BIOINFORMATICS APPLICATION NOTE

Vol. 31 no. 2 2015, pages 279–281
doi:10.1093/bioinformatics/btu624

Systems biology

Advance Access publication September 22, 2014

Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions

Dong-Sheng Cao[1,*,†], Nan Xiao[2,†], Qing-Song Xu[2] and Alex F. Chen[1]

[1]School of Pharmaceutical Sciences and [2]School of Mathematics and Statistics, Central South University, Changsha 410083, P. R. China

Associate Editor: Jonathan Wren

ABSTRACT

Summary: In chemoinformatics and bioinformatics fields, one of the main computational challenges in various predictive modeling is to find a suitable way to effectively represent the molecules under investigation, such as small molecules, proteins and even complex interactions. To solve this problem, we developed a freely available R/Bioconductor package, called Compound–Protein Interaction with R (Rcpi), for complex molecular representation from drugs, proteins and more complex interactions, including protein–protein and compound–protein interactions. Rcpi could calculate a large number of structural and physicochemical features of proteins and peptides from amino acid sequences, molecular descriptors of small molecules from their topology and protein–protein interaction and compound–protein interaction descriptors. In addition to main functionalities, Rcpi could also provide a number of useful auxiliary utilities to facilitate the user's need. With the descriptors calculated by this package, the users could conveniently apply various statistical machine learning methods in R to solve various biological and drug research questions in computational biology and drug discovery.

Availability and implementation: Rcpi is freely available from the Bioconductor site (http://bioconductor.org/packages/release/bioc/html/Rcpi.html).

Contact: oriental-cds@163.com

Received on May 30, 2014; revised on September 12, 2014; accepted on September 15, 2014

1 INTRODUCTION

To develop a powerful model for prediction tasks, one of the most important things to consider is how to effectively represent the molecules under investigation such as small molecules, proteins and even complex interactions, by a descriptor. In the field of chemoinformatics, molecular descriptors for small molecules have frequently been used in quantitative structure-activity/property relationship (QSAR/QSPR), virtual screening, database search, ranking, drug ADME/T prediction and other drug discovery processes (Cao *et al.*, 2011; Cherkasov *et al.*, 2014; Gola *et al.*, 2006, Willett, 2014). These descriptors capture and magnify distinct aspects of molecular topology to investigate how molecular structures affect molecular properties. In the field of bioinformatics, sequence-derived structural and physicochemical features have been widely used for predicting protein structural and functional classes, protein–protein interactions, subcellular locations and peptides of specific properties, etc (Chou *et al.*, 2008; Rangwala *et al.*, 2005; Shen *et al.*, 2007; Ye *et al.*, 2011; Zhang *et al.*, 2005; Zhang *et al.*, 2012). These features are highly useful for representing and distinguishing proteins or peptides of different structural, functional and interaction profiles. Currently, their combinations were routinely used to characterize drug–target interactions and predict new drug–target associations to identify potential drug targets (Cao *et al.*, 2013b; He *et al.*, 2010; Prado-Prado *et al.*, 2011), following the spirit of chemogenomics.

Several programs for computing molecular features have been developed, such as TOPS-MODE, Cinfony, Dragon, CODESSA, PROFEAT, BioJava, BioPython, PseAAC, ProPy, etc (Cao *et al.*, 2013a, c; Cock *et al.*, 2009; Du *et al.*, 2012; Holland *et al.*, 2008; Katritzky *et al.*, 1994; Li *et al.*, 2006; O'Boyle *et al.*, 2008; Pérez-González *et al.*, 2003; Todeschini *et al.*, 2010). Although a number of tools, which are either open sources or commercial softwares, have been developed and widely used in the two fields, their applications only focus on the analysis of either small molecules or proteins. To the best of our knowledge, there is currently no open-source code or tools available for the integration and analysis of increasingly popular interaction problems.

We developed a comprehensive molecular representation tool, called Compound–Protein Interaction with R (Rcpi), to emphasize the integration of chemoinformatics and bioinformatics into a chemogenomics platform for drug discovery. Rcpi mainly focuses on the study of molecular representation techniques for not only small molecules and proteins but also interactions of protein–protein and compound–protein. We recommend Rcpi to analyze and represent various complex molecular data under investigation. Further, we hope that the package will be helpful when exploring questions concerning structures, functions and interactions of various molecular data in the context of systems biology.

2 PACKAGE DESCRIPTION

The Rcpi package aims at offering a unique and comprehensive toolkit for complex molecular representations from small molecules, proteins and more complex interactions (see Table 1). To make the Rcpi package fully functional, we recommend the users to install the Enhances packages by using:

```
source('http://bioconductor.org/biocLite.R')
```

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Table 1. List of various types of descriptors for complex molecular data by Rcpi

Data types	Feature groups	Number of descriptors
Proteins	Amino acid composition	8420
	Autocorrelation	720[a]
	Composition, transition and distribution	147
	Conjoint traid descriptors	343
	Quasi-sequence order	160[a]
	Pseudo-amino acid composition	130[a]
	PSSM profile	–
	PCM	–
	GO similarity	–
	Sequence similarity	–
Compounds	Constitutional	15
	Topological	183
	Geometrical	49
	Electronic	34
	Hybrid	23
	Molecular property	4
	Fingerprints	10
	Maximum common substructure	–
Compound–protein interaction	Type 1	Nc + Np
	Type 2	Nc × Np
Protein–protein interaction	Type 1	Np + Np
	Type 2	Np + Np
	Type 3	Np × Np

Note: [a]The number of descriptors depends on the choice of the number of properties of amino acids and the choice of the parameter in corresponding algorithms. Nc and Np denote the number of molecular descriptors for compounds and proteins, respectively.

```
biocLite['Rcpi', dependencies = c('Imports', 'Enhances')]
```

Rcpi mainly covers the following four functionalities:

(a) For small molecules, Rcpi could (i) calculate >300 molecular descriptors, including constitutional, topological, geometrical, electronic, hybrid and molecular property descriptors; (ii) calculate 10 types of molecular fingerprints, including standard and extended Daylight fingerprints, graph fingerprints based on simple connectivity, hybridization fingerprints based only on hybridization state, FP4 keys, E-state fingerprints, MACCS keys, PubChem fingerprints, KR fingerprints defined by Klekota and Roth, short path fingerprints, etc; (iii) realize parallelized pair-wise similarity computation derived by fingerprints and five types of similarity measures within a list of small molecules; (iv) realize parallelized chemical similarity search with selected similarity metrics and maximum common substructure search between one query molecule and one molecular database.

(b) For protein sequences, Rcpi could (i) calculate a large number of commonly used structural and physicochemical descriptors, such as amino acid composition, autocorrelation, composition, transition, distribution, conjoint traid, quasi-sequence order and pseudo amino acid composition descriptors; (ii) calculate six types of generalized scale-based descriptors for proteochemometric (PCM) modeling, such as generalized scale-based descriptors derived by principal components analysis, amino acid properties, molecular descriptors, factor analysis, multidimensional scaling, and generalized BLOSUM/PAM matrix-derived descriptors; (iii) calculate profile-based protein features based on position-specific scoring matrix (PSSM); (iv) realize parallelized similarity computation derived by protein sequence alignment and Gene Ontology (GO) semantic similarity measures between a list of protein sequences/GO terms/Entrez Gene IDs.

(c) For interaction data, by combining various types of descriptors for drugs and proteins, interaction descriptors representing protein-protein or compound-protein interactions could be conveniently generated with Rcpi, including (i) two types of compound–protein interaction descriptors; (ii) three types of protein–protein interaction descriptors.

(d) Several useful auxiliary utilities are included in Rcpi: (i) parallelized molecule and protein sequence retrieval from several online databases, such as PubChem, ChEMBL, KEGG, DrugBank, UniProt, RCSB PDB, etc; (ii) molecular reading/writing in SMILES/SDF formats for small molecules and FASTA/PDB formats for proteins; (iii) molecular format conversion between ~140 types of molecular formats defined by OpenBabel.

3 DISCUSSION

Rcpi contains a selection of molecular descriptors to analyze, classify and compare complex molecular network in the context of network biology/pharmacology. They facilitate to exploit machine learning techniques to drive hypothesis from complex molecular datasets. The usefulness of these molecular descriptors covered by Rcpi for representing structural features of various molecular data has been sufficiently demonstrated by a number of published studies of the development of machine learning prediction systems.

In the future work, we plan to apply integrated features on various biological and drug research questions, and extend the range of functions with new promising descriptors for the coming versions of Rcpi.

Funding: This study was supported by the National key basic research program (2015CB910700) and the National Natural Science Foundation of China (Grant No. 81402853) and the Postdoctoral Science Foundation of Central South University. The studies meet with the approval of the university's review board.

Conflict of interest: none declared.

REFERENCES

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.* (2013a) ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics*, **29**, 1092–1094.

Cao,D.S. *et al.* (2013b) Genome-scale screening of drug-target associations relevant to Ki binding affinity using a chemogenomics approach. *PLoS One*, **8**, e57680.

Cao,D.S. *et al.* (2013c) propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **29**, 960–962.

Cherkasov,A. *et al.* (2014) QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.*, **57**, 4977–5010.

Chou,K.C. *et al.* (2008) Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **3**, 153–162.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Du,P.F. *et al.* (2012) PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.*, **425**, 117–119.

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

He,Z. *et al.* (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One*, **5**, e9603.

Holland,R.C.G. *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

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

Li,Z.R. *et al.* (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **34**, W32–W37.

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

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.* (2011) 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.

Rangwala,H. *et al.* (2005) Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, **21**, 4239–4247.

Shen,J. *et al.* (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.

Su,C.T. *et al.* (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics*, **7**, 319.

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

Willett,P. *et al.* (2014) The calculation of molecular structural similarity: principles and practice. *Mol. Inf.*, **33**, 403–413.

Ye,X.G. *et al.* (2011) An assessment of substitution scores for protein profile-profile comparison. *Bioinformatics*, **27**, 3356–3363.

Zhang,Q.C. *et al.* (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.

Zhang,Q.D. *et al.* (2005) Improved method for predicting β-turn using support vector machine. *Bioinformatics*, **21**, 2370–2374.