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].

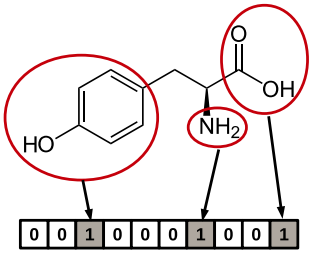

Fig. 1. A representation of a hypothetical 10-bit substructure fingerprint, with three bits set because the substructures they represent are present in the molecule (circled).

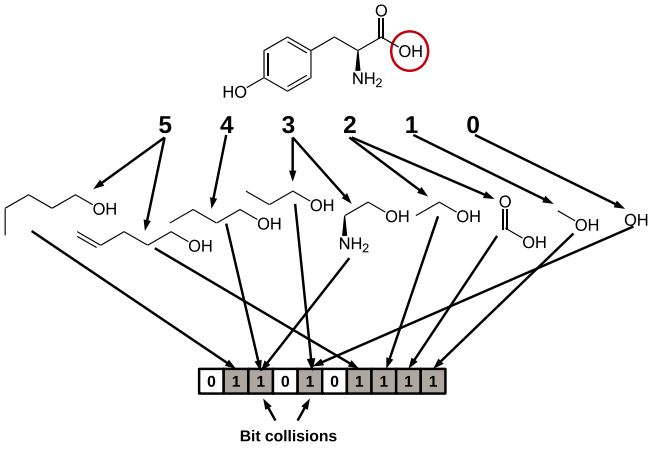

Fig. 2. A representation of a hypothetical 10-bit topological fingerprint, in this case a linear path-based fingerprint with fragments up to a length of 5. All fragments found from the starting atom (circled) are shown, and the fragment length and corresponding bit in the fingerprint are indicated. There 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.