Section III

Cheminformatics - Basics: *Molecular Descriptors and Fingerprints*

Dan C. Fara and Tudor I. Oprea

1. Molecular descriptors

The molecule – thought as a real object – implicitly contains all the chemical information, but only a part of this information can be extracted by experimental measurements. In the last decades, several scientific researches have been focussed on studying how to catch and convert – by a theoretical pathway - the information encoded in the molecular structure into one or more numbers – called molecular descriptors – used to establish quantitative relationships between structures and properties, biological activities and other experimental properties. Nowadays, many contemporary applications in computer-aided drug discovery and cheminformatics rely on such descriptors, which capture the structural characteristics and properties of the chemicals. Literally hundreds of molecular descriptors (~ 2000) that encode the molecular features in very different ways have been reported over the years.

The majority of molecular descriptors can be classified according to their "dimensionality", which refers to the representation of molecules from which descriptor values are computed. According to this scheme, one-dimensional (1D) descriptors capture bulk properties, i.e. molecular weight, molar refractivity, logP (logarithm of the octanol/water partition coefficient) etc., two-dimensional (2D) descriptors describe properties that can be computed from two-dimensional representations of molecules, i.e. number of atoms, number of bonds, connectivity indices[1] etc., and three-dimensional (3D) descriptors depend on conformations of molecules, i.e. solvent accessible surface areas, principal moment of inertia, Van der Waals volume etc.

These categories are heterogeneous and the descriptors in these categories are not always clearly distinguished. For example, some "implicit" 3D descriptors representing surface area terms or shape indices can be calculated from 2D representations of molecules or molecular graphs[2]. Descriptors calculated from 3D structures may require the analysis of many molecular conformations, provided that biologically active conformations are not known from experiment. Popular 3D descriptors include pharmacophore-type representations of molecules where features (e.g., hydrophobic "centers" or hydrogen bond donors) known or thought to be responsible for biological activity are mapped to positions in a molecule. Then conformation-dependent distances between these points are calculated and recorded. Three-point pharmacophores are widely used[3], but more powerful four-point pharmacophores have been introduced[4], which may require the analysis of milions of possible pharmacophores for a test compound. Such complex 3D descriptors are calculated, for example, to identify active conformation(s) of a compound or identify features critical for differences in activity in series of analogs. In the same time, this type of calculations is needed to generate the "pharmacophore shape" of a query molecule in order to search databases for compounds with similar 3D characteristics. In addition, the use of pharmacophore-type descriptors is critical for the derivation of 3D-QSAR or 4D-QSAR models[5,6].

The molecular descriptors can also be classified according to their "nature" into the following five classes: (i) constitutional, (ii) topological, (iii) geometrical, (iv) electronic, and (v) quantum-chemical, as shown in Figure 1. The constitutional descriptors are fragment additive and reflect mostly the general properties of the

compound. The topological descriptors are calculated using the mathematical graph theory applied to the scheme of atoms connections of the structure. The geometrical, electronic and quantum-chemical descriptors are usually derived from the results of empirical schemes or molecular orbital calculations and they encode the molecule's ability to participate in polar interactions or hydrogen bonding (donor, acceptor).

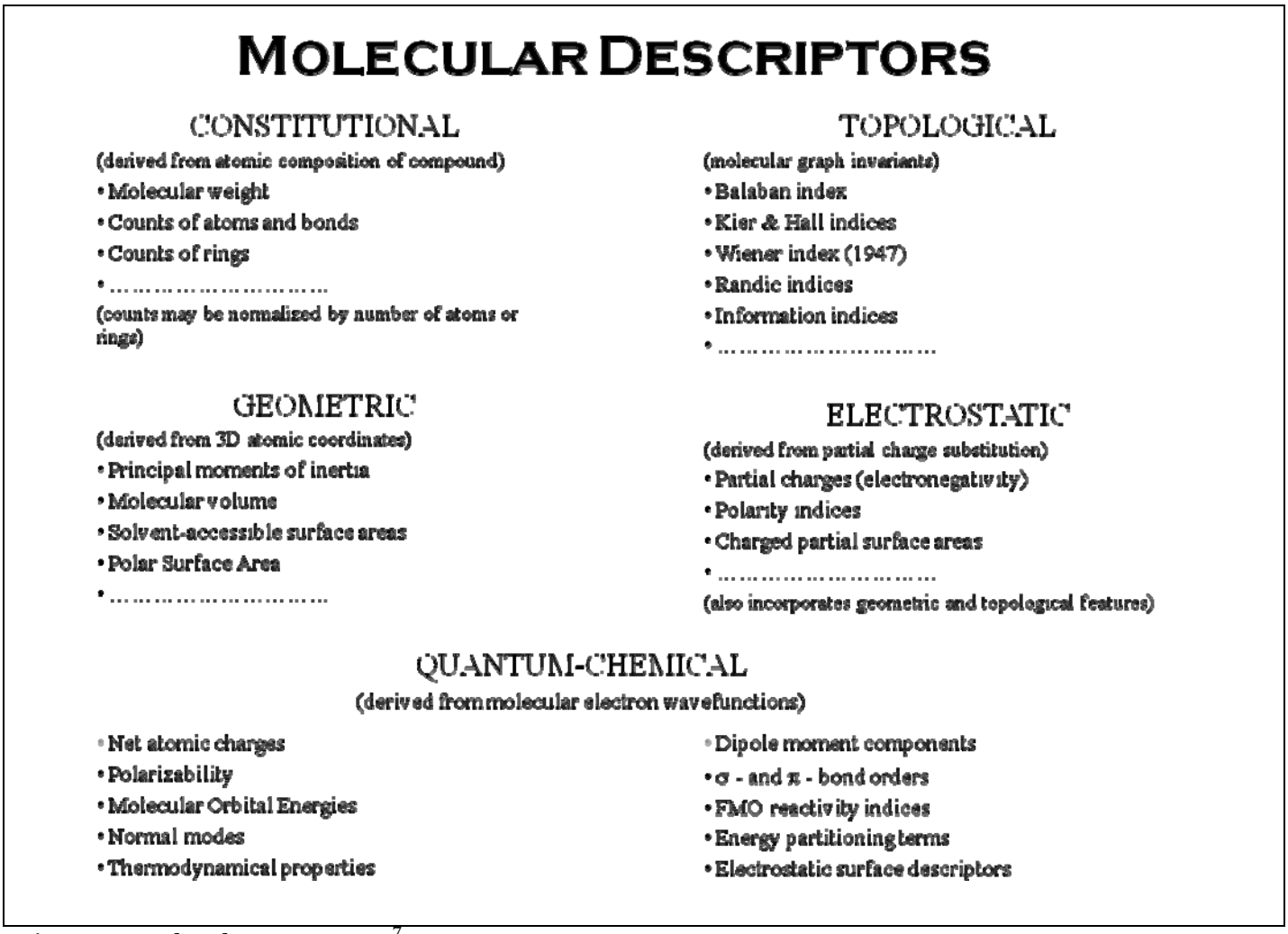

Figure 1. Molecular Descriptors[7]

There are many programs available on the market for calculating molecular descriptors, i.e. CODESSA, ALMOND package, JOELib and so on (see Properties / Parameters at http://www.ndsu.edu/qsar_soc/resource/software.htm). Moreover, some sites provide free on-line tools for performing computational chemistry, ADME/T and chemoinformatics tasks including the building and visualisation of chemical structures, the calculation of molecular properties and the analysis of relationships between chemical structure and properties, i.e. Virtual Computational Chemistry Laboratory.

- Go to Virtual Computational Chemistry Laboratory
- Start the program ALOGPS
- Calculate the AlogP and AlogS for beta-estradiol, 4-hydroxytamoxifen, and dihydrotestosterone

Quiz 1. Which one is more lipophilic? Which one is more soluble?

2. Fingerprints

Fingerprint representations of molecular structure and properties are a particulary complex form of descriptors. Fingerprints are typically encoded as binary bit strings whose settings produce, in different ways, a bit "pattern" characteristic of a given molecule. Fingerprints are designed to account for different sets of molecular descriptors[8], structural fragments[9], possible connectivity pathways through a molecule[10], or different types of pharmacophores[11,12].

If numerical descriptors are encoded in a binary format, their bit settings must account for specific value ranges. Thus, for each descriptor, meaningful value ranges should be determined prior to design of fingerprints, for example, by calculating descriptor values for large compound collections and surveying their distribution in histograms[8]. In this way, appropriate bit settings can be defined that cover the distributions of numerical descriptors in large compound libraries. By contrast, binary encoding of structural fragment-type descriptors is straightforward, since only their presence (1) or absence (0) must be detected.

In other words, the fingerprint (the fixed-length bit-string) contains a fixed number of bits in which each bit can represent the absence (0) or presence (1) of some feature, either on its own or in conjunction with other bits in the bit-string. **Discrete** variables, with more than two values, can be represented in the binary bit-string by using a bit for each possible value or for given ranges of values. **Continuous** variable can be represented by defining ranges of values and then assigning a bit to each range, a process known as *binning*. The ranges covered by each bin can be (i) separate or be overlapping, (ii) equidistant, (iii) equifrequent, or (iv) user-/application-defined. The fingerprints can be (i) directly, (ii) dictionary, or (iii) hash assigned. Descriptors with fixed limits can be *directly* assigned to positions in a fingerprint, with offsets being calculated to assign different groups or different descriptors to separate areas of the fingerprint. Systems based on *dictionary*-assigned bit-strings employ a dictionary that specifies correspondence between particular functional groups or fragments and bit positions in a fingerprint, with each entry (structural key) in the dictionary being assigned a bit position (screen number). Dictionaries of functional groups tend to be fairly small, so all groups can be listed in the dictionary and assigned to a short bit-string. Rather than selecting a subset of fragments for inclusion in a dictionary, so that the number of screens is reduced to the same as the length of the fingerprint, *hash*-assigned bit-strings are created by fitting all of the fragments into the bit-string. This can be achieved by hashing the fragment to generate one or more integers that fall within the length, or a given subrange of the length, of the fingerprint (bit-string). The more integers generated by the hash function, the more unique patterns can be superimposed on the bit-string, so the more fragments can be included.

Fingerprint representations are often used to search for molecules "similar" to a query compound[13]. To do so, a fingerprint is calculated for a query and searched against corresponding fingerprints of compounds in the database. For pairwise comparison of compounds, fingerprint overlap is determined as a measure of similarity and calculated using various coefficients, i.e. Tanimoto[13] - see details on *Molecular Similarity (or Diversity)*.

Several methodologies exist for chemical binary representations. For example, Daylight Chemical Information Systems fingerprint is often referred to as a *path-based* approach. This amounts to a unique subgraph matching of the graph representation of the chemical structure. In the Daylight algorithm the fingerprint is "learned" from the structures themselves. A molecular fingerprint is generated from a hash of all the unique connection paths (subgraphs) up to a maximum size (typically 8) into a fixed length bit string. Fingerprints may be folded to decrease the length and increase the bit density. Typical sizes for Daylight fingerprints are 512 or 1024 bits in length, but any power of two can be generated[10].

Molecular Design Limited (MDL), created a key based fingerprint. This fingerprint uses a pre-defined set of definitions and creates fingerprints based on pattern matching of the structure to the defined "key" set. This key based approach relies on the definitions to encapsulate the molecluar descriptions *apriori* and does not "learn" the keys from the chemical dataset. The MDL original public key set was 166 keys, and their private key set was comprised of 966 keys. Their publication of "drug -like" keys contains a subset of 320 keys from their 966 set[14]. So, MDL fingerprints could take on a maximum bit length of 966. No folding occurs with this type of fingerprint.

Barnard Chemical Information Systems (BCI) uses a dictionary approach in which the keys for the fingerprinter are first generated from the set and then implemented in the description. This combines a bit of both of the Daylight and MDL approaches. Typically the BCI dictionary generates thousands of keys, resulting in molecular fingerprint bit lengths on the order of 5,000 bits.

Mesa Analytics & Computing, LLC uses the 320 "drug-like" published by MDL to generate 320 bit string representations as well as the 166 bit string representations based on MDL's original public dataset. The keys are generated from SMARTS pattern matching against the chemical dataset using the SMARTS matching algorithm in OEChem from OpenEye Scientific Software.

Table 1 lists the fingerprints of dihydrotestosterone, β-estradiol, and 4-hydroxytamoxifen generated using **gen_maccs320** (former **MACCSKeys320Generator**) program of the **Fingerprint Module** in the **Mesa Suite**, a product of Mesa Analytics & Computing, LLC.

Table 1. SMILES and FINGERPRINTS of dihydrotestosterone, β-estradiol, and 4-hydroxytamoxifen

NAME	SMILES	FINGERPRINTS (MDL MACCS 320 bit)
dihydrotestosterone	C1CC(=O)CC2C1(C3C(CC2)C4C(CC3)(C(CC4)O)C)C	00111110111000000001000000000000000000010 01111111100000000000000000000000000000000 00000000000001011110000111100000000100111 11111111000010000111100000000000001110000 00011110000000000000000000001000000000000 00000000000000000000000000000000000000000 00000000000000000000000000000000001001110 00001000100000010001000010000000
β-estradiol	c1cc(cc2c1C3C(CC2)C4C(CC3)(C(CC4)O)C)O	00111110111000000001000000000000000000000 00000000000000000000000000000000000000000 00000000000001011110000111100000000110111 11101111000010000111100000000000001111001 11011110000000000000000000001101011111111 10000000000000000000000000000000000000000 00100000000000000000000000000000001001110 00001000100000000000000100000000
4-hydroxytamoxifen	CCC(=C(c1ccc(cc1)O)c2ccc(cc2)OCCN(C)C)c3ccccc3	01100000000011111001001000000000000000000 00001000000000000000000000000000000000000 00000000000000000010000111100000000100110 01111100110110000000000011000000000000000 00000000101000000010100000001000011111111 00000111100100000000001100000000010100000 10100000000000000001000000000000000001110 00011001100000000000000000000000

- Go to Mesa Analytics & Computing, LLC, *Home*
- Enter the Cheminformatics Virtual Classroom
- Click on Fingerprint Viewer and watch the short video

[1] Balaban, A.T. *Chem. Phys. Lett.*, **1982**, *89*, 399-404.
[2] Kier, L.B. *Med. Res. Rev.*, **1987**, *7*, 417-440.
[3] Pickett, S.D.; Mason, J.S.; McLay, I.M. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 1214-1223.
[4] Mason, J.S.; Morize, I.; Menard, P.R.; Cheney, D.L.; Hulme, C.; Labaudiniere, R.F *J. Med. Chem.*, **1999**, *42*, 3251-3264.
[5] Hopfinger, A.T.; Wang, S.; Tobarski, J.S.; Jin, B.; Albuquerque, M.; Madhav, P.J.; Duraiswami, C. *J. Am. Chem. Soc.*, **1997**, *119*, 10509-10524.
[6] Hopfinger, A.J.; Reaka, A.; Venkatarangan, P.; Duca, J.S.; Wang, S. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 1151-1160.
[7] Fara, D.C. QSPR Modeling of Complexation and Distribution of Organic Compounds, *Dissertationes Chimicae Universitatis Tartuensis*: Tartu, 2004.
[8] Xue, L.; Godden, J.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 881-886.
[9] McGregor, M.J.; Pallai, P.V. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 443-448.
[10] James, C.A.; Weininger, D. Daylight theory manual. Daylight Chemical Information Systems, Inc., Irvine, CA, 1995.
[11] Pickett, S.D.; Luttman, C.; Guerin, V.; Laoui, A.; James, E. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 144-150.
[12] Sheridan, R.P.; Miller, M.D.; Underwood, D.J.; Kearsley, S.K. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 128-136.
[13] Willett, P.; Barnard, J.M.; Downs, G.M. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.
[14] Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273-1280.