BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–3

Chemistry Aware Model Builder (camb): An R package for bioactivity and property modeling of small molecules and proteins

Daniel S. Murrell [1*], Isidro Cortes-Ciriano [2,*], Gerard J.P. van Westen [3], Thérèse E. Malliavin [2,†], Andreas Bender[1,†], Robert C. Glen[1,†]

[1]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, United Kingdom.
[2]Unite de Bioinformatique Structurale, Institut Pasteur and CNRS UMR 3825, Structural Biology and Chemistry Department, 2528, rue Dr. Roux, 75 724 Paris, France.
[3]European Molecular Biology Laboratory European Bioinformatics Institute Wellcome Trust Genome Campus, Hinxton, United Kingdom.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: *camb* is an R package that can be used for the rapid generation of quantitative predictive models in the area of medicinal chemistry (QSAR, QSPR, QSAM, proteochemometrics and chemogenomics). It is aimed at both amateur and advanced R users. Its capabilities include the standardisation and representation of chemical structures, computation of 905 two-dimensional and 14 fingerprint type descriptors for small molecules, the computation of 8 types of amino acid descriptors and 13 different whole protein sequence descriptors, f[illegible]ased statistical preprocessing, generation of predictive models (R package *caret*), as well as techniques to ensemble these models (R package *caretEnsemble*). Results can be visualised through high-quality, customisable plots (R package *ggplot2*).

Availability: *camb* is written in R, C++, Python and Java and is available open source at https://github.com/cambDI/camb. Two tutorials are also included.

Contact: d[illegible]ell@gmail.com or isidrolauscher@gmail.com

1 INTRODUCTION

The advent of high-throughtput technologies over the last two decades has led to a vast [illegible]ase of compound, bioactivity and genomic databases (Bender, 2010). This rampant increase in the amount of chemical and biological information has been exploited by emergent fields in drug discovery such as chemogenomics or proteochemometrics (PCM) (van Westen *et al.*, 2011; Cortes Ciriano *et al.*, 2014).

The R programming language provides a excellent platform for statistical analyses (R Core Team, 2013), and its applicability in medicinal chemistry has been reviewed elsewhere (Mente and Kuhn, 2012). Although R is extensively used in diverse biological domains, *e.g.* genomics (Gentleman *et al.*, 2004), the availability of R packages for cheminformatics and medicinal chemistry is limited. Nonetheless, R still constitutes the most frequent choice in the medicinal chemistry literature for compounds bioactivity and property modelling (Mente and Kuhn, 2012). In general, these type of studies share a common structure, which can be summarised in 4 model generation steps: (i) compound standardisation, (ii) descriptor calculation, (iii) preprocessing, model training and validation, and (iv) bioactivity/property prediction for new molecules.

Currently available R packages provide the capability for a subset of the previous steps. For instance, R packages *chemmineR* (Cao *et al.*, 2008) and *rcdk* (Guha, 2007) enable the manipulation of SDF and SMILES files, the calculation of physicochemical descriptors, the clustering of molecules, and the retrieval of compounds from PubChem (Wang *et al.*, 2012). On the machine learning side, the *caret* package provides a unified platform for the training of machine learning models (Kuhn, 2008).

Here, we present the R package *camb*: **C**hemistry **A**ware **M**odel **B**uilder, which aims to address the current lack of an R framework encompasing all four steps mentioned above. The package has been conceived in a way that users with little programming skill are able to generate competitive predictive models and high-quality plots under default operation. However, each function can be utilised to fulfil the more versatile needs of more experienced users.

Overall, *camb* enables the generation of predictive models (QSAR, QSPR, QSAM, PCM and chemogenomics) starting from chemical structure files, optional protein sequences, and the associated properties or bioactivities. Moreover, *camb* is the first R package [illegible]ing the manipulation of chemical structures *via* the C-writt[illegible]ligo API (GGA Software Services, 2013), and the calculation of: (i) PaDEL descriptors and fingerpints (Yap, 2011), (ii) hashed and unhashed Morgan fingerprints (Rogers and Hahn, 2010), and (iii) 8 types of amino acid descriptors. Two case studies illustrating the application of *camb* for QSPR modelling and

*Equal contributors

†to whom correspondence should be addressed

© Oxford University Press 2014.

PCM are available in the online supplementary information. In the following section we detail the main functionalities provided by *camb*.

2 DESCRIPTION

This section describes the tools provided by *camb* for (i) compound standardisation, (ii) descriptor calculation, (iii) preprocessing, model training and validation, and (iv) visualisation.

2.1 Compound standardization

In order to represent all molecules in a given dataset in the same way (compound standardisation), *camb* provides the function *StandardiseMolecules* which utilises Indigo's C API (GGA Software Services, 2013). SDF and SMILES formats are provided as molecule input options. The maximum number of fluorines, chlorines, bromines and iodines that a compound can exhibit in order to pass the standardisation process can be defined by the user. Additional arguments of this function include the removal of inorganic molecules or those compounds with a molecular mass above or below a given thresholds. *camb* makes use of Indigo's InChI plugin to standardise tautomers to the same SMILES representation by converting to InChI, discarding tautomeric information, and converting back to SMILES.

2.2 Descriptor calculation

Currently, *camb* supports the calculation of compounds descriptors and fingeprints from PaDEL (Yap, 2011), and circular Morgan fingerprints (Rogers and Hahn, 2010) as implemented in the RDkit (Landrum, 2006). The function *GeneratePadelDescriptors* permits the calculation of 905 2-dimensional descriptors and 10 PaDEL fingerprints.

Morgan fingerprints can be computed with the function *MorganFPs* through the python library RDkit (Landrum, 2006). Hashed fingerprints are calculated in binary format and with counts. Additionally, this function computes unhashed (keyed) fingerprints. In this case, each substructure in the dataset is assigned a unique position in a binary fingerprint. To calculate the fingerprint for each compound those positions in the fingerprint corresponding to the substructures present in a given compound are set to 1 (binary format) or the number of times the substructure appears in that compound (counts format).

From the above, it is apparent that the position of the set bits in an unhashed fingerprint directly depends on the dataset. To facilitate the application of predictive models trained on unhashed fingerprints, the function *MorganFPs* also allows the calculation of unhashed fingerprints for new compounds using a basis defined by the substructures present in a given chemical training set.

camb also enables the calculation of 13 types of whole protein sequence descriptors from UniProt identifiers (Xiao and Xu, 2014), as well as the calculation of 8 types of amino acid descriptors (van Westen *et al.*, 2013).

2.3 Model training and validation

Prior to model training, descriptors should to be statistically preprocessed (Andersson *et al.*, 2011). To this aim, several functions (see package documentation and tutorials) are provided, *e.g.* the removal of non-informative predictors or their conversion to z-scores.

camb invokes the R package *caret* to train individual machine learning models. Additionally, two ensemble modelling approaches, namely: greedy and stacking optimisation, have been integrated from the R package *caretEnsemble* (Mayer, 2013). Statistical metrics for model validation have also been included (Golbraikh and Tropsha, 2002).

2.4 Visualization

All plots are generated using the R package *ggplot2* (Wickham, 2009). Default options for plotting functions allow the generation of high-quality plots, however, the layer-based structure of ggplot objects allows for further tweaking by the addition of customisation layers. Visual depiction of compounds is also possible with the function *PlotMolecules*, utilises Indigo's C API. Visualization functions are explained in the tutorials.

3 CONCLUSIONS

In silico predictive models have proved a valuable tool for the optimisation of compounds potency, selectivity and safety profiles. In this context, *camb* provides a complete framework to (i) manipulate compound structures, (ii) generate compound and protein descriptors, and (iii) train and validate QSAR, QSPR, QSAM, PCM and chemogenomic models.

4 ACKNOWLEDGEMENTS

ICC thanks the Paris-Pasteur International PhD Programme for funding. ICC and TM thank CNRS, Institut Pasteur and ANR bipbip for funding. DM thanks Unilever for funding. GvW thanks EMBL (EIPOD) and Marie Curie (COFUND) for funding. AB thanks Unilever and the European Research Commission (Starting Grant ERC-2013-StG 336159 MIXTURE) for funding.

REFERENCES

Andersson, C. R., Gustafsson, M. G., and Strmbergsson, H. (2011). Quantitative chemogenomics: machine-learning models of protein-ligand interaction. *Current topics in medicinal chemistry*, **11**(15), 1978–1993. PMID: 21470169.

Bender, A. (2010). Databases: Compound bioactivities go public. *Nature Chemical Biology*, **6**(5), 309–309.

Cao, Y., Charisi, A., Cheng, L. C., Jiang, T., and Girke, T. (2008). Chemminer: a compound mining framework for r. *Bioinformatics*, **24**(15), 1733–1734.

Cortes Ciriano, I., U., Subramanian, V., Lenselink, E. B., Mendez Lucio, O., IJzerman, A. , Wohlfahrt, G., Prusis, P., Malli avin, T., van Westen, G. J., and Bender, A. (2014). Polypharmacology modelling using proteochemometrics: Recent developments and future prospects. *About to be submitted to Med. Chem. Comm.*

Gentleman, R. C., Carey, V. J., Bates, D. M., and others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

GGA Software Services (2013). Indigo cheminformatics library.

Golbraikh, A. and Tropsha, A. (2002). Beware of q2! *Journal of molecular graphics & modelling*, **20**(4), 269–276. PMID: 11858635.

Guha, R. (2007). Chemical informatics functionality in r. *Journal of Statistical Software*, **18**(6).

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, **28**(5), 1–26.

Landrum, G. (2006). Rdkit: Open-source cheminformatics.

Mayer, Z. (2013). caretensemble: Framework for combining caret models into ensembles. [r package version 1.0].

Mente, S. and Kuhn, M. (2012). The use of the r language for medicinal chemistry applications. *Current topics in medicinal chemistry*, **12**(18), 1957–1964. PMID: 23110531.

R Core Team (2013). R: A language and environment for statistical computing.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, **50**(5), 742–754. PMID: 20426451.

van Westen, G. J., Swier, R. F., Cortes-Ciriano, I., Wegner, J. K., Overington, J. P., Ijzerman, A. P., van Vlijmen, H. W., and Bender, A. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): Modeling performance of 13 amino acid descriptor sets. *J. Cheminf*, **5**(1), 42.

van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W. T., and Bender, A. (2011). Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, **2**, 16–30.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., Bolton, E., Gindulyte, A., and Bryant, S. H. (2012). PubChem's BioAssay database. *Nucleic acids research*, **40**(Database issue), D400–412. PMID: 22140110 PMCID: PMC3245056.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis.

Xiao, N. and Xu, Q. (2014). protr: Protein sequence descriptor calculation and similarity computation with r. R package version 0.2-1.

Yap, C. W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, **32**(7), 1466–1474.