BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 8 2013, pages 1092–1094
doi:10.1093/bioinformatics/btt105

Systems biology

Advance Access publication March 14, 2013

ChemoPy: freely available python package for computational biology and chemoinformatics

Dong-Sheng Cao[1], Qing-Song Xu[2], Qian-Nan Hu[3] and Yi-Zeng Liang[1,*]

[1]Research Center of Modernization of Traditional Chinese Medicines, Central South University, Changsha 410083, P. R. China, [2]School of Mathematics and Statistics, Central South University, Changsha 410083, P. R. China and [3]Key Laboratory of Combinatorial Biosynthesis and Drug Discovery (Wuhan University), Ministry of Education, Wuhan 430071, P. R. China

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Molecular representation for small molecules has been routinely used in QSAR/SAR, virtual screening, database search, ranking, drug ADME/T prediction and other drug discovery processes. To facilitate extensive studies of drug molecules, we developed a freely available, open-source python package called chemoinformatics in python (ChemoPy) for calculating the commonly used structural and physicochemical features. It computes 16 drug feature groups composed of 19 descriptors that include 1135 descriptor values. In addition, it provides seven types of molecular fingerprint systems for drug molecules, including topological fingerprints, electro-topological state (E-state) fingerprints, MACCS keys, FP4 keys, atom pairs fingerprints, topological torsion fingerprints and Morgan/circular fingerprints. By applying a semi-empirical quantum chemistry program MOPAC, ChemoPy can also compute a large number of 3D molecular descriptors conveniently.

Availability: The python package, ChemoPy, is freely available via http://code.google.com/p/pychem/downloads/list, and it runs on Linux and MS-Windows.

Contact: yizeng_liang@263.net

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 27, 2012; revised on February 22, 2013; accepted on February 25, 2013

1 INTRODUCTION

Molecular features for small molecules have frequently been used in the development of machine learning in QSAR/QSPR, virtual screening, database search, similarity search, ranking, drug ADME/T prediction and other drug discovery processes (Cao *et al.*, 2010, 2011, 2012a, b; Dea-Ayuela *et al.*, 2008; Du *et al.*, 2005, 2008a, b, 2009; Gola *et al.*, 2006, González-Díaz *et al.*, 2005; Prado-Prado *et al.*, 2008, 2009, 2010; van de Waterbeemd *et al.*, 2003; Wang *et al.*, 2011; Wei *et al.*, 2009; Yan *et al.*, 2012; Zhu *et al.*, 2011). These descriptors capture and magnify distinct aspects of molecular topology to investigate how molecular structure affects molecular property. Currently, these features were widely used to characterize ligand molecules in the protein–ligand network and predict new protein–ligand associations to identify potential drug targets (Campillos *et al.*, 2008; Cao *et al.*, 2012a, b; Chou *et al.*, 2006; Izrailev *et al.*, 2004; Keiser *et al.*, 2007; Prado-Prado *et al.*, 2011a, b; Viña *et al.*, 2009), following the spirit of chemogenomics.

Several programs for computing molecular descriptors have been developed, such as MARCH-INSIDE, TOPS-MODE, TOMO-COMD, Dragon, CODESSA, Molconn-Z (http://www.edusoft-lc.com/molconn/), Chemistry Development Kit (CDK), Indigo (http://ggasoftware.com/opensource/indigo), JOELib, RDKit (http://www.rdkit.org/) and Avogadro (González-Díaz *et al.*, 2008; Hanwell *et al.*, 2012; Katritzky *et al.*, 1994; Marrero-Ponce *et al.*, 2002; Pérez-González *et al.*, 2003; Steinbeck *et al.*, 2003; Todeschini *et al.*, 2010; Wegner, 2005). Unfortunately, some of these tools are not comprehensive, or are limited to only a certain kind of features. Additionally, some are not freely and easily accessible.

We implemented a selection of sophisticated molecular features and provided them as a package for the free and open-source software environment python. The ChemoPy package aims at providing the user with comprehensive implementations of these descriptors in a unified framework to allow easy and transparent computation. To our knowledge, ChemoPy is the first open-source package computing a large number of molecular features based on the MOPAC optimization. We recommend ChemoPy to analyse and represent the drugs or ligand molecules under investigation. Further, we hope that the package will be helpful when exploring questions concerning drug activity, drug ADME/T and drug–target interactions in the context of computational biology. After accomplishing the previous goal, we expect that our/other groups may use the free code of ChemoPy and the new QSAR models to implement public web servers, such as MIND-BEST (González-Díaz *et al.*, 2011). The users can run predictions of libraries of compounds using SMILES codes as input.

2 CHEMOPY FEATURES

The ChemoPy package contains several functions and modules manipulating drug molecules. To obtain molecular structures easily, ChemoPy provides a download module, by which the user could easily get molecular structures from four databases (i.e. KEGG, PubChem, DrugBank and CAS) by providing IDs.

```
>>> from pychem.pychem import getmol
>>> smi1 = getmol.GetMolFromNCBI('2244')
>>> print smi1
```

*To whom correspondence should be addressed.

```
CC(Oc1ccccc1C(O)=O)=O
>>> smi2 = getmol.GetMolFromKegg('D00109')
>>> print smi2
CC(Oc1ccccc1C(O)=O)=O
```

ChemoPy can compute a large number of 2D and 3D descriptors. A list of structural and physicochemical features covered by ChemoPy is summarized in Table 1 (see also detailed descriptor list in Supplementary Material S1). There are two means to compute these molecular descriptors from small molecules. One is to use the built-in modules. There exist 19 modules responding to the calculation of descriptors from 16 feature groups. The instruction for each module is provided in the form of HTML in ChemoPy (see Supplementary Material S2). We could import related functions to compute these features. For example, the topology module includes a number of functionalities used for calculating various topological descriptors. The user could conveniently use them as need.

```
>>> from pychem.pychem import Chem, topology
>>> mol = Chem.MolFromSmiles("CC(Oc1ccccc1C(O)=O)=O")
>>> Weiner = topology.CalculateWeiner(mol)
>>> Alltopology = topology.GetTopology(mol)
```

Table 1. List of ChemoPy computed features for small molecules

Feature group	Features	Number of descriptors
Constitution	Constitutional descriptors[a]	30
Topology	Topological descriptors[a]	35
Connectivity	Connectivity indices[a]	44
E-state	E-state descriptors	245
Kappa	Kappa shape descriptors[a]	7
Basak	Basak information indices	21
Burden	Burden descriptors	64
Autocorrelation	Moreau-Broto autocorrelation	32
	Moran autocorrelation	32
	Geary autocorrelation	32
Charge	Charge descriptors	25
Property	Molecular property[a]	6
MOE-type	MOE-type descriptors[b]	60
Geometric	Geometric descriptors	12
CPSA	CPSA descriptors	30
RDF	RDF descriptors	180
MoRSE	MoRSE descriptors	210
WHIM	WHIM descriptors	70
Fingerprints	Topological fingerprints[b]	2048
	MACCS keys[b]	166
	FP4 keys[c]	307
	E-state fingerprints[b]	79
	Atom pairs fingerprints[b]	—
	Topological torsions[b]	—
	Morgan fingerprints[b]	—

[a]Some of these features are from RDKit.
[b]These features are from RDKit.
[c]These features are from OpenBabel.

The other is to call the PyChem2d or PyChem3d class by importing the pychem module, which encapsulates commonly used descriptor calculation methods. PyChem2d and PyChem3d are responsible for the calculation of 2D and 3D molecular descriptors, respectively. We could construct a PyChem2d or PyChem3d object with a molecule input and then call corresponding methods to calculate these features.

```
>>> from pychem.pychem import PyChem2d, PyChem3d
>>> des1 = PyChem2d()
>>> des1.ReadMolFromSmile("CC(Oc1ccccc1C(O)=O)=O")
>>> AllConnectivity = des1.GetConnectivity()
>>> des2 = PyChem3d()
>>> des2.ReadMol("CC(Oc1ccccc1C(O)=O)=O")
>>> All3D = des2.GetAllDescriptor()
```

In ChemoPy, molecular structures are optimized by the AM1 method in MOPAC. MOPAC input file is directly prepared by Pybel and OpenBabel. The detailed introductions for all descriptors are provided in the ChemoPy manual (see Supplementary Material S3). A user guide for the use of ChemoPy is included to guide how the user uses it to calculate the needed features (see Supplementary Material S4).

ChemoPy is written by the pure python language. We chose to use python because it is open source, and there already exist packages to handle small molecules [e.g. Pybel (O'Boyle *et al.*, 2008b), PyMol and Cinfony (O'Boyle *et al.*, 2008a)]. It is convenient for ChemoPy to analyse drug molecules processed by Cinfony or RDKit. ChemoPy is available for two operating systems: Linux and Windows. ChemoPy depends on Pybel, RDKit, OpenBabel (O'Boyle *et al.*, 2011) and MOPAC (Stewart, 1990). Moreover, it needs the support of scientific library for python (SciPy).

3 DISCUSSION

ChemoPy contains a selection of molecular descriptors to analyse, classify and compare complex molecular network. They facilitate to exploit machine-learning techniques to drive hypothesis from complex molecular datasets. The usefulness of these molecular descriptors covered by ChemoPy for representing structural features of small molecules has been sufficiently demonstrated by a number of published studies of the development of machine-learning classification systems. The ChemoPy implementation of each of these algorithms was extensively tested by using a number of test molecules. The computed descriptor values were compared with the known values for these molecules to ensure that our computation is accurate.

Owing to the modular structure of ChemoPy, extensions or new functionalities can be implemented easily without complex and time-consuming alterations of the source code. In future work, we plan to apply the integrated features on various biological research questions, and extending the range of functions with new promising descriptors for the coming versions of ChemoPy.

ACKNOWLEDGEMENTS

The authors thank three anonymous referees for their constructive comments, which greatly helped improve on the original version of the manuscript.

Funding: National Natural Science Foundation of China (21075138, 21275164 and 11271374).

Conflict of Interest: none declared.

REFERENCES

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Cao,D.S. *et al.* (2010) Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J. Chemometr.*, **24**, 584–595.

Cao,D.S. *et al.* (2011) In silico classification of human maximum recommended daily dose based on modified random forest and substructure fingerprint. *Anal. Chim. Acta.*, **692**, 50–56.

Cao,D.S. *et al.* (2012a) Computer-aided prediction of toxicity with substructure pattern and random forest. *J. Chemometr.*, **26**, 7–15.

Cao,D.S. *et al.* (2012b) Large-scale prediction of drug-target interactions using protein sequences and drug topological structures. *Anal. Chim. Acta.*, **752**, 1–10.

Chou,K.C. *et al.* (2006) Predicting networking couples for metabolic pathways of *Arabidopsis*. *EXCLI J.*, **5**, 55–65.

Dea-Ayuela,M.A. *et al.* (2008) HP-Lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a *Leishmania infantum* sequence. *Bioorg. Med. Chem.*, **16**, 7770–7776.

Du,Q.S. *et al.* (2005) Heuristic molecular lipophilicity potential (HMLP): a 2D-QSAR study to LADH of molecular family pyrazole and derivatives. *J. Comput. Chem.*, **26**, 461–470.

Du,Q.S. *et al.* (2008a) Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). *J. Comput. Chem.*, **29**, 211–219.

Du,Q.S. *et al.* (2008b) Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. *Curr. Protein Pept. Sci.*, **9**, 248–259.

Du,Q.S. *et al.* (2009) Fragment-based quantitative structure-activity relationship (FB-QSAR) for fragment-based drug design. *J. Comput. Chem.*, **30**, 295–304.

Gola,J. *et al.* (2006) ADMET property prediction: the state of the art and current challenges. *QSAR Comb. Sci.*, **25**, 1172–1180.

González-Díaz,H. *et al.* (2005) Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. *Bioorg. Med. Chem.*, **13**, 1119–1129.

González-Díaz,H. *et al.* (2008) Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr. Top. Med. Chem.*, **8**, 1676–1690.

González-Díaz,H. *et al.* (2011) MIND-BEST: web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical experimental study of G3PDH protein from *Trichomonas gallinae*. *J. Proteome. Res.*, **10**, 1698–1718.

Hanwell,M.D. *et al.* (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Chemoinform.*, **4**, 17.

Izrailev,S. *et al.* (2004) Enzyme classification by ligand binding. *Proteins*, **57**, 711–724.

Katritzky,A.R. *et al.* (1994) *CODESSA Comprehensive Descriptors for Structural and Statistical Analysis*. Reference manual.

Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotech.*, **25**, 197–206.

Marrero-Ponce,Y. *et al.* (2002) *TOMOCOMD software, version 1.0, 2002*. Central University of Las Villas.

O'Boyle,N. *et al.* (2008a) Cinfony—combining open source cheminformatics toolkits behind a common interface. *Chem. Cent. J.*, **2**, 24.

O'Boyle,N. *et al.* (2008b) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 1–5.

O'Boyle,N. *et al.* (2011) Open babel: an open chemical toolbox. *J. Cheminform.*, **3**, 1–14.

Pérez-González,M. *et al.* (2003) TOPS-MODE based QSARs derived from heterogeneous series of compounds. Applications to the design of new herbicides. *J. Chem. Inf. Comput. Sci.*, **43**, 1192–1199.

Prado-Prado,F.J. *et al.* (2008) Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg. Med. Chem.*, **16**, 5871–5880.

Prado-Prado,F.J. *et al.* (2009) Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg. Med. Chem.*, **17**, 569–575.

Prado-Prado,F.J. *et al.* (2010) Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. *Bioorg. Med. Chem.*, **18**, 2225–2231.

Prado-Prado,F.J. *et al.* (2011a) Using entropy of drug and protein graphs to predict FDA drug-target network: theoretical-experimental study of MAO inhibitors and hemoglobin peptides from *Fasciola hepatica*. *Eur. J. Med. Chem.*, **46**, 1074–1094.

Prado-Prado,F.J. *et al.* (2011b) 2D MI-DRAGON: a new predictor for protein-ligands interactions and theoretic-experimental studies of US FDA drug-target network, oxoisoaporphine inhibitors for MAO-A and human parasite proteins. *Eur. J. Med. Chem.*, **46**, 5838–5851.

Steinbeck,C. *et al.* (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.

Stewart,J.P. (1990) MOPAC: a semiempirical molecular orbital program. *J. Comput. Aided Mol. Des.*, **4**, 1–103.

Todeschini,R. *et al.* (2010) *Molecular Descriptors for Chemoinformatics*. Wiley-VCH GmbH & Co. KGaA, Weinheim.

van de Waterbeemd,H. *et al.* (2003) ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.*, **2**, 192–204.

Viña,D. *et al.* (2009) Alingment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. *Mol. Pharm.*, **6**, 825–835.

Wang,J.M. *et al.* (2011) Recent advances on aqueous solubility prediction. *Comb. Chem. High Throughput Screen.*, **14**, 328–338.

Wegner,J.K. (2005) *JOELib: Graph/Data Mining and Clustering*. ACS Meeting.

Wei,H. *et al.* (2009) Investigation into adamantane-based M2 inhibitors with FB-QSAR. *Med. Chem.*, **5**, 305–317.

Yan,J. *et al.* (2012) Comparison of quantitative structure-retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds. *J. Chromatogr. A*, **1223**, 118–125.

Zhu,J.Y. *et al.* (2011) Recent developments of in silico predictions of oral bioavailability. *Comb. Chem. High Throughput Screen.*, **14**, 362–374.