Chemometrics and Intelligent Laboratory Systems 169 (2017) 12–18

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

PyDescriptor: A new PyMOL plugin for calculating thousands of easily understandable molecular descriptors

Vijay H. Masand [a,*], Vesna Rastija [b]

[a] Department of Chemistry, Vidya Bharati Mahavidyalaya, Amravati, Maharashtra 444 602, India
[b] Department of Chemistry, Faculty of Agriculture, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia

ARTICLE INFO

Keywords:
Molecular descriptors
PyMOL
PyDescriptor
QSAR
Interpretation

ABSTRACT

The field of Quantitative Structure-Activity Relationship (QSAR) relies heavily on molecular descriptors. Among various guidelines suggested by Organisation for Economic Co-operation and Development (OECD), a very important guideline demands the mechanistic interpretation of a QSAR model. For this, a very attractive idea is to build a QSAR model using easily understandable molecular descriptors. To address this important issue, in the present work, we present an innovative chem-informatics tool, *PyDescriptor*. It can calculate a diverse pool of 11,145 molecular descriptors comprising easily understandable 1D- to 3D- descriptors encoding pharmacophoric patterns, atomic fragments and a variety of fingerprints. It is a new Python based plugin implemented within the commonly used visualization software PyMOL. *PyDescriptor* has several advantages like easy to install, open source, works on all major platforms (Windows, Linux, MacOS), easy to use through graphical user interface (GUI) and command-line, and output is saved in comma separated values (CSV) file format for further QSAR procedure. The plugin is freely available for academia.

1. Introduction

Computer Aided Drug Designing (CADD) has advanced with innovations in its thriving branches viz. Quantitative Structure-Activity Relationship (QSAR), molecular docking, pharmacophore modelling. The field of QSAR is among the oldest branches of CADD with its emphasis on prediction of activity/property (quantitative QSAR) and determination of pharmacophoric features or mechanistic interpretation (qualitative QSAR) [1–4].

Structure drawing and optimization, molecular descriptor calculations, model building and model validation are four basic steps of a typical QSAR analysis [5–8]. Molecular descriptors, which are used to represent the structural features in terms of numbers, encode valuable information about structure or patterns in the molecular structures [9–16].

Molecular descriptors have occupied unique place in chemistry, pharmaceutical sciences, quality control, etc. to provide valuable representation of molecular features in numerical and computational form for further evaluations [9–18]. With the progress of QSAR field, the types of descriptors have changed from simple and easily interpretative like number of carbon atoms, number of nitrogen atoms, log*P*, etc. to very complex descriptors like WHIM, BCUT, 3D-MoRSE, RDF, GETAWAY, and others [17,18]. These molecular descriptors are mostly classified as 1D-, 2D- and 3D- descriptors. The 1D- molecular descriptors represent bulk properties of compounds, such as the number of particular atoms, molecular weight, etc., and can be computed using molecular formula. 2D-molecular descriptors characterize structural information that can be calculated from 2D- structure of a molecule, such as the number of rings, the number of hydrogen bond acceptors, etc. 3D- molecular descriptors stand for structural information that has to be obtained from 3D- structure of a molecule, such as solvent accessible surface area with negative partial charge in the structure [17,18].

Manual calculation of descriptors like 3D-MoRSE, WHIM, BCUT, and similar complex (or esoteric [5]) descriptors was a very time consuming and laborious process [1,9–12,15,16]. To overcome this difficulty, computer programs were developed for computing descriptors either as independent software or as a part of QSAR software. The rapid developments in the field of computers and algorithms have made exact and

Abbreviations: OECD, Organisation for Economic Co-operation and Development; WHO, World health organisation; ADMET, Absorption, Distribution, Metabolism, Excretion and Toxicity; OLS, Ordinary Least Square; QSARINS-Chem, QSAR Insubria-Chemistry; GA, Genetic algorithm; MLR, Multiple linear Regression; OFS, Objective Feature Selection; SFS, Subjective Feature Selection; MMFF94, Molecular Mechanics Force Field 94; MAE, mean Absolute Error; CCC, Concordance Correlation Coefficient.

* Corresponding author.
E-mail address: vijaymasand@gmail.com (V.H. Masand).

http://dx.doi.org/10.1016/j.chemolab.2017.08.003
Received 11 May 2017; Received in revised form 15 July 2017; Accepted 10 August 2017
Available online 13 August 2017
0169-7439/

precise calculations of theoretical molecular descriptors possible in shorter time and cost-effective [1,9–12,15,16]. At present, there are many free and commercial softwares like Dragon (Talete) [17,18], PaDEL [19], MOE [20], Schrodinger [21], ChemDes [22], etc. which can calculate a variety of molecular descriptors viz. 1D- to 3D-, constitutional, topological, fingerprints. Some of these have been developed exclusively for the calculation of molecular descriptors only such as PaDEL-Descriptor [19], ChemDes [22], etc. while others are QSAR softwares which have descriptor calculation as one of their features (e.g., CODESSA Pro [23], Accelrys Discovery Studio [24], Sybyl-x [25], MOE [20]). Also, there are some open source libraries, such as JOELib [26,27], Chemistry Development Kit [28], and Chemical Descriptors Library [29], to name a few, which have molecular descriptor calculation functionality. It is reasonable that a good descriptor calculation software should have following features [19]:

1. Free or low-priced so that it is easy to purchase it.
2. Open source so that researchers could introduce their specific molecular descriptor calculations.
3. Has an easy to use graphical user interface (GUI).
4. Independent of operating system.
5. Possibly processes different molecular file formats like mol2, mol, sdf, etc.
6. Ability to compute numerous types of molecular descriptors.

A careful analysis of various currently available molecular descriptor calculating softwares reveals that many softwares lack one or more above mentioned features, besides, having its own advantages and limitations. An important area of research in the field of molecular descriptors is introduction of new descriptors or improvements in existing descriptors with easy correlation in terms of structural and pharmacophoric patterns [1,10–16,22,28,29]. Therefore, the field of molecular descriptors is dynamic and open for future developments like introduction of new softwares with ease of use and better user control functionalities, new descriptors with enhanced abilities to capture structural features [1,10–16,22,28,29].

Among various guidelines suggested by Organisation for Economic Co-operation and Development (OECD), a very important guideline demands the mechanistic interpretation of a QSAR model. For this, a very attractive idea is to build a QSAR model using easily understandable molecular descriptors. Unfortunately, the physical correlation of esoteric descriptors like WHIM, GETAWAY, RDF, etc. with one or more structural features/patterns is very complicated and an active area of qualitative and quantitative QSAR [5]. Therefore, there is need for introduction of easily understandable molecular descriptors. In the present work, we present a new PyMOL plugin, *PyDescriptor*, which has capacity to calculate 11,145 easily understandable molecular descriptors. It is a new chem-informatics tool which transforms a variety of structural features and local environment of a molecule to understandable 1D- to 3D- descriptors, which include encoding pharmacophoric patterns, atom-centred descriptors and a variety of fingerprints. These descriptors are either available in costly commercial softwares or in operating system dependent free softwares, thereby restricting their wide use. *PyDescriptor* possesses many advantageous features and plethora molecular descriptors, which justify its usefulness and wide acceptance in the field of QSAR and allied areas.

2. Experimental details

2.1. Plugin design and availability

PyDescriptor has been written in the object-orientated programming language Python 2.7.10 (64 bit) as a plugin for the three-dimensional molecular viewer PyMOL 1.8.2 and higher versions (Schrödinger, LLC. http://www.pymol.org/). Therefore, the advantages and limitations of Python 2.7.10 and PyMOL are associated with this plugin also. PyMOL is a widely-used software proficient in rendering and ray-tracing high resolution molecular representations in publication quality [30]. Due to availability of an open-source version of PyMOL, it is an attractive choice for academic and educational use [30]. Apart from visualizations of molecular structures, PyMOL has emerged as a calculation software due to availability of different open source plugins for a variety of purposes for example APBS for electrostatic map calculation, CAVER for calculation and visualization of tunnels, MIPTOOL for LogP calculation, DYNAMICS for molecular dynamic simulations with Gromacs, a few to mention [30]. In addition, LIQUID is an open source plugin for PyMOL, which is capable of generating pharmacophore model for a molecule. The output of LIQUID is available in the form of spheres and ellipsoids in the 3D- viewer of PyMOL [31]. Though, *PyDescriptor* uses the framework of PyMOL, it has been fully coded by our group. Practical information, such as a user guide/manual and application notes, along with the plugin '*PyDescriptor*', are available free of charge from authors.

2.2. System requirements and installation

In order to use *PyDescriptor*, a working installation of PyMOL version ≥1.8.2 on a standard Linux or Windows or MacOS installations with Python 2.7.10 is essential. *PyDescriptor* can be used without any dependencies i.e. there is no need to install any other module or software. At present, the plugin has been built to use MOL2 file format containing single molecule only. MOL2 format has the benefit of storing all the essential information for atom type, position, partial charges, and connectivity. In addition, it is also a well-known standardized format that many programs can read. It is one of the few public formats capable of supporting both a chemically-accurate description of small organic molecules as well as protein or nucleic acid also. Other formats for representing molecular structure have to be converted to an MOL2 file format for use in *PyDescriptor*. For this purpose, users can use open-source programs (e.g. Open babel, Avogadro) to convert other file formats into MOL2 format. While using MOL2 file format, all atom-typing and atomic partial charges assignments need to be performed correctly with all hydrogen atoms added. After successful completion of the descriptor calculations, the molecular descriptor values are automatically saved in CSV file format.

2.3. Parsing and calculations

PyDescriptor performs the main task of reading the MOL2 files and calculating the molecular descriptor value for all the MOL2 files located in the folder (for windows users, C:\PyDescriptor). As shown in Fig. 1, when the user clicks 'Compute descriptors', the plugin executes the calculation of molecular descriptors. The values for all the molecular descriptors are entered and automatically saved iteratively into the CSV columns along with the name of MOL2 file in the first column.

The following set of codes is used to read MOL2 files:

```
import os, glob, csv, pymol
os.chdir('C:\PyDescriptor')
from pymol import cmd, stored, util
path = os.path.dirname(pymol.__script__)
cmd.delete('all')
mol_files = glob.glob(os.path.join(path, '*.mol2'))
```

***PyDescriptor* Protocol:**

- Read all the MOL2 files from a particular folder (for windows users, C:\PyDescriptor)
- Calculate the molecular descriptors for all the molecules in the given folder
- Read the name of the files and enter in the CSV file together with their corresponding descriptor values