Methods 71 (2015) 58–63

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Molecular fingerprint similarity search in virtual screening

Adrià Cereto-Massagué [a], María José Ojeda [a], Cristina Valls [a], Miquel Mulero [a], Santiago Garcia-Vallvé [a,b], Gerard Pujadas [a,b,*]

[a] Group of Cheminformatics & Nutrition, Biochemistry and Biotechnology Department, Universitat Rovira i Virgili (URV), Campus de Sescelades, N4 Building, 43007 Tarragona, Catalonia, Spain
[b] Centre Tecnològic de Nutrició i Salut (CTNS), TECNIO, CEICS, Avinguda Universitat, 1, 43204 Reus, Catalonia, Spain

ARTICLE INFO

Article history:
Available online 15 August 2014

Keywords:
Fingerprints
Virtual screening
Similarity search
Data fusion
Comparison

ABSTRACT

Molecular fingerprints have been used for a long time now in drug discovery and virtual screening. Their ease of use (requiring little to no configuration) and the speed at which substructure and similarity searches can be performed with them – paired with a virtual screening performance similar to other more complex methods – is the reason for their popularity. However, there are many types of fingerprints, each representing a different aspect of the molecule, which can greatly affect search performance. This review focuses on commonly used fingerprint algorithms, their usage in virtual screening, and the software packages and online tools that provide these algorithms.

1. Introduction

Computational advances during the past two decades have enabled the extensive use of virtual screening for drug discovery [1]. Virtual screening is an *in silico* method that consists of screening large small-molecule databases for bioactive molecules. This enables the researcher to avoid the cost of experimentally testing hundreds or thousands of compounds by reducing the number of candidate molecules to be tested to manageable numbers.

The screening can be conducted using several methods or their combination, which can be classified as structure-based methods (which are based on matching the compounds to a target binding site, the most common of these approaches being protein–ligand docking) or ligand-based methods (which involves retrieving those compounds from the database that are similar in some ways to known active molecules and vary greatly depending on the molecular features taken into account for similarity assessment). The main ligand-based approaches involve the use of pharmacophores (abstractions of the features needed for the molecule to be active) [2], shape-based similarity [3], fingerprint similarity, and also machine learning using molecular properties and data from any of the former approaches [4].

Fingerprint-based similarity searching is also used outside of the virtual screening and drug discovery fields. One such example is the application of the method to flavor chemistry [5].

2. Methods for molecular fingerprints

Similarity in itself is subjective and can be measured and their results interpreted in several ways [6–8]. One of the most important problems encountered when trying to measure the similarity between two compounds is the complexity of the task, which depends on the complexity of the molecular representation used. In order to make the comparison between molecular representations computationally easier, some level of simplification or abstraction is required. The most commonly used of these abstractions are molecular fingerprints, which involve turning the molecule into a sequence of bits that can then be easily compared between molecules.

This comparison must then be expressed in a way that can be quantified. There are many ways to assess the similarity between two vectors, the most common overall being Euclidean distance. But for molecular fingerprints, the industry standard is the Tanimoto coefficient, which consists of the number of common bits set to 1 in both fingerprints divided by the total number of bits set to 1 between both fingerprints. This means that the Tanimoto coefficient will always have a value between 1 and 0, regardless of the length of the fingerprint, which causes it to loose representativity as the fingerprints become longer. This loss also means that how similar two fingerprints with a given Tanimoto coefficient actually will greatly depend on the type of fingerprint used, which makes it

* Corresponding author at: Group of Cheminformatics & Nutrition, Biochemistry and Biotechnology Department, Universitat Rovira i Virgili (URV), Campus de Sescelades, N4 Building, 43007 Tarragona, Catalonia, Spain.
E-mail address: gerard.pujadas@urv.cat (G. Pujadas).

http://dx.doi.org/10.1016/j.ymeth.2014.08.005
1046-2023/

Table 1
Some similarity coefficients and distances used with fingerprints.

Measure	Expression	Range
Tanimoto/Jaccard coefficient	$\frac{c}{a+b-c}$	0 to 1
Euclidean distance	$\sqrt{a+b-2c}$	0 to N
City-block/Manhattan/Hamming distance	$a+b-2c$	0 to N
Dice coefficient	$\frac{2c}{a+b}$	0 to 1
Cosine similarity	$\frac{c}{\sqrt{ab}}$	0 to 1
Russell–RAO coefficient	$\frac{c}{m}$	0 to 1
Forbes coefficient	$\frac{cm}{ab}$	0 to 1
Soergel distance	$\frac{a+b-2c}{a+b-c}$	0 to 1

Where, given the fingerprints of two compounds, A and B, m equals the total amount of bits present in the fingerprints, a equals the amount of bit set to 1 in A, b equals the amount of bits set to 1 in B and c equals the amount of bits set to 1 in both A and B.

impossible to select a universal cutoff criterion for determining whether two fingerprints are similar or dissimilar. However, the performance of molecular fingerprints could be improved by combining them with other similarity coefficients [9]. Several similarity and distance metrics that have been used with fingerprints are listed in Table 1.

2.1. Types of molecular fingerprint

There are several types of molecular fingerprints depending on the method by which the molecular representation is transformed into a bit string. Most methods use only the 2D molecular graph and are thus called 2D fingerprints; however, some methods are capable of storing 3D information, most notably pharmacophore fingerprints. The main approaches are substructure keys-based fingerprints, topological or path-based fingerprints, and circular fingerprints.

- Substructure keys-based fingerprints set the bits of the bit string depending on the presence in the compound of certain substructures or features from a given list of structural keys. This usually means that these fingerprints are most useful when used with molecules that are likely to be mostly covered by the given structural keys, but not so much when the molecules are unlikely to contain the structural keys, as their features would not be represented. Their number of bits is determined by the number of structural keys, and each bit relates to presence or absence of a single given feature in the molecule (Fig. 1), which does not happen with other (hashed) types of fingerprints. Some of the most commonly used substructure keys-based fingerprints are:
  - MACCS [10,11]: It comes in two variants, one with 960 and the other with 166 structural keys based on SMARTS patterns. The shorter one is the most commonly used, as it is relatively small in length (only 166 bits) but covers most of the interesting chemical features for drug discovery and virtual screening. Additionally several software packages are able to calculate it, which is not true for the longer version.
  - PubChem fingerprint [12]: this fingerprint, with 881 structural keys covers a wide range of different substructures and features. It is the fingerprint used by PubChem for similarity searching and neighboring. Other than PubChem's own code, it is also implemented in ChemFP [13] (although deemed "experimental") and in CDK [14,15].
  - BCI fingerprints [16]: BCI fingerprints can be generated using different numbers of bits and can be modified by the user in several ways, but the standard substructure dictionary includes 1052 keys [17]. BCI fingerprints are only available in BCI toolkits.
  - TGD [18] and TGT fingerprints: These are two-point and three-point pharmacophoric fingerprints calculated from a 2D molecular graph, consisting, respectively of 735 and 13,824 bits. TGD encodes atom-pair descriptors using seven-atom features and distances up to 15 bonds [17,18]. TGT encodes triplets of four-atom features using three graph distances divided into six distance ranges [17]. They are both available in MOE software package [19].
- Topological or path-based fingerprints work by analyzing all the fragments of the molecule following a (usually linear) path up to a certain number of bonds, and then hashing every one of these paths to create the fingerprint (Fig. 2). This means that any molecule can produce a meaningful fingerprint, and its length can be adjusted. They can also be used for fast substructure searching and filtering. These are hashed fingerprints, which means that a single bit cannot be traced back to a given feature. A given bit may be set by more than one different feature, which is called "bit collision". The Daylight fingerprint [20]: is the most prominent of these types of fingerprints. They consist of up to 2048 bits and encode all possible connectivity pathways through a molecule up to a given length. Most software packages implement these fingerprints or fingerprints based on them, which can sometimes reach higher number of bits or use non-linear connectivity paths, such as OpenEye's Tree fingerprints [21].

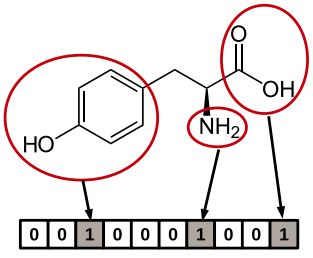

Fig. 1. A representation of a hypothetical 10-bit substructure fingerprint, with three bits set because the substructures they represent are present in the molecule (circled).

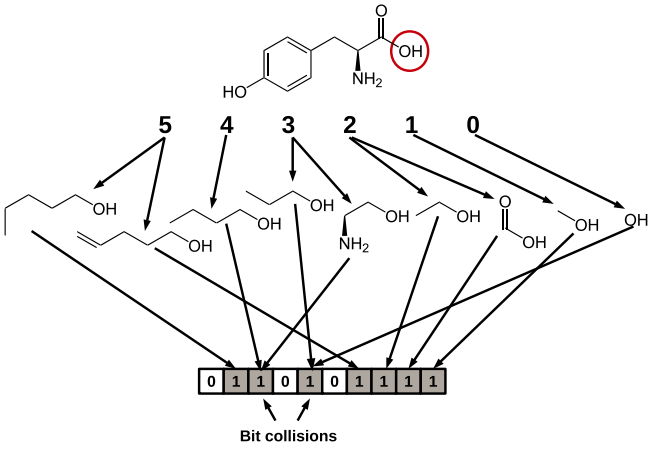

Fig. 2. A representation of a hypothetical 10-bit topological fingerprint, in this case a linear path-based fingerprint with fragments up to a length of 5. All fragments found from the starting atom (circled) are shown, and the fragment length and corresponding bit in the fingerprint are indicated. There are two bit collisions, which are bits that are set by more than one fragment; these are likely in fingerprints with a reduced number of bits. Only fragments and bits for a single starting atom are shown; for the full fingerprint, this process would be carried out for every atom in the molecule. Circular fingerprints use a similar approach, but building fragments within a radius of the starting atom instead of linear fragments.

- Circular fingerprints are also hashed topological fingerprints, but they are different in that instead of looking for paths in the molecule, the environment of each atom up to a determined radius is recorded. They are therefore not suitable for substructure queries (as the same fragment may have different environments) but are widely used for full structure similarity searching.
  - Molprint2D [22,23]: Molprint2D encodes the atom environments of each atom of the molecular connectivity table, which are represented by strings of varying size. This fingerprint is available in several software packages, such as Open Babel [24] and jCompoundMapper [25].
  - ECFP: The *de facto* standard circular fingerprints are the Extended-Connectivity Fingerprints (ECFPs), based on the Morgan algorithm [26], which were specifically designed for their use in structure–activity modeling [27]. They represent circular atom neighborhoods and produce fingerprints of variable length. They are most commonly used with a diameter of 4 and referred to as ECFP4. A diameter of 6 (ECFP6) is also commonly used, although some benchmarks have shown small performance differences between the two [28]. Additionally, there is a variation that keeps track of the frequency counts of the ECFP features, recording each identifier as many times as it appears in the molecule instead of only once. This variation is often denoted as ECFC. Notable software programs that provide these fingerprints are Pipeline Pilot [29], Chemaxon's JChem [30], the CDK [14] and the RDKit [31] (referred to as "Morgan fingerprints").
  - FCFP (Functional-Class Fingerprints): FCFP are a variation of ECFP, which are further abstracted in that instead of indexing a particular atom in the environment, they index that atom's role. So, different atoms or groups with the same or similar function are not distinguished by the fingerprint. This enables them to be used as pharmacophoric fingerprints. There is also a FCFC variation, akin to the ECFC variation to the ECFP. All major software packages supporting ECFP fingerprints also support these variations.
- There are also some hybrid fingerprints that combine the same bits string bits set using different approaches. Some commonly used fingerprints that fall into this category are the following:
  - UNITY 2D [32]: This is a 988-bit long fingerprint based both on structural keys and connectivity path fragments.
  - MP-MFP [33]: MP-MFP is a 171-bit fingerprint with 110 bits set from structural keys and 61 bits set from property descriptors.
- Pharmacophore fingerprints are also commonly used. A pharmacophore represents the relevant features and interactions needed for a molecule to be active against a given target. Pharmacophoric fingerprints usually encode the information for the features from a list that a molecule presents, in a manner similar to substructure-key based fingerprints, but taking into account the distance between these features, usually classifying it using a list of distance ranges. In this way, 3D information can be encoded into the fingerprint [34].
- Lastly, there are also other types of fingerprints that try totally different approaches. For example, LINGO [35] and SMIfp [36] are fingerprints that are text-based and are calculated based on the canonical SMILES [37] of the molecule. Protein–ligand interaction fingerprints (PLIF), as their name suggests, encode information about protein–ligand interactions, such as hydrogen bonds, ionic interactions and surface contacts with their residue of origin [19]. Structural Interaction Fingerprint (SIFt) is also one of these fingerprints [38].

In general, fingerprints with longer bit strings have been found to perform better during similarity searching, because they contain an increased amount of stored information (due to a reduction of bit collision for hashed fingerprints) [39].

2.2. Software for fingerprint-based virtual screening

There are many software packages that can be used for fingerprint-based virtual screening, from whole drug discovery suites including fingerprint functionality to software libraries or tools centered specifically in dealing with fingerprints and similarity searching. Each software package supports a different set of fingerprints, and most of them implement fingerprints not present in any other package. However, the most commonly used fingerprinting algorithms can be found in most software packages. Here is a list of the main software packages used when doing ligand-based virtual screening with fingerprint similarity, in no specific order:

- OEChem TK: This OpenEye toolkit [21] is able to produce 166-bit MACCS, LINGO, Circular, Path (Daylight-like) and Tree (Daylight-like with non-linear, "tree" fragments) fingerprints. It has interfaces to C++, Java, Python, and C#.
- JChem from ChemAxon [30]: This is a java library that provides access to several hashed fingerprints, ECFP fingerprints with all their variants (ECFC, FCFP, FCFC), and pharmacophoric fingerprints. ChemAxon also provides packages for .NET and is usable in Python through cinfony [40].
- Open Babel [24,41]: This is a free and open-source cheminformatics toolkit, which implements MOLPRINT2D, 166-bit MACCS, a Daylight-like fingerprint (FP2), and 2 structural key fingerprints with 55 (FP3) and 307 bits. It can be used from C++, Java, Python, C#, and Perl.
- RDKit [31]: This is also a free and open-source cheminformatics toolkit that provides access to several fingerprints: 166-bit MACCS, "Topological" (Daylight-like), "Atom pairs" (based on the atomic environments and shortest path separations of every atom pair in the molecule [42]),"Morgan" (ECFP and its variations), "Torsion" (based on the topological torsion descriptor [43]), and "Layered" (an experimental topological fingerprint intended to make fingerprinting queries more straightforward). It is usable from C++, Python, Java, and C#.
- CDK [14,15,44]: This is another free and open-source toolkit, which features several fingerprints, the most notable being ECFP, LINGO, Daylight-like fingerprint, 166-MACCS, PubChem, and other structural keys fingerprints such as E-State [45] and Klekota–Roth [46]. It is a Java library but can be used in regular Python through cinfony [40].
- Indigo [47]: This is another free and open-source cheminformatics toolkit that offers several hashed fingerprints and their combination. It can be used from C++, Java, Python and C#.
- Cinfony [40,48] This is not a toolkit in itself and does not implement any fingerprint, but it gives the user access to several toolkits (Open Babel, RDKit, CDK, JChem, and Indigo) through a common API in Python and to some extent in Jython (JVM) and IronPython (.NET).
- ChemFP [13] This is a tool that can be used as a back-end database with either Open Babel, RDKit or OEChem, thus supporting most of their fingerprints, and implementing on top of that a 166-bit MACCS and a PubChem-like fingerprint. But what is special about Chemfp is its ability to store the fingerprints in a standard file format (FPS) and then to perform high-speed Tanimoto similarity searches. It provides a Python library and command-line tools.
- Canvas from Schrödinger offers MACCS, customizable SMARTS-based keys fingerprints, and seven types of hashed fingerprints, including MOLPRINT2D, ECFP, and linear (Daylight-like), as well as fingerprints derived from pharmacophore models [39,49,50]

- Molecular Operating Environment (MOE) implements 2 (TGD), 3 (TGT), and 4-point pharmacophore fingerprints in 2D/3D, MACCS keys, and EigenSpectrum shape fingerprints among others [19].
- jCompoundMapper [25,51]: This is an open-source command-line tool and a library for chemical fingerprints, featuring support for many fingerprint types, including MOLPRINT2D, atom pairs, and pharmacophore fingerprints among others. It also provides several machine learning tools and uses CDK.
- Pipeline Pilot from Accelrys [29]: This is an authoring tool with a visual and dataflow authoring language. It can calculate a wide variety of fingerprints, including both MACCS versions, ECFP, and its variants.
- SYBYL-X Suite from Tripos [32]: this is a molecular modeling suite that includes the UNITY 2D fingerprints for similarity searches.
- DecoyFinder [52,53]: DecoyFinder is a graphical tool that helps find decoy sets for virtual screening validation. It uses MACCS fingerprints and molecular descriptors to find the decoy molecules.
- FLAP [54] (Fingerprints for Ligands and Proteins): FLAP is a tool that provides a common reference framework for comparing molecules using GRID Molecular Interaction Fields (MIFs). The fingerprints are characterized by quadruplets of pharmacophoric features and can be used for ligand–ligand, ligand–receptor, and receptor–receptor comparison.
- MayaChemTools is a free collection of Perl scripts, modules and classes that support day-to-day computational discovery needs [55] The collection of scripts can compute several molecular fingerprints, including ECFP, MACCS, path-based fingerprints and many others; it can also be used directly for similarity searching with fingerprints.

2.3. Online tools for fingerprint-based virtual screening

In comparison to the large number of software packages offering fingerprint functionality, the number of online services doing so is far lower, mostly consisting of databases that include a similarity searching option using some fingerprint. A brief enumeration of the most interesting services is as follows:

- PubChem [56] provides a fast chemical structure similarity search tool. Any small molecule may be used as query, and a Tanimoto coefficient threshold can be chosen above which molecules will be deemed similar enough. The fingerprint used for this similarity searches is the PubChem fingerprint [12].
- ChemSpider [57–59] also supports similarity searching with Tanimoto (and other metrics) thresholds. It uses a fingerprint calculated by GGA's BINGO database cartridge, which uses the Indigo toolkit [49].
- The ZINC database [60–62] also supports similarity search. The fingerprint used is the path-based ChemAxon fingerprint from JChem [30,61]. It uses the same fingerprint for the generation of clusters with molecules of up to a given similarity cutoff, which produces clusters with guaranteed molecular diversity and chemical space coverage.
- The Multi-Fingerprint Browser for ZINC [63,64] is a tool that enables rapid identification of close analogs among commercially available compounds in the ZINC database [60]. The browser retrieves nearest neighbors in multi-dimensional chemical spaces defined by four different fingerprints (fingerprint = a vector composed of several numerical descriptors of molecular structure and properties), each of which represents relevant structural and pharmacophoric features in a different way: sFP (substructure fingerprint), ECFP4 (Extended connectivity fingerprint), MQN (Molecular Quantum Numbers), and SMIfp (SMILES fingerprint). Distances are calculated using the city-block distance (CBD; see Table 1), a similarity measure which, according to Awale et al. [63], performs as well as Tanimoto similarity.

3. Usual fingerprint-based virtual screening scenarios

To conduct a virtual screening based on fingerprint similarity, the following things are needed:

- At least one known active molecule, which will be the reference molecule(s).
- A molecular database with potential actives.
- Software capable of generating and comparing fingerprints.

Once the reference molecules are chosen, the next step would be to choose the most appropriate fingerprint. The choice is usually limited by the available options in the software being used. The most appropriate option would also depend mostly on the reference molecules, as a fingerprint should be able to properly represent the reference molecules (which is generally not a concern for hashed fingerprints). Whether the database and the available fingerprints account for stereochemistry, tautomeric forms, and the conformations of both the reference molecules and the molecules in the database to be screened should also be taken into account. Stereochemistry-sensitive methods should be used preferably to screen stereochemistry-sensitive databases. The presence of conformations enables the use of fingerprints that depend on them [34]. Tautomerism of the studied molecules should also be taken into account, as different tautomers of the same molecule could have substantially different fingerprints.

With the chosen algorithm, fingerprints would be calculated for every molecule and reference in the database, and then the similarity coefficient is calculated between the reference molecule and every other molecule. After this, the molecules can be ranked in descending order using the similarity coefficient. The top molecules of the rank would be expected to exhibit a similar activity as the reference molecule.

4. Comparing fingerprint similarity search with other virtual screening methods

In a comparison by Tresadern et al. [65] ECFP6 fingerprints were compared to several other virtual screening methods: feature trees, topomers, ROCS shape Tanimoto, EON electrostatic Tanimoto, OpenEye ComboScore (a combination of shape Tanimoto and color-score), and Cresset-Fieldscreen. All of these, other than those that feature trees, are 3D methods and require substantially more computation time than fingerprints. The results were as expected: the ECFP6 fingerprint was the weakest performing method with 3 out of the 4 queries, although it exhibited one of the highest performances with the remaining query. However, the 3 queries, where the fingerprint was outperformed, all showed very similar performances for all the methods, which may imply that the performance of the methods depends on the selected queries.

In a different comparison, by McGaughey et al. [66], the Daylight fingerprint was put to test against many other virtual screening methods, including protein–ligand docking. The Daylight fingerprint outperformed most of the other methods. The authors conclude that "as measured by EF, the 2D similarity methods (TOPOSIM, Daylight) perform well at lead-hopping when applied to a diverse database. [...] One may ask how it is possible for 2D similarity methods to perform nearly as well as 3D methods at lead hopping." They also noted how sensitive the performance is in Daylight fingerprints regarding path length, and that the default

settings (minimum path length of 0 and maximum of 7) is too easy to outperform making them poor standards for 2D similarity.

In yet another comparison [67], several fingerprints (Open Babel FP2, BCI, MACCS, Daylight and MOLPRINT2D) were compared against 3D molecular shape-based methods (ESHAPE3D, ROCS, PARAFIT, SHAEP and USR). Given the results, the authors state that "Overall, we find that the 2D fingerprint-based methods give better Virtual Screening performance than the 3D shape-based approaches for many of the DUD targets". This shows how 3D methods do not always outperform simple fingerprint similarity search.

However, when comparing fingerprint similarity searching to other virtual screening approaches, the use of fingerprints has several advantages:

- It requires minimal setup and configuration. Some fingerprints can be fine-tuned in several ways, but it will still require a lot less work than creating pharmacophores or selecting and preparing a binding site for a protein–ligand docking.
- Most of the commonly used fingerprints are calculated based on 2D structures. Therefore, for these, conformations do not need to be generated as opposed to shape-similarity or docking approaches. This also means that 3D information will be mostly missing from the screening, although that may not impact the performance at all [67].
- It is less CPU-intensive than other methods. This means that it can be carried out in a regular computer, and with the same hardware, it will be a lot faster than other methods, especially protein–ligand docking.

Nonetheless, fingerprint-based similarity searching also has some pitfalls that users should be aware of:

- Activity cliffs: Activity cliffs are defined as pairs of compounds with very high similarity yet highly different activity; therefore, their presence can negatively impact the performance of the similarity searching. Activity cliffs are dependent on the dataset and the descriptors used to calculate similarity, so different approaches will show different activity cliffs in the same dataset, and finding the best solution can be tricky [68].
- Choice of descriptors: Similarity search performance depends greatly on the descriptors used to calculate similarity, and in the case of fingerprints, different fingerprints can yield very different performance results [69]. The obtained results can also vary depending on the algorithm implementation.
- Reference molecules: For similarity searching, at least one known active molecule is needed for use as a reference molecule. However, it is often the case that not all parts of the reference molecules are equally relevant to overall activity. If this redundancy is not taken into account, one may obtain inactive molecules similar in irrelevant aspects to the reference molecules ranked similarly or even higher than bona fide active molecules that are only similar to the reference molecules in the activity-relevant aspects. A proper fingerprint choice based on the knowledge of the reference compounds may help alleviate this problem.
- Conformation coverage: When using 3D fingerprints, the conformations of each molecule should adequately cover its conformational space, which requires the testing and optimization of several parameters [70].

In addition, there are also many other pitfalls that are not specific to similarity searching, but common to almost all virtual screening methods, as thoroughly explained by Scior et al. [70].

5. Conclusion

There are many types of fingerprints, and thus there is also interest in knowing which fingerprints perform better. There are open-source platforms to benchmark fingerprints for ligand-based virtual screening that have been tested with 14 2D fingerprints [28]. Studies have found that the overall performance of all the fingerprints was similar, though, the inter-target difference in performance was greater than the intra-target difference between fingerprints. After ranking the fingerprints by performance, these studies found that ECFP0 (with a diameter of 0 when only taking the single atom as the environment) and 166-bit MACCS were the worst when using early recognition evaluation methods. Using the same methods, circular fingerprints were ranked higher, and the topological torsions fingerprint was always highly ranked regardless of the evaluation methods.

The current trend regarding similarity searching with molecular fingerprints seems to be to combine different approaches through data fusion [71] (either by combining different fingerprints [63,72,73] or by combining fingerprints with other virtual screening methods [73,74], specially structure-based methods [75]). The advantage of this approach is that, by combining methods that capture different chemical information, the highest-ranked hits will be those that are highly ranked by several approaches, making them more relevant and reducing the amount of artifacts that would be introduced by a single approach. This could possibly lead to the optimal search and combination of methods in data fusion, with increased virtual screening performance.

Acknowledgements

This manuscript was edited for English language fluency by American Journal Experts. This study was supported by grant AGL2011-25831/ALI from the Spanish Government and ACC1Ó program and XRQTC Grant from 'Generalitat de Catalunya'.

References

[1] U. Rester, Curr. Opin. Drug Discov. Devel. 11 (7) (2008) 559–568.
[2] H. Sun, Curr. Med. Chem. 15 (1) (2008) 1018–1024.
[3] J. Kirchmair, S. Distinto, P. Markt, D. Schuster, G.M. Spitzer, K.R. Liedl, et al., J. Chem. Inf. Model. 49 (3) (2009) 678–692, http://dx.doi.org/10.1021/ci8004226.
[4] J.L. Melville, E.K. Burke, J.D. Hirst, Comb. Chem. High Throughput Screen. 12 (5) (2009) 332–343.
[5] M. Dunkel, U. Schmidt, S. Struck, L. Berger, B. Gruening, J. Hossbach, et al., Nucleic Acids Res. 37 (1) (2009) D291–D294, http://dx.doi.org/10.1093/nar/gkn695.
[6] G.M. Maggiora, V. Shanmugasundaram, Methods Mol. Biol. 672 (1) (2011) 39–100, http://dx.doi.org/10.1007/978-1-60761-839-3_2.
[7] G.M. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, J. Med. Chem. (10) (2013), http://dx.doi.org/10.1021/jm401411z.
[8] H. Eckert, J. Bajorath, Drug Discov. Today 12 (3) (2007) 225–233, http://dx.doi.org/10.1016/j.drudis.2007.01.011.
[9] N. Salim, J. Holliday, P. Willett, J. Chem. Inf. Comput. Sci. 43 (1) (2002) 435–442, http://dx.doi.org/10.1021/ci025596j.
[10] Accelrys, MACCS structural keys, (n.d.).
[11] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, J. Chem. Inf. Model. 42 (11) (2002) 1273–1280, http://dx.doi.org/10.1021/ci010132r.
[12] E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, Annu. Rep. Comput. Chem. 4 (2008) 217–241, http://dx.doi.org/10.1016/S1574-1400(08)00012-1.
[13] A. Dalke, ChemFP. <http://chemfp.com> (accessed on 08/04/2014).
[14] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, J. Chem. Inf. Comput. Sci. 43 (1) (2003) 493–500, http://dx.doi.org/10.1021/ci025584y.
[15] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, E.L. Willighagen, Curr. Pharm. Des. 12 (1) (2006) 2111–2120.
[16] J.M. Barnard, G.M. Downs, J. Chem. Inf. Model. 37 (1) (1997) 141–142, http://dx.doi.org/10.1021/ci960090k.
[17] A. Tovar, H. Eckert, J. Bajorath, ChemMedChem 2 (2007) 208–217, http://dx.doi.org/10.1002/cmdc.200600225.

[18] R.P. Sheridan, M.D. Miller, D.J. Underwood, S.K. Kearsley, J. Chem. Inf. Model. 36 (1) (1996) 128–136, http://dx.doi.org/10.1021/ci950275b.
[19] Chemical Computing Group Inc., Molecular operating environment (MOE), (2013).
[20] I. Daylight chemical information systems, Daylight. <http://www.daylight.com/> (accessed on 08/04/2014).
[21] OpenEye scientific software, OEChem, (2013).
[22] A. Bender, H.Y. Mussa, R.C. Glen, S. Reiling, J. Chem. Inf. Comput. Sci. 44 (1) (2003) 170–178, http://dx.doi.org/10.1021/ci034207y.
[23] A. Bender, H.Y. Mussa, R.C. Glen, S. Reiling, J. Chem. Inf. Comput. Sci. 44 (1) (2004) 1708–1718, http://dx.doi.org/10.1021/ci0498719.
[24] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, J. Cheminform. 3 (1) (2011) 33, http://dx.doi.org/10.1186/1758-2946-3-33.
[25] G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner, A. Zell, J. Cheminform. 3 (1) (2011) 3, http://dx.doi.org/10.1186/1758-2946-3-3.
[26] H.L. Morgan, J. Chem. Doc. 5 (5) (1965) 107–113, http://dx.doi.org/10.1021/c160017a018.
[27] D. Rogers, M. Hahn, J. Chem. Inf. Model. 50 (5) (2010) 742–754, http://dx.doi.org/10.1021/ci100050t.
[28] S. Riniker, G.A. Landrum, J. Cheminform. 5 (1) (2013) 26, http://dx.doi.org/10.1186/1758-2946-5-26.
[29] Accelrys, Accelrys – scientific enterprise software for chemical research, material science R&D. <http://accelrys.com/> (accessed on 08/04/2014).
[30] ChemAxon – cheminformatics platforms and desktop applications. <https://www.chemaxon.com/> (accessed on 08/04/2014).
[31] G. Landrum, RDKit: open-source cheminformatics. <http://www.rdkit.org> (accessed on 08/04/2014).
[32] Tripos: a Certara™ company. <http://www.tripos.com/> (accessed on 08/04/2014).
[33] L. Xue, J.W. Godden, F.L. Stahura, J. Bajorath, J. Chem. Inf. Comput. Sci. 43 (1) (2003) 1151–1157, http://dx.doi.org/10.1021/ci030285+.
[34] M.J. McGregor, S.M. Muskal, J. Chem. Inf. Comput. Sci. 40 (1999) 117–125.
[35] D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. 45 (1) (2005) 386–393, http://dx.doi.org/10.1021/ci0496797.
[36] J. Schwartz, M. Awale, J.-L. Reymond, J. Chem. Inf. Model. 53 (8) (2013) 1979–1989, http://dx.doi.org/10.1021/ci400206h.
[37] D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Model. 29 (5) (1989) 97–101, http://dx.doi.org/10.1021/ci00062a008.
[38] Z. Deng, C. Chuaqui, J. Singh, J. Med. Chem. 47 (1) (2004) 337–344, http://dx.doi.org/10.1021/jm030331x.
[39] M. Sastry, J.F. Lowrie, S.L. Dixon, W. Sherman, J. Chem. Inf. Model. 50 (5) (2010) 771–784, http://dx.doi.org/10.1021/ci100062n.
[40] N.M. O'Boyle, G.R. Hutchison, Chem. Cent. J. 2 (2008) 24.
[41] Open Babel. <http://openbabel.org> (accessed on 08/04/2014).
[42] R.E. Carhart, D.H. Smith, R. Venkataraghavan, J. Chem. Inf. Model. 25 (5) (1985) 64–73, http://dx.doi.org/10.1021/ci00046a002.
[43] R. Nilakantan, N. Bauman, J.S. Dixon, R. Venkataraghavan, J. Chem. Inf. Model. 27 (5) (1987) 82–85, http://dx.doi.org/10.1021/ci00054a008.
[44] The chemistry development kit. <http://sourceforge.net/projects/cdk/> (accessed on 08/04/2014).
[45] L.H. Hall, L.B. Kier, J. Chem. Inf. Model. 35 (11) (1995) 1039–1045, http://dx.doi.org/10.1021/ci00028a014.
[46] J. Klekota, F.P. Roth, Bioinformatics 24 (12) (2008) 2518–2525, http://dx.doi.org/10.1093/bioinformatics/btn479.
[47] Indigo – GGA software services. <http://ggasoftware.com/opensource/indigo> (accessed on 08/04/2014).
[48] Cinfony – A common API for several cheminformatics toolkits – google project hosting. <https://code.google.com/p/cinfony/> (accessed on 08/04/2014).
[49] J. Kiener, J. Cheminform. 5 (1) (2013) 48, http://dx.doi.org/10.1186/1758-2946-5-48.
[50] Canvas – Product features. <http://www.schrodinger.com/Canvas/> (accessed on 08/04/2014).
[51] jCompoundMapper. <http://jcompoundmapper.sourceforge.net/> (accessed on 08/04/2014).
[52] A. Cereto-Massagué, L. Guasch, C. Valls, M. Mulero, G. Pujadas, S. Garcia-Vallve, Bioinformatics 4 (2012) 2–3, http://dx.doi.org/10.1093/bioinformatics/bts249.
[53] DecoyFinder. <http://urvnutrigenomica-ctns.github.io/DecoyFinder/> (accessed on 08/04/2014).
[54] FLAP – (Fingerprints for Ligands and Proteins). <http://www.moldiscovery.com/soft_flap.php> (accessed on 08/04/2014).
[55] M. Sud, MayaChemTools: Home. <http://www.mayachemtools.org/> (accessed on 08/04/2014).
[56] The PubChem Project. <http://pubchem.ncbi.nlm.nih.gov/> (accessed on 08/04/2014).
[57] A.J. Williams, Curr. Opin. Drug Discov. Devel. 11 (5) (2008) 393–404.
[58] H.E. Pence, A. Williams, J. Chem. Educ. 87 (11) (2010) 1123–1124, http://dx.doi.org/10.1021/ed100697w.
[59] ChemSpider | Search and share chemistry. <http://www.chemspider.com/> (accessed on 08/04/2014).
[60] J.J. Irwin, B.K. Shoichet, J. Chem. Inf. Model. 45 (n.d.) 177–82. doi:10.1021/ci049714+.
[61] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, J. Chem. Inf. Model. 52 (7) (2012) 1757–1768, http://dx.doi.org/10.1021/ci3001277.
[62] Welcome to ZINC Is Not Commercial – a database of commercially-available compounds. <https://zinc.docking.org/> (accessed on 08/04/2014).
[63] M. Awale, J.-L. Reymond, Nucleic Acids Res. 4 (2014) gku379, http://dx.doi.org/10.1093/nar/gku379.
[64] ZINC Browser. <http://dcb-reymond23.unibe.ch:8080/MCSS/> (accessed on 08/04/2014).
[65] G. Tresadern, D. Bemporad, T. Howe, J. Mol. Graph. Model. 27 (1) (2009) 860–870, http://dx.doi.org/10.1016/j.jmgm.2009.01.003.
[66] G.B. McGaughey, R.P. Sheridan, C.I. Bayly, J.C. Culberson, C. Kreatsoulas, S. Lindsley, et al., J. Chem. Inf. Model. 47 (1) (2007) 1504–1519, http://dx.doi.org/10.1021/ci700052x.
[67] V. Venkatraman, V.I. Pérez-Nueno, L. Mavridis, D.W. Ritchie, J. Chem. Inf. Model. 50 (12) (2010) 2079–2093, http://dx.doi.org/10.1021/ci100263p.
[68] M. Cruz-Monteagudo, J.L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M.N.D.S. Cordeiro, F. Borges, Drug Discov. Today 2 (2014), http://dx.doi.org/10.1016/j.drudis.2014.02.003.
[69] A. Bender, J.L. Jenkins, J. Scheiber, S.C.K. Sukuru, M. Glick, J.W. Davies, J. Chem. Inf. Model. 49 (1) (2009) 108–119, http://dx.doi.org/10.1021/ci800249s.
[70] T. Scior, A. Bender, G. Tresadern, J.L. Medina-Franco, K. Martínez-Mayorga, T. Langer, et al., J. Chem. Inf. Model. 52 (4) (2012) 867–881, http://dx.doi.org/10.1021/ci200528d.
[71] P. Willett, Biotechnol. J. 5 (1) (2013) e201302002, http://dx.doi.org/10.5936/csbj.201302002.
[72] A. Ahmed, F. Saeed, N. Salim, A. Abdo, J. Cheminform. 6 (1) (2014) 19, http://dx.doi.org/10.1186/1758-2946-6-19.
[73] P. Willett, J. Chem. Inf. Model. 53 (1) (2013) 1–10, http://dx.doi.org/10.1021/ci300547g.
[74] G.M. Sastry, V.S.S. Inakollu, W. Sherman, J. Chem. Inf. Model. 53 (7) (2013) 1531–1542, http://dx.doi.org/10.1021/ci300463g.
[75] F. Broccatelli, N. Brown, J. Chem. Inf. Model. 5 (2014), http://dx.doi.org/10.1021/ci5001604. 140530101617007.