6.1: Molecular Descriptors

Molecular Similarity

Molecular similarity [1-3] is one of the most heavily exploited concepts in cheminformatics and related areas (such as medicinal chemistry and drug discovery). It is applied to multiple tasks, including similarity searching [1], property prediction [4], synthesis design [5], virtual screening [2,3,6], cluster analysis [7,8], and molecular diversity analysis [9-11]. However, because molecular similarity is a concept, not a physical observable, "measuring" molecular similarity is inherently subjective and context-dependent. There is no correct or authoritative measure of molecular similarity. As a result, various similarity measures have been proposed to quantify the degree of structural similarity between molecules. In general, these measures involve two principal components [12]:

- **Molecular descriptors** that represent the structures of the molecules being compared.
- **Similarity coefficient** (metric) used to compute a quantitative score for the degree of similarity based on the weighted values of structural descriptors.

The molecular descriptors may need to be pre-processed before the similarity calculation, using a weighting scheme that assigns differing degrees of importance to various components of molecular descriptors. For this reason, some papers list the weighting scheme as a third component of similarity measures [1,13]. While some studies [14,15] have focused on the effects of the weighting schemes upon similarity calculations, much more attention has been given to molecular descriptors and similarity coefficients. Therefore, this chapter also focuses on these two components.

Molecular descriptors

There are many molecular descriptors that capture different aspects of molecules, but they are broadly classified according to their "dimensionality" [16]. One-dimensional (1-D) descriptors include bulk properties and physicochemical parameters (e.g., log P, molecular weight, polar surface area). Two-dimensional (2-D) descriptors include structural fragments or connectivity indices derived from the 2-D representation of the molecule. Three-dimensional (3-D) descriptors, such as molecular shape, are derived from 3-D molecular structures (i.e., 3-D coordinates of the atoms in the molecule). In this chapter, we focus on 2-D molecular fingerprints, which encodes the 2-D structure of molecules. While many molecular fingerprints have been developed, we discuss two types of molecular fingerprints, structural keys and hashed fingerprints, because they are more widely used than others.

Structural keys

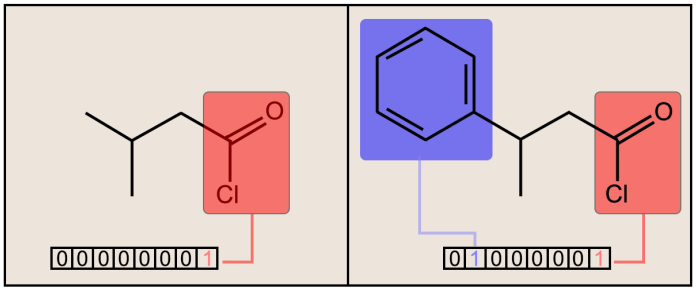

Fig. 1. (above) Two molecules are shown along with the respective bit substructures highlighted for comparison. The number of bits and designations used for this figure is simply for display and illustrative purposes. The true fingerprint would be much longer.

In structural keys, the structure of a molecule is encoded into a binary bit string (that is, a sequence of 0's and 1's), each bit of which corresponds to a "pre-defined" structural feature (e.g., substructure or fragment). If the molecule has a pre-defined feature, the bit position corresponding to this feature is set to 1 (ON). Otherwise, it is set to 0 (OFF). It is important to understand that structural keys cannot encode structural features that are not pre-defined in the fragment library. Examples are the MACCS keys [17,18] and PubChem Fingerprints [19].

- #### MACCS keys

  The MACCS (Molecular ACCess System) keys [17,18] are one of the most commonly used structural keys. They are sometimes referred to as the MDL keys, named after the company that developed them [the MDL Information Systems (now BIOVIA)]. While there are two sets of MACCS keys (one with 960 keys and the other containing a subset of 166 keys), only the shorter fragment definitions are available to the public. These 166 public keys are implemented in popular open-source

cheminformatics software packages, including RDKit [20], OpenBabel [21,22], CDK [23,24], etc. The fragment definitions for the MACCS 166 keys can be found in this document:

https://github.com/rdkit/rdkit/blob/master/rdkit/Chem/MACCSkeys.py

- PubChem fingerprints

  The PubChem fingerprint [19] is a 881-bit-long structural key, which is used by PubChem for similarity searching (interactively through the PubChem Homepage or programmatically through PUG-REST). It is also used for structure neighboring, which "pre-computes" a list of similar chemical structure for each compound. This pre-computed list is accessible through the Compound Summary page (the Related Compounds and Related Compounds with Annotation sections). The fragment dictionary of the PubChem fingerprint is organized in seven sections, as described in the following document:

  ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf

Hashed Fingerprints

An alternative to structural keys is hashed fingerprints. Contrary to structural keys, hashed fingerprints do not require a pre-defined fragment library. Instead, they are generated by enumerating through the molecule all possible fragments that are not bigger than a certain size and then converting these fragments into numeric values using a "hash" function (https://en.Wikipedia.org/wiki/Hash_function). These numeric values can be used to indicate bit positions in the hashed fingerprints.

Hash functions are used to map data of arbitrary size to "fixed-size" values. Enumerating all possible fragments with a molecule may result in a very large number of fragments. Hashing them into values within a fixed range inevitably results in "bit collisions", in which different fragments are converted into the same numeric value (and the same bit position). Because of this, there is no one-to-one correspondence between fragments and fingerprint bits (contrary to structural keys).

Hashed fingerprints may be further classified into topological or path-based fingerprints and circular fingerprints, according to the way by which the fragments are enumerated.

- **Path-based fingerprints**

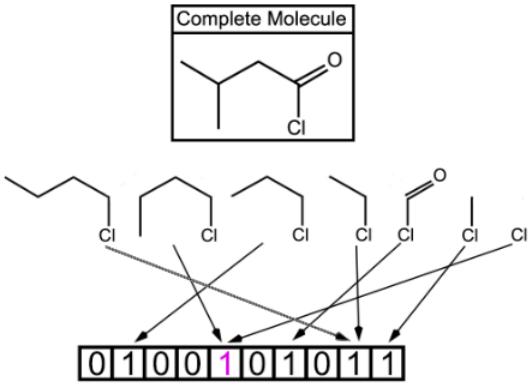

Fig. 2. Shown above is a topological fingerprint with multiple collisions between fragments. A bit collision is represented by having two or more arrows from the molecular fragments pointing to the same bit value. Starting with the chlorine atom, all of the possible fragments are shown. However in a true fingerprint, each atom could be the starting point which would allow for many more fragments than this example shows. The more bits allowed, the less likely for the bit collisions, which is represented by having two collisions due to only 10 bits being used.

  In this type of fingerprints, fragments of the molecule are generated by following a (usually linear) path up to a certain number of bonds within the molecule. The most well-known example of path-based fingerprints is the Daylight fingerprint [25,26].

-

Circular fingerprints

Circular fingerprints are generated by considering the "circular" environment of each atom up to a given "radius" or "diameter". Examples of circular fingerprints are extended-connectivity fingerprints (ECFPs) [27]. ECFPs are generated using a variant of the Morgan algorithm [28], which is a method for solving the molecular isomorphism problem (i.e., how to identify identical molecules that have different atom numberings). Different flavors of ECFPs may be generated by selecting different maximum diameter of the circular atom neighborhood and they are referred to as ECFP2, ECFP4, ECFP6, etc., where the digit at the end indicates the maximum diameter value employed to generate the fingerprint. The most commonly used ones are ECFP4 and ECFP6.

Another example of circular fingerprints is functional-class fingerprints (FCFPs) [27], which are a variation of ECFPs. FCFPs are further abstracted in that FCFPs encodes atom's roles (not atoms). At the initial stage of FCFP generation, each atom in the molecule is assigned a special code that represents one of the atom roles (e.g., hydrogen-bond acceptor and donor, negatively or positively ionizable, aromatic, and halogen), and these codes (not the atoms) are used to generate FCFPs, through the same process as ECFPs.

References

1. Willett P, Barnard JM, Downs GM: **Chemical similarity searching.** *J Chem Inf Comput Sci* 1998, **38:**983-996.
2. Cereto-Massague A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallve S, Pujadas G: **Molecular fingerprint similarity search in virtual screening.** *Methods* 2015, **71:**58-63.
3. Muegge I, Mukherjee P: **An overview of molecular fingerprint similarity search in virtual screening.** *Expert Opin Drug Discov* 2016, **11:**137-148.
4. Brown RD, Martin YC: **Use of structure Activity data to compare structure-based clustering methods and descriptors for use in compound selection.** *J Chem Inf Comput Sci* 1996, **36:**572-584.
5. Wipke WT, Rogers D: **ARTIFICIAL-INTELLIGENCE IN ORGANIC-SYNTHESIS - SST - STARTING MATERIAL SELECTION-STRATEGIES - AN APPLICATION OF SUPERSTRUCTURE SEARCH.** *J Chem Inf Comput Sci* 1984, **24:**71-81.
6. Eckert H, Bojorath J: **Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches.** *Drug Discov Today* 2007, **12:**225-233.
7. Cruz R, Lopez N, Quintero M, Rojas G: **Cluster analysis from molecular similarity matrices using a non-linear neural network.** *J Math Chem* 1996, **20:**385-394.
8. Pan DH, Iyer M, Liu JZ, Li Y, Hopfinger AJ: **Constructing optimum blood brain barrier QSAR models using a combination of 4D-molecular similarity measures and cluster analysis.** *J Chem Inf Comput Sci* 2004, **44:**2083-2098.
9. Golbraikh A: **Molecular dataset diversity indices and their applications to comparison of chemical databases and QSAR analysis.** *J Chem Inf Comput Sci* 2000, **40:**414-425.
10. Klein CT, Kaiser D, Ecker G: **Topological distance based 3D descriptors for use in QSAR and diversity analysis.** *J Chem Inf Comput Sci* 2004, **44:**200-209.
11. Koutsoukas A, Paricharak S, Galloway W, Spring DR, Ijzerman AP, Glen RC, Marcus D, Bender A: **How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space.** *J Chem Inf Model* 2014, **54:**230-242.
12. Holliday JD, Hu CY, Willett P: **Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings.** *Comb Chem High Throughput Screen* 2002, **5:**155-166.
13. Chen X, Reynolds CH: **Performance of similarity measures in 2D fragment-based similarity searching: Comparison of structural descriptors and similarity coefficients.** *J Chem Inf Comput Sci* 2002, **42:**1407-1414.
14. Bath PA, Morris CA, Willett P: **EFFECT OF STANDARDIZATION ON FRAGMENT-BASED MEASURES OF STRUCTURAL SIMILARITY.** *J Chemometr* 1993, **7:**543-550.
15. Turner DB, Willett P, Ferguson AM, Heritage TW: **Similarity Searching in Files of Three-Dimensional Structures: Evaluation of Similarity Coefficients and Standardisation Methods for Field-Based Similarity Searching.** *SAR and QSAR in Environmental Research* 1995, **3:**101-130.
16. Xue L, Bajorath J: **Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening.** *Comb Chem High Throughput Screen* 2000, **3:**363-372.
17. Durant JL, Leland BA, Henry DR, Nourse JG: **Reoptimization of MDL keys for use in drug discovery.** *J Chem Inf Comput Sci* 2002, **42:**1273-1280.

18. **THE KEYS TO UNDERSTANDING MDL KEYSET TECHNOLOGY.** https://www.3dsbiovia.com/products/pdf/keys-to-keyset-technology.pdf. Accessed Oct. 2019.
19. **PubChem Substructure Fingerprint.** ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf. Accessed Oct. 2019.
20. **RDKit.** https://www.rdkit.org/. Accessed October 2019.
21. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: An open chemical toolbox.** *J Cheminformatics* 2011, **3:**33.
22. **The Open Babel Package.** https://openbabel.org. Accessed October, 2019.
23. **Chemistry Development Kit (CDK).** https://cdk.github.io/. Accessed October 2019.
24. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Cherto M, Spjuth O, Torrance G, Evelo CT, Guha R, Steinbeck C: **The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching.** *J Cheminformatics* 2017, **9:**33.
25. **Daylight Theory: Fingerprints.** https://www.daylight.com/dayhtml/doc/theory/theory.finger.html. Accessed October 2019.
26. **Daylight Fingerprints.** https://www.daylight.com/meetings/summerschool01/course/basics/fp.html. Accessed October 2019.
27. Rogers D, Hahn M: **Extended-Connectivity Fingerprints.** *J Chem Inf Model* 2010, **50:**742-754.
28. Morgan HL: **GENERATION OF A UNIQUE MACHINE DESCRIPTION FOR CHEMICAL STRUCTURES-A TECHNIQUE DEVELOPED AT CHEMICAL ABSTRACTS SERVICE.** *Journal of Chemical Documentation* 1965, **5:**107-&.

Tags recommended by the template: article:topic
