BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 23 2010, pages 3000–3001
doi:10.1093/bioinformatics/btq556

Data and text mining

Advance Access publication September 30, 2010

Chembench: a cheminformatics workbench

Theo Walker[1], Christopher M. Grulke[1], Diane Pozefsky[2] and Alexander Tropsha[1,*]

[1]Division of Medicinal Chemistry and Natural Products, Eshelman School of Pharmacy and [2]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Advances in the field of cheminformatics have been hindered by a lack of freely available tools. We have created Chembench, a publicly available cheminformatics portal for analyzing experimental chemical structure–activity data. Chembench provides a broad range of tools for data visualization and embeds a rigorous workflow for creating and validating predictive Quantitative Structure–Activity Relationship models and using them for virtual screening of chemical libraries to prioritize the compound selection for drug discovery and/or chemical safety assessment.

Availability: Freely accessible at: http://chembench.mml.unc.edu

Contact: alex_tropsha@unc.edu

Received on August 11, 2010; revised on September 18, 2010; accepted on September 26, 2010

1 INTRODUCTION

Within the last decade, cheminformatics has emerged as a burgeoning discipline combining computational, statistical and informational methodologies with key concepts in chemistry and biology (Brown, 2005; Varnek and Tropsha, 2008). Cheminformatics addresses the fundamental problem of structure–activity (property) relationships as applied to many areas of chemical and biological research, providing the ability to use models for imputation of target activities or properties of untested compounds.

Opportunities for cheminformatics research have grown significantly with the advent of parallel chemical synthesis and high-throughput screening and publicly available data from projects such as the Molecular Libraries Initiative (Austin *et al*., 2004). For instance, PubChem (http://pubchem.ncbi.nlm.nih.gov/) currently contains nearly 27 million chemical compound records; almost one million of these have been tested in over 2600 bioassays with nearly 300 000 found active. Many other similarly structured databases have emerged recently (Oprea and Tropsha, 2006), providing a corpus of data rivaling the size and complexity of biological databases that established the need for bioinformatics.

Despite the abundance of databases of biologically active compounds in the public domain, the data remain largely underexplored because of the dearth of public domain tools for data analysis. Along with other recently emerging tools and toolkits such as CDK (Kuhn *et al*., 2010) and OCHEM (http://ochem.eu/), Chembench is poised to advance experimental research in chemical genomics, drug discovery and chemical safety assessment.

*To whom correspondence should be addressed.

2 METHODS

Chembench is a Java-based system, built with freely available technologies carefully chosen to ensure a stable, maintainable system. The front end of the website uses Java Server Pages (JSPs; McPherson, 2000) with Javascript. The Struts 2 framework (Roughley, 2007) provides the interface between data on the JSPs and Java objects. Java objects are mapped to a relational database using HIBERNATE (King *et al*., 2004).

Chembench implements several Quantitative Structure–Activity Relationship (QSAR) modeling methods and uses several commercial packages, i.e. MOLCONNZ (eduSoft, 2008), DRAGON (Talete, 2007), MOE (Lin, 2000) and MACCS keys (Symyx, 2005) for descriptor generation. The JChem suite (ChemAxon, 2010) is used for image generation and standardization of compounds. Scripts for dataset visualization are executed using MATLAB and R. Ensembles of QSAR models are built following a well-established workflow (shown as a diagram under the Modeling module) incorporating rigorous validation procedures (Tropsha, 2010). All calculations are executed on a 350-node Beowulf Linux cluster provided by UNC-Chapel Hill.

3 RESULTS

Chembench supports the following cheminformatics data analysis tasks structured as modules. Each module can be used independently or as part of an integrated study design.

- *Dataset Creation:* Chembench allows users to upload, store and standardize (Fourches *et al*, 2010) a set of chemical structures. To enable the QSAR modeling of a dataset, activity data for each compound must also be provided. Available descriptors are generated for each compound upon upload. An external set to validate models can be selected manually or automatically.
- *Dataset Visualization*: Several tools are available. The user can view the chemical structures, examine the distribution of activities, and generate a structure–activity heat map, using either Tanimoto similarity (Tanimoto, 1957) or Mahalanobis distance measure (Mahalanobis, 1936), to check for obvious relationships between global compound similarity and activity.
- *Modeling:* The modeling function allows the user to select a modeling dataset (either one of his uploaded datasets or a provided benchmark set) and build an ensemble of statistically validated models (i.e. a predictor) of the target property. Chembench currently supports model building with kNN (Zheng and Tropsha, 2000) and random forest (Breiman, 2001) techniques; support vector machines (Chang and Lin, 2001) are currently under development. As listed in Section 2, several commercial packages are used for descriptor generation.
- *Model Validation:* When selecting a completed predictor, the user is provided with the detailed statistics for estimating the

predictor's robustness such as a plot of the predicted versus actual activity for the external set, and the results of the *y*-randomization test.

- *Virtual screening*: The user may predict a specific activity or a spectrum of activities for a virtual chemical library or a single compound; available libraries include NCI diversity set (http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html) DrugBank (Wishart *et al*., 2008), ChEMBL (http://www.ebi.ac.uk/chembldb/) and Wombat (Olah *et al*., 2007); the user may also upload his own library. Several predictors developed by UNC's Molecular Modeling Lab are available and more are being added continuously. Prediction of activity is limited by the applicability domain (Tropsha, 2010), which may be tuned to provide more conservative or liberal predictions.

The user has control over many of the modeling parameters influencing the choice of descriptors, modeling algorithms, feature selection and the internal validation. We distinguish typical and advanced users, who are provided with differential options to control modeling parameters. Upon submission, the job is placed in a queue for execution and the user can monitor the status of the task or request email notification when the job completes.

Eleven benchmark datasets with continuous activity values and five datasets with binary activity values previously modeled and published by our group are included under the Modeling module. To illustrate the use of the portal, we have executed the embedded workflow using all available QSAR techniques, Dragon descriptors and default parameters for two benchmark sets. The highest external R^2-value for the blood–brain barrier permeability dataset (Zhang *et al*., 2008) was 0.73 and the test set prediction accuracy for discriminating Pgp substrates from inhibitors (de Cerqueira *et al.*, 1996) was 90%. Both results were in agreement with published values; calculations took from several minutes to several hours depending on the algorithm (random forest was faster than kNN).

Because there is a single workflow that supports a range of different techniques, it is easy to re-do a modeling run with simple changes. The presentation of statistics then allows the user to make direct comparison between the alternative selections made in modeling parameters. This is a significant difference from the current practice in cheminformatics, where workflows tend to rely on a single method or bundle a broad range of choices that are hard to investigate individually.

4 DISCUSSION

Covering the expanse of cheminformatics tools, ranging from chemical data visualization to creation of robust QSAR models to identification of novel chemicals with a desired activity profile, Chembench serves both the seasoned cheminformatician as well as the bench scientist. With the abundance of publicly available chemocentric data, this portal will enable knowledge mining and hypothesis generation across the breadth of biomolecular inquiries, from chemical properties and ADME characteristics to specific target binding/phenotype to chemical toxicity.

ACKNOWLEDGEMENTS

We thank Chemical Computing Group, Talete srl, eduSoft, ChemAxon and Sunset Molecular for their software licenses. We also thank Steven Fishback and UNC Information Technology Services for their support and members of the Molecular Modeling Lab for their input and help in testing.

Funding: National Institutes of Health grants (P20HG003898 and R01GM066940); Environmental Protection Agency grants (R832720 and RD83382501).

Conflict of Interest: none declared.

REFERENCES

Austin,C.P. *et al*. (2004) NIH molecular libraries initiative. *Science*, **306**, 1138–1139.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **1**, 5–32.

Brown,F. (2005) Editorial opinion: chemoinformatics—a ten year update. *Curr. Opin. Drug Discov. Dev.,* **8**, 298–302.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (last accessed date September 18, 2010).

ChemAxon (2010) JChem User's Guide, Version 5.3.5. Available at http://www.chemaxon.com/jchem/doc/user/ (last accessed date September 18, 2010).

eduSoft (2008) Software package for molecular topology analysis user's guide. Available at http://www.edusoft-lc.com/molconn/manuals/400/ (last accessed date September 18, 2010).

Fourches,D. *et al*. (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model*., **50,** 1189–1204.

King,G. *et al*. (2004) HIBERNATE – relational persistence for idiomatic java. Red Hat. Available at http://docs.jboss.org/hibernate/stable/core/reference/en/html/ (last accessed date September 18, 2010).

Kuhn,T. *et al*. (2010) CDK-Taverna: an open workflow environment for cheminformatics. *BMC Bioinform*., **11**,159–169.

Lin,A. (2000) QuaSAR-Descriptor. Available at http://www.chemcomp.com/journal/descr.htm (last accessed date September 18, 2010).

Mahalanobis,P. (1936) On the generalised distance in statistics. *Proc. Natl Inst. Sci. India,* **2,** 49–55.

McPherson,S. (2000) JavaServer pages: a developer's perspective. Available at http://java.sun.com/developer/technicalArticles/Programming/jsp/ (last accessed date September 18, 2010).

Olah,M. *et al*. (2007) WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery. In Schreiber,S. *et al*. (eds), *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, Wiley-VCH, New York, pp. 760–786.

Oprea,T. *et al*. (2006) Target, chemical and bioactivity databases – integration is key. *Drug Discov. Today,* **3**, 357–365.

Roughley,I. (2007) *Starting Struts 2*; *Lulu.com, Raleigh*.

Symyx (2005) *MACCS Structural Keys*, MDL Information Systems Inc., San Ramon, CA.

de Cerqueira Lima,P. *et al*. (2006) Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Info. Model*., **46**, 1245–1254.

Talete (2007) DRAGON for Windows and Linux. Available at http://www.talete.mi.it/help/dragon_help/ (last accessed date September 18, 2010).

Tanimoto,T. (1957) IBM Internal Report, 17 November, IBM Corp, Armonk.

Tropsha,A. (2010) Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.,* **29**, 476–488.

Wishart,D. *et al*. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.,* **6**, D901–D906.

Varnek,A. and Tropsha,A. (2008) *Cheminformatics Approaches to Virtual Screening*, RSC, London.

Zhang,L. *et al*. (2008) QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res*., **25**, 1902–1914.