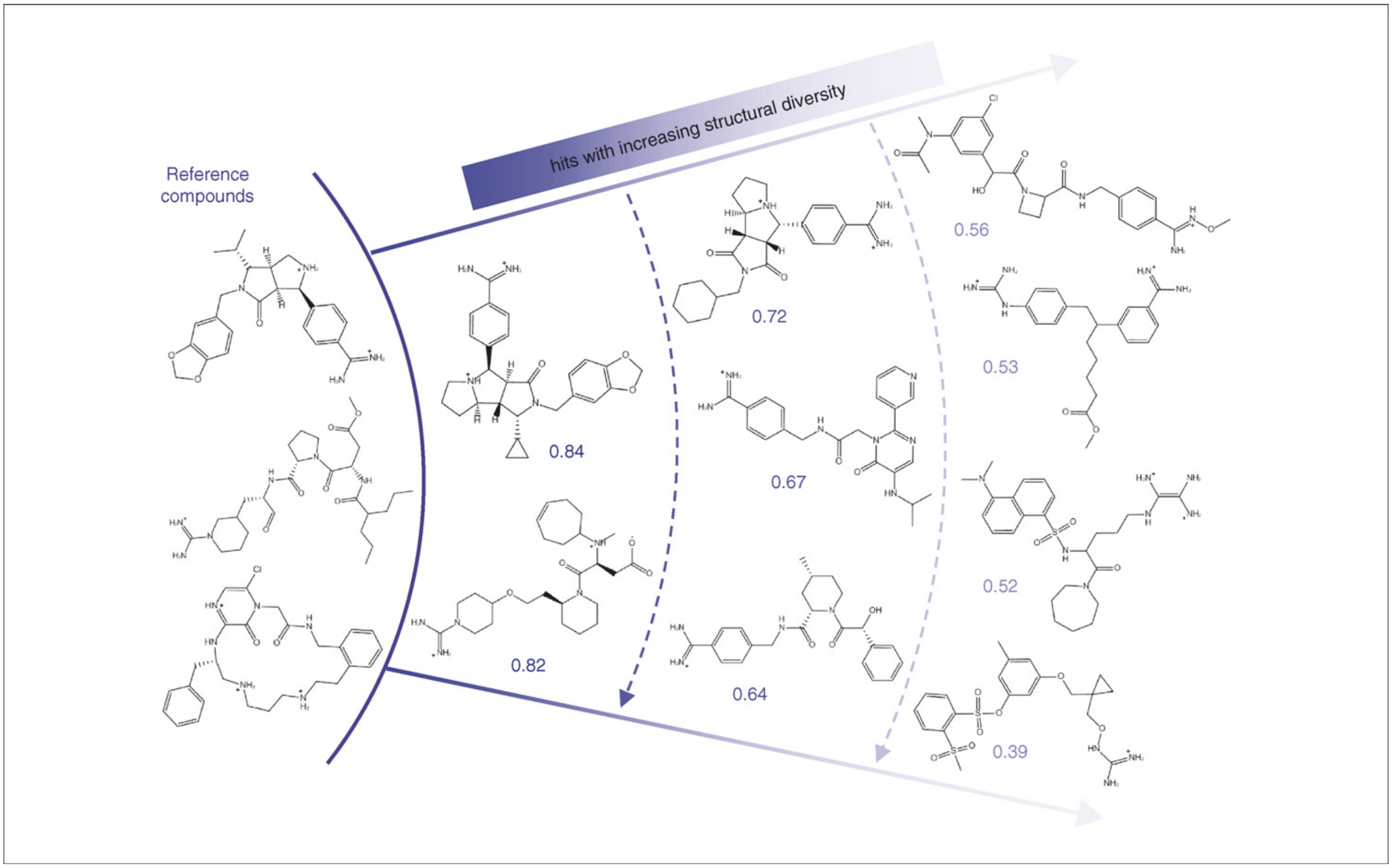


FIGURE 1

Structural spectrum of thrombin inhibitors. Starting with known inhibitors of thrombin, a sequence of hits ranging from close analogs to increasingly diverse structures was identified in a simulated LBVS campaign in a large screening database containing ~1.4 million compounds. Left: three of the five reference molecules. Right: examples of hits identified in a selection set of 250 database compounds. Hits are arranged in layers of increasing structural diversity (top down, from left to right). As a measure of structural similarity, the Tanimoto coefficient (Tc) [29] is reported for each hit relative to the most similar of the five reference molecules. Reference molecules and hits were compared using a fingerprint consisting of the publicly available set of 166 MACCS structural keys (MDL Elsevier). Virtual screening calculations used the DynaMAD algorithm described in the text.

changes in structure have dramatic effects. For LBVS, recognizing increasingly diverse structures that have similar activity is a major goal [6].

Molecular similarity and target–ligand interactions

The principles of molecular similarity can be evaluated further, going beyond the small molecule-centric view. The specific biological activity of a synthetic molecule is, first and foremost, the result of well-defined interactions with a macromolecular target, most often a protein, and small-molecule SARs should take into account knowledge of specific protein–ligand interactions [7]. From structural biology and structure-based design we know that protein–ligand interactions are determined by the formation of specific interactions of different chemical nature (i.e. polar/charged, hydrophobic/aromatic), a high degree of shape complementarity, and other entropic effects. It is also known that a single interaction, such as a hydrogen bond, can dramatically alter the selectivity and/or potency of a compound.

Given the structural constraints on specific protein–ligand interactions, why are not all small molecule SARs discontinuous? Or are they? Clearly, any gently rolling activity landscape will, ultimately, face a 'cliff' (probably more than one) that corresponds to structural alterations that either abolish or increase specific binding. Thus, any activity radius, as illustrated in Figure 1, has its boundaries. However, the architectures of protein–ligand complexes are not determined entirely by rules of static molecular engineering. Many ligand-binding sites are characterized by a degree of structural plasticity, and even similar ligands can differ in binding conformation and/or orientation [8]. Also, structurally distinct ligands can be accommodated in an adaptable binding site, as illustrated in Figure 2. In this example, the structural divergence of inhibitors of HIV reverse transcriptase inhibitors can be rationalized on the basis of their different binding modes. In principle, the specific binding of different structural motifs to a target site is often indicative of a continuous SAR.

Heterogeneous SARs

What do we conclude from the above considerations? At the atomic level, individual interactions that are crucial for the formation of protein–ligand complexes introduce cliffs in activity landscapes but do not transform rolling hills into desert canyons. 'All-or-none' binding events are rare, because most binding sites accommodate at least some analogs of active compounds and are permissive to structural variations. Therefore, SARs should, in

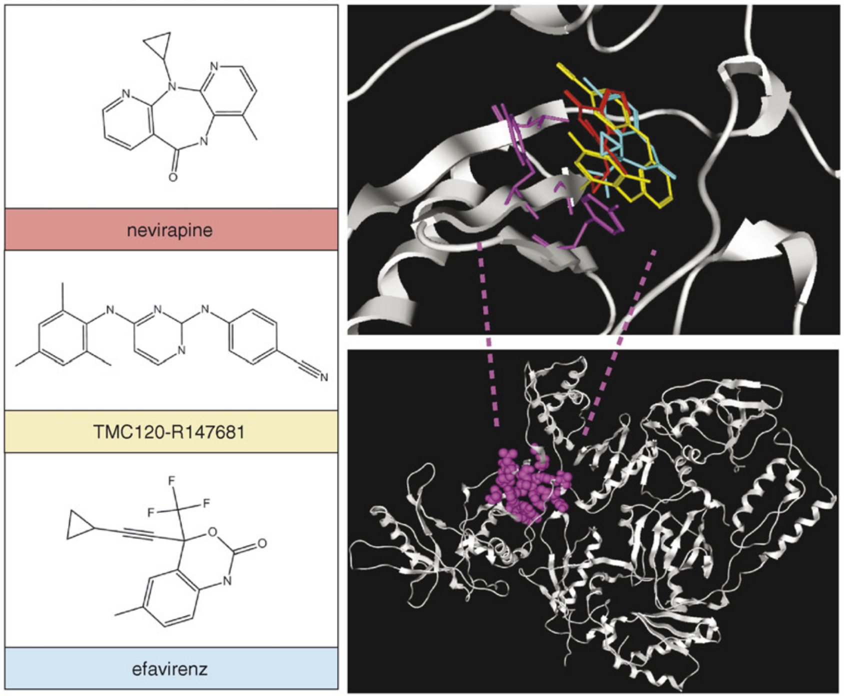


FIGURE 2

Binding modes of non-nucleoside reverse transcriptase inhibitors (NNRTIs). Lower right: the structure of HIV-1 reverse transcriptase (pdb id '1vrt') and the location of the binding pocket (purple) of NNRTIs. Upper right: close-up view comparing the binding of three NNRTIs (based on superposition of the enzyme). Nevaripine (pdb id '1vrt'), a first-generation NNRTI, is in red and exhibits a 'butterfly-like' binding mode that is different from that of efavirenz (pdb id '1fk9') in blue. TMC120-R147681 (pdb id '1s6q'), in yellow, is a diarylpyrimidine (DAPY) analogue with a binding mode that differs from nevaripine and efavirenz. Moreover, DAPY analogues change conformation, and reorient and reposition themselves when natural mutations alter the shape of NNRTI binding pocket (for example, of residues shown in purple). Typically, such mutations lead to high resistance against nevaripine but do not prevent the DAPY analogues from binding with EC_{50} values of <0.01 μM [53].

principle, be heterogeneous in nature and their activity surfaces should contain both flat (continuous) and steep (discontinuous) regions. Figure 3 shows an example of a heterogeneous SAR. Different structural motifs represent potent, selective tyrosine kinase inhibitors (within a flat region of the activity landscape), but close analogs of these molecules have dramatically reduced potency (indicating the proximity of cliffs).

If an activity surface contains flat regions, the SSP applies and molecular-similarity calculations that focus on these regions, through the use of appropriate reference molecules, can be expected to identify different structures with similar activity. By contrast, in steep regions and close to cliffs, similarity analysis is meaningless. Figure 3 presents an extreme example: a similarity method must recognize the two nearly identical analogs on the right in this figure as being 'similar', although one of them is essentially inactive.

Implications for similarity methods and their relative performance

The proposed presence of differently balanced and heterogeneous SARs helps to explain the success of molecular similarity analysis [6], although similarity calculations have limitations in many, if not all, cases. If SARs are predominantly discontinuous and their activity radii are small, similarity methods are likely to fail, irrespective of their complexity or specific features, and one of the unsolved key problems is how to select reference molecules that focus similarity analysis on sparse continuous segments. Other questions include why does the relative performance of similarity methods generally depend on compound classes [9], and why do all methods not succeed (or fail) equally? These questions can be answered by taking into account that the nature of SARs depends on the chosen molecular representations and the reference spaces into which compound sets are projected. Generally, different similarity methods rely on different descriptors, representations and reference spaces, and the SAR landscapes of compound activity classes are influenced strongly by changing reference frames. This observation has been made, for example, when selecting compounds for follow-up evaluation after initial high-throughput screening [10]. Different molecular representations that generated distinct chemical spaces produced candidate compound lists with only ~15% overlap. After compound screening, most of the newly identified hits were selected by only one of the alternative methods.

Developing an understanding of the complications described above is helpful in putting the opportunities and limitations of similarity calculations into perspective. One possible conclusion is that it is worth continuing to design novel algorithms and similarity methods to benefit from their complementary nature and to explore SARs more thoroughly. Therefore, in the second part of this review we discuss some novel developments that depart from conventional schemes.

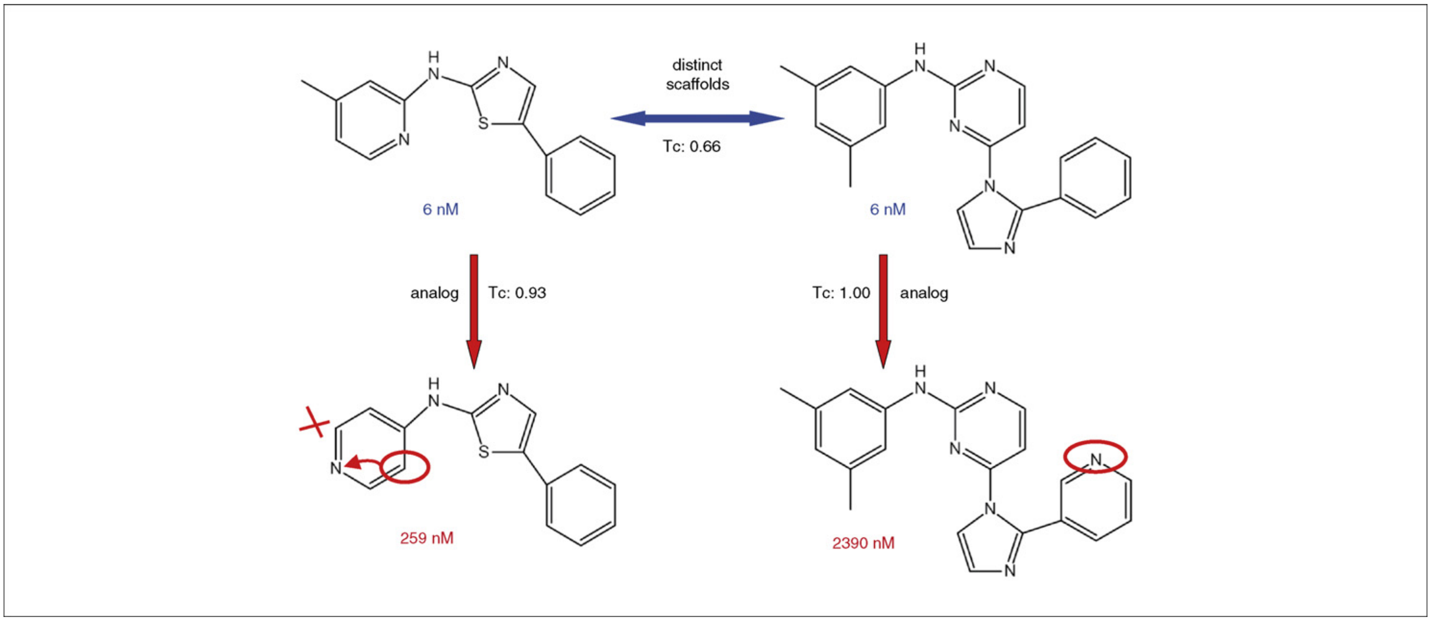


FIGURE 3

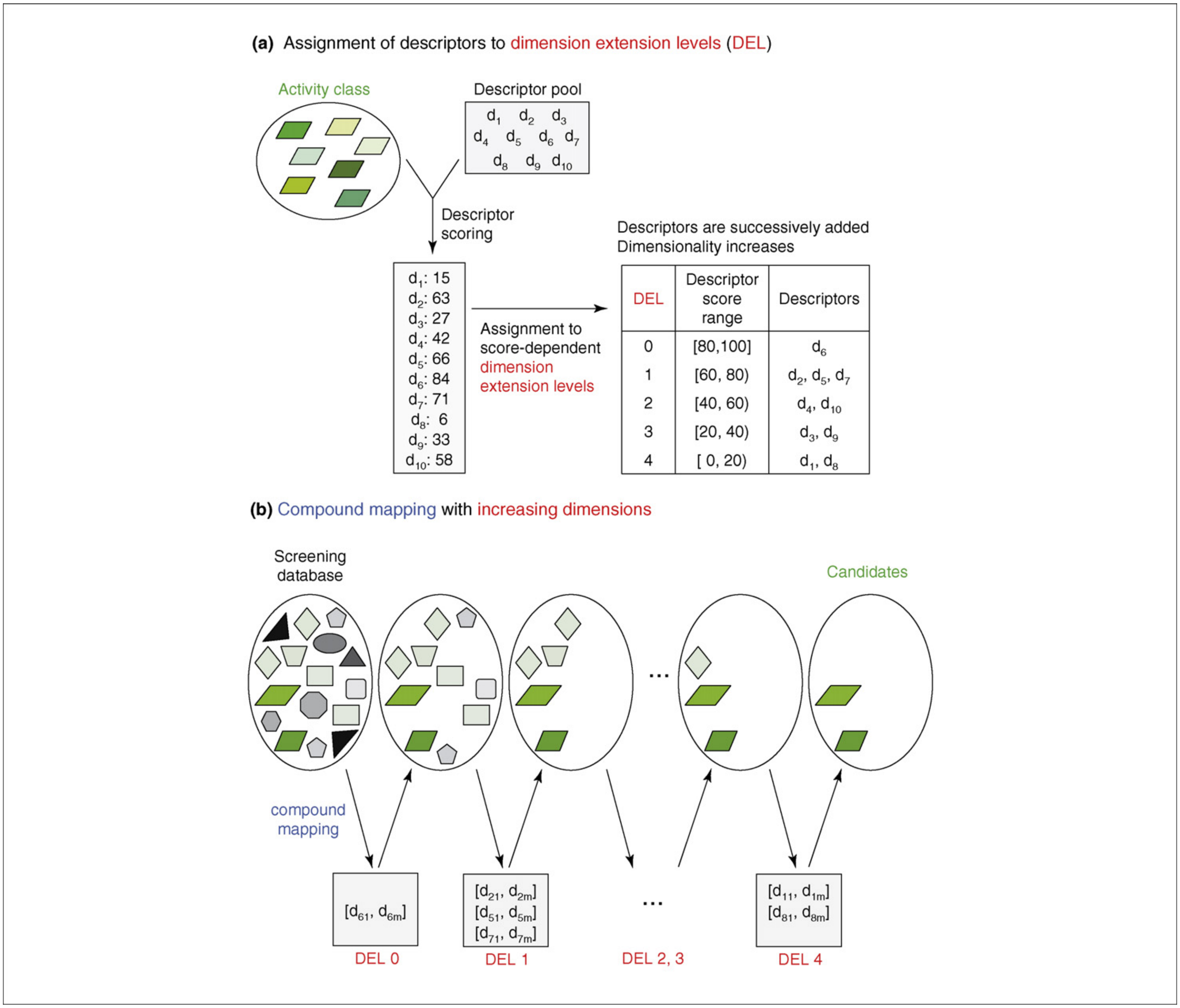
Heterogeneous SARs. Shown are four vascular endothelial growth-factor receptor (VEGFR-2) tyrosine kinase inhibitors with different structures and potencies. The two inhibitors at the top are potent and bind with IC_{50} values of 6 nM, although they have different core structures. However, subtle structural modifications of each inhibitor decrease their potency by two to three orders of magnitude. MACCS Tc values are reported for pairwise structural comparisons.

High-dimensional similarity methods

Because of the pioneering contributions of Pearlman [11], Agrafiotis [12] and others, calculations in low-dimensional reference spaces have become a paradigm in chemoinformatics. Advantages of low-dimensional space representations include, for example, creation of orthogonal reference frames, removal of descriptor correlation effects, controlled occupancy of subregions or cells in chemical space, ease of interpretation of compound distributions and possible visualization without substantial loss of information. Accordingly, one might ask whether low-dimensional space representations are essential for the success of molecular similarity analysis or virtual screening. The answer is no. Clustering or partitioning algorithms have been applied using a relatively large number of descriptors and a few methods have been developed specifically to navigate high-dimensional descriptor spaces. For example, there has been much interest in support vector machines (SVMs) for compound classification and class label predictions [13,14]. Methods of molecular similarity analysis utilize information that is provided predominantly by active compounds, whereas machine learning techniques such as SVM require training sets that include active and inactive molecules. Initially, the SVM method projects compounds as descriptor vectors into high-dimensional spaces and then constructs a maximum-margin hyperplane by linear combination of training set vectors to optimally separate two classes of compounds (active/inactive). If no linear separation of the training classes is possible a 'kernel trick' is applied that introduces additional dimensions to enable linear classification in the transformed space. In class label predictions, SVMs currently achieve at least 80% prediction accuracy, which makes them attractive for binary compound classification. A recent study [14] adopted SVMs for virtual screening: the SVM classification function was modified to generate real numbers instead of yes/no decision values. These numbers were then used to rank a screening database, providing a strategy that was more effective at enriching selection sets for active compounds with 'novel chemistries' than fingerprint-based methods.

Novel, high-dimensional similarity methods have been a focus in our laboratory and we have pursued two avenues, the design of distance functions [15,16] and mapping algorithms [17–20]. A surprising finding has been that a simple distance function can successfully capture SARs in unrefined descriptor spaces of >100 dimensions. The approach, called distance in activity-centered chemical space (DACCS) involves a scaling procedure that centers high-dimensional descriptor spaces on a subspace populated by a set of active compounds and then superimposes an approximated orthogonal coordinate system onto this subspace [15]. As a measure of similarity, DACCS calculates Euclidian-like distances from the center of the 'active subspace' to compounds in a screening database, and generates a distance-based ranking of candidates that corresponds to decreasing molecular similarity. Through Bayesian modeling, the DACCS function has also been transformed into a likelihood estimate (BDACCS) where the probability that a molecule is active decreases with its distance from the active subspace [16]. These distance functions have been tested on >50 compound classes and performed either as well as or better than different 2D fingerprints (which also produce a similarity ranking of test compounds) [16].

Mapping algorithms originated with the introduction of dynamic mapping of consensus positions (DMC), a method to determine and iteratively refine consensus positions of classes of active compounds in simplified descriptor spaces of gradually increasing dimensionality [17]. Database compounds that match these positions are selected as candidates for hit identification. Subsequently, a potency scaling function was implemented in DMC that tunes search calculations towards the recognition of potent database hits [18]. The idea behind operating in descriptor spaces of progressively increasing dimensionality is that irrelevant database compounds are removed iteratively from activity-dependent consensus positions until only similar compounds remain. Therefore, it is not necessary to determine or ‘guess’ an appropriate number of dimensions before the calculations. Second-generation mapping algorithms, such as mapping to activity-selective descriptor value ranges (MAD) [19] and dynamic MAD (DynaMAD) [20], operate in high-dimensional spaces without the need to simplify the representations, similar to distance functions. Crucially important for the development of these algorithms is the finding that activity class-selective value ranges can be systematically identified for many molecular property descriptors [19]. The MAD algorithm utilizes a predefined number of such descriptors to map database compounds to multiple value ranges that have a selective tendency for a given activity class, and ranks database compounds according to the proportion of ranges that they match. DynaMAD then adds a DMC-like dimension extension routine to MAD to map molecules to descriptor value ranges in spaces of increasing dimensionality. The DynaMAD approach is summarized in Figure 4. In benchmark calculations, mapping algorithms recognized diverse structures having similar activity, as also illustrated in Figure 1. Furthermore, a characteristic feature of mapping algorithms is enrichment of active compounds in small selection sets. For example, MAD frequently recovered >50% of potential hits from a source database containing ~1.34 million molecules in selection sets of fewer than 50 database compounds [19].



DEL	Descriptor score range	Descriptors
0	[80,100]	d_6
1	[60, 80)	d_2, d_5, d_7
2	[40, 60)	d_4, d_{10}
3	[20, 40)	d_3, d_9
4	[0, 20)	d_1, d_8

FIGURE 4

Schematic outline of DynaMAD. **(a)** Initially, the algorithm uses known, active compounds to score descriptors according to the presence of activity class-selective value ranges and then assigns descriptors to different scoring layers or dimension extension levels (DEL). Descriptors that fall within a specified score interval represent a scoring layer. During dimension extension descriptors of the next scoring layer are added, which increases the dimensionality of the chemical reference space in a stepwise manner. **(b)** Database molecules are then iteratively mapped to the activity-selective descriptor value ranges of each scoring layer and only the compounds that match all value ranges qualify for the next dimension extension step. Dimension extension and compound mapping are continued until a small compound selection set is obtained.

Virtual screening using mapping algorithms

In general, retrospective benchmarking has limited value in evaluating compound identification methods. Ultimately, the ability to identify novel active compounds needs to be assessed. Therefore, practical applications are as important as novel developments to advance the virtual-screening field. In collaboration with a pharmaceutical company, a parallel virtual screen using mapping algorithms was carried out for antagonists of an ion channel. Of the ~6 million compounds screened with DMC, MAD and DynaMAD, 76 compounds were selected for testing. This small test-set contained three active molecules, one with (undesired) agonist activity that was identified with MAD, and two antagonists, one identified using both MAD and DynaMAD and the other with DMC. Both antagonists were active in the low micromolar range and structurally distinct from the reference molecules. These studies confirm the ability of mapping algorithms to identify active compounds in small selection sets.

These findings also reflect a trend observed when similarity methods are applied successfully: typically novel hits are active in the micromolar range and are not highly potent. This situation is rationalized by considering principles of molecular similarity analysis, as discussed above. Reference molecules for similarity calculations are usually optimized and, thus, highly potent molecules. Because the goal is to depart from optimized structural motifs and identify different structures with similar activity, novel hits are not optimized for potency and are most likely to represent starting points for a new optimization effort.

Molecular fingerprints

Since the early days of chemoinformatics, there has been much debate whether 2D or 3D descriptors and methods are superior [21]. This discussion continues [22,23] and, depending on the test cases, different conclusions are often drawn. For fingerprints, which are bit-string representations of molecular structure and properties, the dimensionality of encoded descriptors has also been studied intensely and, in this case, some firm conclusions can be drawn: 2D fingerprints are often powerful similarity search tools [6], and even simple search strings and atom count vectors recognize active compounds successfully [24,25]. It is, therefore, not surprising that 2D-similarity searching continues to be a topic in chemo-informatics research. State-of-the-art 2D fingerprints include, for example, hashed connectivity pathways [26], and structural dictionary-based [27] and layered atom environment fingerprints [28]. In many publications and for historical reasons, daylight fingerprints [26] are used as a standard for benchmarking. Scientifically, it is difficult to accept any 2D fingerprint as a standard for similarity searching.

Originally, 2D fingerprints were developed for similarity searching using single template molecules, but independent studies have shown that search performance is enhanced if multiple reference compounds are used [29]. Preferably, all templates are known actives. But even molecules found to be most similar to a single reference compound in an initial similarity search can be included, irrespective of activity. This is known as 'turbo' similarity searching [30]. Recent investigations to increase fingerprint search performance utilizing multiple reference compounds have much concentrated on strategies to either scale [31] or average [32] fingerprints and on the evaluation of alternative scoring schemes, in particular, nearest neighbor methods [32] and data fusion [33]. In data fusion and nearest neighbor methods, similarity values are determined individually for each available reference compound and for each database compound, the similarity score is either calculated as the average similarity against a pre-specified number of nearest neighbors in the reference set or as the maximum [32]. The latter approach is termed 1-NN or 'sum fusion rule' and has often produced the best results in comparative studies [33]. However, nearest neighbor methods, in particular 1-NN, might have the drawback that they often show less potential to identify structurally diverse active compounds than methods utilizing multiple compound information as a whole [34]. For an extensive discussion of similarity coefficients and data fusion techniques, see the recent review by Peter Willett [35].

Another fingerprint-based, machine-learning methodology that is used increasingly for either compound classification or LBVS is binary kernel discrimination (BKD) [36]. Following this approach, binary fingerprints are used to estimate the probability that a molecule is active. Fingerprint bit positions that differ between pairs of test molecules are determined as input for a kernel function to derive probability density functions for known active and inactive compounds. These density functions are then used to estimate the probability of whether a molecule is active, based on its fingerprint settings. In benchmark calculations, BKD compared favorably to other multiple-template, fingerprint-based methods [33].

In contrast to the development of different fingerprint-search strategies discussed above, few novel types of 2D fingerprints have been designed in recent years. One approach follows an original idea of experimentally fingerprinting a test compound against a panel of proteins [37]. This process was mimicked by docking of compounds into arrays of protein-binding sites and scoring them as a measure of similarity [38]. A recent study extends this approach exclusively to the ligand level and explores the similarity of a test molecule to different compound activity classes using Bayesian modeling, which generates a quasi-fingerprint that consists of the resulting activity class scores [39].

Recent 2D-fingerprint designs include Molprint 2D, a complex atom environment fingerprint consisting of up to 2^{50} theoretically possible strings [28], and property descriptor value range-derived fingerprint (PDR-FP), a low-complexity fingerprint consisting of 500 bits [40]. PDR-FP is designed specifically for similarity searching using multiple reference compounds. It encodes the screening database value ranges of property descriptors that display a general tendency to respond to compound activity classes through a process termed 'equifrequent' binning. This procedure divides the value range into several intervals, such that each interval (bit) is matched by the same number of database compounds. For every test compound and descriptor, exactly one bit is set and the constant bit setting renders PDR-FP calculations independent of molecular size. For a series of reference compounds, an activity class-specific search string is created by recording their bit frequencies in PDR-FP. Bit positions with high frequencies indicate substantial deviations between active reference compounds and database compounds. Thus, the activity-dependent search string represents a fingerprint trained on a given activity class that is then compared with bit strings of individual database compounds. PDR-FP performed better on structurally diverse active compounds than other 2D fingerprints [40]. In this study, PDR-FP also produced meaningful results on peptide-like molecules, which are notoriously difficult for similarity searching.

3D similarity methods

Many of the methodologies described herein can make use of either 2D or 3D molecular descriptors. In addition, exclusive 3D-similarity methods have been developed, including shape-matching algorithms [41], shape-based fingerprints [42], fuzzy, 3D-feature representations derived from cluster analysis of molecular conformations [43], molecular field descriptors [44] and pharmacophore fingerprints [45]. These fingerprints systematically monitor potential pharmacophore arrangements in molecules and, thereby, transform local molecular views into a global view. Ensemble pharmacophore methods make 3D-similarity searching independent of detailed predictions of the conformation of a bioactive compound, which continues to be a major bottleneck for meaningful applications of many 3D descriptors and methods.

Descriptor-independent similarity methods

Although most similarity methods depend on the use of predefined chemical descriptors and chemical reference spaces, a few approaches do not. These include string-based similarity searching [46], reduced (or simplified) 2D molecular graph representations for similarity searching [29,47] and the MolBlaster methodology [48]. Reduced graphs of molecules can be compared for overlap as a similarity criterion. By contrast, systematic matching of combinations of nodes and edges in regular graphs is computationally unfeasible on a large scale (because of the combinatorial problem that is commonly referred to as ‘subgraph isomorphism’). MolBlaster produces random-fragment profiles of molecules from their connectivity tables and records them in histogram representations. The fragmentation scheme is reminiscent of, but distinct from, mass spectrometry. MolBlaster fragment profiles of different molecules are compared quantitatively using information entropic metrics as a measure of molecular similarity. Histogram comparisons accurately reproduced similarity-based compound rankings that were generated using 2D fingerprints [48], and the approach has also been adopted for large-scale LBVS [49].

Conclusions and future perspectives

There is no doubt that the concept of ‘molecular similarity’ is more complex than it might first appear, and there is room for further investigations and developments. Although the postulated heterogeneous nature of many SARs is, in part, consistent with the SPP, it puts severe constraints on molecular similarity analysis. Regardless of this, similarity-based methods are cornerstones of

TABLE 1

Classification of representative LBVS methods

Method	Descriptors	Approach	Refs
Compound classification			
Clustering	Continuous, also fingerprints	Groups compounds by means of distances in descriptor or fingerprint space	[29]
Cell-based partitioning	Continuous, binned	Compounds are mapped to subsections of chemical space	[11]
BKD	Binary transformed or fingerprints	Machine learning technique to estimate class label probabilities	[36]
SVM	Continuous, also fingerprints	Class label prediction using a maximum-margin hyperplane	[13,14]
Mapping algorithms and distance functions			
DACCS	Continuous	Determines compound distances from the center of an ‘active subspace’	[15]
BDACCS	Continuous	Transforms DACCS into a likelihood estimate	[16]
DMC	Binary transformed	Maps compounds to consensus bit positions in chemical space	[17,18]
MAD	Continuous	Maps compounds to activity-selective descriptor value ranges	[19]
DynaMAD	Continuous	Adds dimension extension to MAD	[20]
Fingerprints			
BCI	Predefined structural fragments	Quantitative comparison of bit strings using a similarity coefficient for single or multiple reference compounds; also input for kernel methods or clustering	[27]
Daylight	Hashed connectivity pathways		[26]
Molprint 2D	Layered atom environments		[28]
Shape fingerprint	Set of reference shapes		[42]
3D Pharmacophore fingerprint	Set of potential pharmacophore arrangements		[1,45]
Bayes affinity fingerprint	Conventional fingerprints	Generates quasi-fingerprints of Bayes scores for correlation analysis	[39]
PDR-FP	Continuous, binned to equifrequent intervals	Generates compound class-specific search strings for similarity evaluation	[40]
Others			
LINGO	SMILES substrings	Comparison of strings using similarity coefficient	[46]
MolBlaster	Random fragment profiles	Uses information entropic metrics to compare fragment profiles	[48,49]
Reduced graphs	Simplified 2-D molecular graphs	Directly compares simplified graph similarity	[29,47]
ROCS	Gaussian shape models	Determines volume overlap as a measure of similarity	[41]

Fingerprints are considered to be both methods and descriptors. If not specified, descriptors refer to molecular property descriptors.
Abbreviations: BCI, Barnard Chemical Information; ROCS, rapid overlay of chemical structure.

chemoinformatics and computer-aided pharmaceutical research. The conceptual diversity of successful methods is astonishing. Table 1 classifies the representative methods described here, and additional methods that have not been discussed in detail because of space constraints. Successful benchmark calculations are reported using low-dimensional and high-dimensional reference spaces, 2D and 3D descriptors, and methods of low- and high-computational complexity. However, in most cases the results depend strongly on the test cases used. Considering the spectrum of similarity methods, it looks as if 'anything goes'. Being optimistic, we might conclude that many powerful methodologies are available. An alternative conclusion is that many methods are successful because SARs are 'easy' to study, in qualitative terms at least. However, given the considerations above, this seems unlikely. Being more pessimistic, we might assume that the performance of diverse methods is limited principally by the nature of SARs, and that benchmark calculations provide artificial insights. Perhaps the only firm conclusion that can be drawn is that LBVS has identified novel, active compounds in many applications, which indicates that similarity methods do have substantial 'selectivity' in recognizing diverse, active compounds. Their principal limitation in 'real-life' applications is the tendency to detect false-positives [50], which corresponds to low 'specificity'. This indicates future directions for research. For example, systematic deselection of either inactive or irrelevant compounds is as important for the success of molecular similarity analysis as the selection of active compounds [51]. Furthermore, given the complex nature of many SARs and the inherent approximation of similarity methods, finding 'active needles in chemical haystacks' might not be the most promising application scenario of LBVS. Other applications, for example enriching moderately sized subsets of databases with desired compounds and, thereby, interfacing similarity analysis with biological screening are thought to be particularly attractive [6,52], and provide further opportunities for basic and applied research.

Acknowledgements

The authors are grateful to Lu Tan for her help in preparing Figure 2.

References

1 *Pharm. Des.* 7, 567–597

2 Esposito, E.X. *et al.* (2004) Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* 275, 131–214

3 Johnson, M. and Maggiora, G.M., eds (1990) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons

4 Kubinyi, H. (1998) Similarity and dissimilarity – a medicinal chemist's view. *Perspect. Drug Discov. Des.* 11, 225–252

5 Maggiora, G.M. (2006) On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model.* 46, 1535

6 Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894

7 Bender, A. and Glen, R.C. (2004) Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* 2, 3204–3218

8 Boström, J. *et al.* (2006) Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* 49, 6716–6725

9 Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* 7, 903–911

10 Shanmugasundaram, V. *et al.* (2005) Hit-directed nearest neighbour searching. *J. Med. Chem.* 48, 240–248

11 Pearlman, R.S. and Smith, K.M. (1998) Novel software tools for chemical diversity. *Perspect. Drug Discov. Des.* 9, 339–353

12 Agrafiotis, D.K. *et al.* (2002) Combinatorial informatics in the post-genomics era. *Nat. Rev. Drug Discov.* 1, 337–346

13 Warmuth, M.K. *et al.* (2003) Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* 43, 667–673

14 Jorissen, R.N. and Gilson, M.K. (2005) Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* 45, 549–561

15 Godden, J.W. and Bajorath, J. (2006) A distance function for retrieval of active molecules from complex chemical space representations. *J. Chem. Inf. Model.* 46, 1094–1097

16 Vogt, M. *et al.* Bayesian interpretation of a distance function for navigating high-dimensional descriptor spaces. *J. Chem. Inf. Model.* DOI:10.1021/ci600280b (http://pubs.acs.org/journals/jcisd8/index.html)

17 Godden, J.W. *et al.* (2004) Molecular similarity analysis and virtual screening in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* 44, 21–29

18 Godden, J.W. *et al.* (2004) POT-DMC: a virtual screening method for the identification of potent hits. *J. Med. Chem.* 47, 5608–5611

19 Eckert, H. and Bajorath, J. (2006) Determination and mapping of activity-specific descriptor value ranges (MAD) for the identification of active compounds. *J. Med. Chem.* 49, 2284–2293

20 Eckert, H. *et al.* (2006) Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: design of DynaMAD and comparison with MAD and DMC. *J. Chem. Inf. Model.* 46, 1623–1634

21 Brown, R.D. and Martin, Y.C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 36, 572–584

22 Zhang, Q. and Muegge, I. (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem.* 49, 1536–1548

23 Nettles, J.H. *et al.* (2006) Bridging chemical and biological space: 'target fishing' using 2D and 3D molecular descriptors. *J. Med. Chem.* 49, 6802–6810

24 Wang, N. *et al.* (2005) Fast small molecule similarity searching with multiple alignment profiles of molecules represented in one-dimension. *J. Med. Chem.* 48, 6980–6990

25 Bender, A. and Glen, R.C. (2005) A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* 45, 1369–1375

26 James, C.A. and Weininger, D. (2006). Daylight theory manual, chapter 6. Daylight Chemical Information Systems

27 Barnard, J.M. and Downs, G.M. (1997) Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* 37, 141–142

28 Bender, A. *et al.* (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* 44, 1708–1718

29 Willett, P. (2005) Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* 48, 4183–4199

30 Hert, J. *et al.* (2005) Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information. *J. Med. Chem.* 48, 7049–7054

31 Xue, L. *et al.* (2003) Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.* 43, 1218–1225

32 Schuffenhauer, A. *et al.* (2003) Similarity metrics for ligands reflecting the similarity of the target protein. *J. Chem. Inf. Comput. Sci.* 43, 391–405

33 Hert, J. *et al.* (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* 44, 1177–1185

34 Tovar, A. *et al.* Comparison of 2D fingerprint methods for multiple-template similarity searching on compound classes of increasing structural diversity. *ChemMedChem.* 10.1002/cmdc.200600225 (http://www3.interscience.wiley.com/cgi-bin/jhome/110485305)

35 Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* 11, 1046–1053

36 Harper, G. *et al.* (2001) Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* 41, 1295–1300

37 Kauvar, L.M. *et al.* (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* 2, 107–118

38 Briem, H. and Lessel, U. (2000) *In vitro* and *in silico* affinity fingerprints: finding similarities beyond structural classes. *Perspect. Drug Discov. Des.* 20, 231–244

39 Bender, A. *et al.* (2006) Bayes affinity fingerprints'' improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *J. Chem. Inf. Model.* 46, 2445–2456

40 Eckert, H. and Bajorath, J. (2006) Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J. Chem. Inf. Model.* 46, 2515–2526

41 Rush, T.S., III *et al.* (2005) A shape-based 3D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* 48, 1489–1495

42 Haigh, J.A. *et al.* (2005) Small molecule shape-fingerprints. *J. Chem. Inf. Model.* 45, 673–684

43 Jenkins, J.L. *et al.* (2004) A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* 47, 6144–6159

44 Cheeseright, T. *et al.* (2006) Molecular field extrema as descriptors of biological activity: definition and validation. *J. Chem. Inf. Model.* 46, 665–676

45 Saeh, J.C. *et al.* (2005) Lead hopping using SVM and 3D pharmacophore fingerprints. *J. Chem. Inf. Model.* 45, 1122–1133

46 Grant, J.A. *et al.* (2006) Lingos, finite state machines, and fast similarity searching. *J. Chem. Inf. Model.* 46, 1912–1918

47 Gillet, V.J. *et al.* (2003) Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* 43, 338–345

48 Batista, J. *et al.* (2006) Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.* 46, 1937–1944

49 Batista, J. and Bajorath, J. (2007) Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *J. Chem. Inf. Model* 10.1021/ci600377m (http://pubs.acs.org/journals/jcisd8/index.html)

50 Lang, P.T. *et al.* (2005) Evaluating the high-throughput screening competitions. *J. Biomol. Screen.* 10, 649–652

51 Schreyer, S.K. *et al.* (2004) Data shaving: a focused screening approach. *J. Chem. Inf. Comput. Sci.* 44, 470–479

52 Parker, C.N. and Bajorath, J. (2006) Towards unified compound screening strategies: a critical evaluation of error sources in experimental and virtual high-throughput screening. *QSAR Comb. Sci.* 12, 1153–1161

53 Das, K. *et al.* (2004) Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *J. Med. Chem.* 47, 2550–2560