SBMLDock: Docker Driven Systems Biology Tool Development and Usage

Etienne Z. Gnimpieba[1(✉)], Mathialakan Thavappiragasam[1], Abalo Chango[2], Bill Conn[1], and Carol M. Lushbough[1]

[1] Computer Science Department, University of South Dakota, Vermillion, SD, USA
{Etienne.Gnimpieba, Mathialakan.Thavappi, Bill.Conn, Carol.Lushbough}@usd.edu
[2] UPSP EGEAL, Institut Polytechnique LaSalle Beauvais, Beauvais, France
abalo.chango@lasalle-beauvais.fr

Abstract. A glut of Systems Biology tools and their lack of accessibility has significantly delayed bioscience advances that depend on the analysis of large scale systems with big datasets and High Performance Computing (HPC) resources. This work presents SBMLDock, the first Systems Biology Docker image that aims to advance scalability, usability and reproducibility in Systems Biology by making tools much more immediately available to the biological domain scientist, student, and educator, without requiring special training for use, and without losing the reproducibility aspect of their research. SBMLDock consists of one Docker image containing basic tools developed for Systems Biology Model manipulation (parallel model similarity analyzer, model checker, model splitter, model annotation, model extractor). The user can then pull up the Docker image, customize it and/or run each tool as service. Stored on the Docker hub, the image version is managed to assure research reproducibility. SBMLDock is available as a Docker file under CC licence at github https://github.com/USDBioinformatics/SBMLDock and the Docker image can be found in Docker hub at https://registry.hub.docker.com/u/usdbioinformatics/sbmldock/ with supplementary documents.

Keywords: SBMLDocker · Docker image · Systems biology · Reproducible research

1 Introduction

Emerging developments in Big Data, Systems Biology, and Integrative Biology introduce an increasing number of challenges in life science research. The primary objective of Software as a Service (SaaS) and platform as a service (PaaS) initiatives such as Workflow Management Systems (WMS) or Docker is to simplify researchers' ability to access, apply, and share analytic tools, workflows and data [1]. Executing an analytic tool can be very difficult if the researchers are not well prepared. Additionally, it is not always optimal to use systems biology tools due to deployment times that degrade the tool usability [1].

O. Roux and J. Bourdon (Eds.): CMSB 2015, LNBI 9308, pp. 282–285, 2015.
DOI: 10.1007/978-3-319-23401-4_24

The development of a container system (Docker) allows bioscience tool developers to hide complexity from researchers by providing a distributed container to embed any development module (service, tool, workflow, data storage) (https://www.docker.com/whatisdocker/). This method has been adopted in bioinformatics areas including the Galaxy infrastructure [2].

System Biology Markup Language(SMBL) is a machine-readable XML format for representing computational models of biological processes [3]. Software tools that support SBML as a format for reading and writing biological systems models facilitate their cooperative sharing, evaluation, and development. The XML-based SBML is the de facto standard file format for the storage and exchange of quantitative computational models in systems biology, supported by more than 220 software packages to date (March 2014) [3]. This includes several biological systems modeling tools (e.g. Systems Biology toolbox for Matlab, COPASI, EPISIM, Virtual Cell) and several databases for the representation and knowledge sharing (e.g. BioModels, BRENDA, KEGG).

2 SBMLDock

SBMLDock is the first systems biology Docker container for researchers, educators, and developers. We developed the first set of tools for SBML file manipulation including SBMLSplit, SBMLModeler, SBMLAnnotate. In order to complete our toolkit, we integrated recently published tools in the same series, such as ParaABioS [4], SBMLMerge, SBMLChecker [5], SBMLCompare [6]. Each tool has been integrated into a Docker image with a test dataset. The researcher can use this test data set to test each *tool.*

ParaABioS is an implementation of a parallel algorithm for bioscience elements similarity estimation [4]. This parallelization is critical when you involve the synonyms of bioscience terms because the curse of dimensionality becomes worse and requires HPC resources. ParaABioS uses heuristic techniques to measure similarity parameter values (distance and ratio) of the elements. The algorithm was implemented using SIMD data parallelization techniques in java.

ParaABioS requires four parameters to run, and provides the similarity results in a text file. Running in Docker, the syntax is `ParaABioS <inputfile1> <inputfile2> <distance> <ratio> Where <inputfile1>, <inputfile2>` are two bioscience element lists (metabolite, compound, protein, gene, etc.), `<distance> and <ratio>` are threshold values for edit distance and the ratio respectively.

`E.g. docker run -v /home/wjconn/SBMLDock/mount:/tmp -w /tmp usdbioinformatics/sbmldock ParaABioS file1.txt file2.txt 6 0.7`

SBMLChecker is a Systems Biology Markup Language model checker. SBMLChecker improves the online SBML validator by integrating meaning using semantic (ontology and database) checking [5, 7]. It uses the annotated URL ids of each element to measure the semantic strength of the reliability score. In order to execute SBMLChecker in Docker use the following command `SBMLChecker <sbmlinputfile>`. This will return a checking report printed in the system out or in a report output files store on your mounted directory.

E.g. docker run -v /home/wjconn/SBMLDock/mount:/tmp -w /tmp usdbioinformatics/sbmldock SBMLChecker one.xml

SBMLCompare is an implementation of ParABioS algorithm specific for SBML model comparison. In addition to naming similarity techniques used in ParABioS, SBMLCompare use biological annotated meanings to ensure the semantic similarity between models. SBMLCompare on the Docker can be use as follow SBMLCompare <inputfile1> <inputfile2> . This will provide a comparison report in 3 formats (text, excel or xml) in files named *sbml_compare_report*.

E.g. docker run -v /home/wjconn/SBMLDock/mount:/tmp -w /tmp usdbioinformatics/sbmldock SBMLCompare one.xml two.xml

SBMLMerge is an automatic merging tool for SBML models. Other existing merging tools for SBML models require human interaction. Using a heuristic algorithm, SBMLmerge provides a consistent merged model. This tool helps biologists combine sub-model from different sub-biosystems into a targeted biosystem. To execute SBMLMerge on Docker use the following syntax SBMLMerge <edit distance int[0-10]> <similarity ratio float[0-1]> <inputfile1> <inputfile2> <optional input files up to 6>. This will provide a merged SBML model mergedmodel.xml file in your mounted directory.

E.g. docker run -v /home/wjconn/SBMLDock/mount:/tmp -w /tmp usdbioinformatics/sbmldock SBMLMerge 6 0.7 /opt/SBMLMerge/one.xml /opt/SBMLMerge/two.xml

SBMLSplit is an SBML model extractor. A researcher can extract a sub-model based on reaction or compound (metabolite, species) list. SBMLSplit can be run on the Docker as SBMLSplit <flag> <inputfile> where your <flag> is C or R to split on Compound or Reaction respectively, and the <inputfile> is the SBML file you want to split. This provide 2 split SBML files (e.g. *S0.xml* and *S1.xml*), that are stored in your mounted folder.

E.g. docker run -v /home/wjconn/SBMLDock/mount:/tmp -w /tmp usdbioinformatics/sbmldock SBMLSplit C one.xml

SBMLModeler is an implementation of a data mining workflow for SBML model design from multiple data repositories (e.g. KEGG, SABIO-RK, BRENDA, …), using a top down approach with the pathway name as the entry. The current version of SBMLModeler focuses on a short pathway list for accuracy purposes. The list named Pathwayslist.txt can be found in the directory /opt/SBMLModeler/ in the SBMLDock image. Once you have your pathway picked out you can run SBMLModeler using the following command SBMLModeler <Path to store file> <Pathway name>.

E.g. docker run -v/home/wjconn/SBMLDock/mount:/tmp -w/tmp usdbioinformatics/sbmldock SBMLModeler. "folate biosynthesis"

SBMLAnnotate is an automatic annotation tool for SBML models. SBMLAnnotate evaluates the existing annotation degree of your SBML model (i.e. number of element annotated with ontologies or common databases such as SBO, KEGG) and proposes a reliable annotation to improve the model. To execute SBMLAnnotate use: SBMLAnnotate <inputfile> <outputfile>. This will save an out.xml file in your mounted directory as output.

```
E.g. docker run -v/home/wjconn/SBMLDock/mount:/tmp -w/tmp
usdbioinformatics/sbmldock SBMLAnnotate one.xml out.xml
```

3 Conclusion

Systems integration in life science research has become a complex challenge as data sets have grown. The ability to minimize the tools usage can be a tremendous asset for bioscience scientist. SBMLDock provides systems biology tools that allow developers and users to work together in minimizing the complexity of tool deployment and version management. This also greatly contributes toward the development of reproducible research.

Acknowledgement. This work has been partially supported by the National Science Foundation/EPSCoR Award No. IIA-1355423 and by the state of South Dakota, through BioSNTR.

References

1. Beasley, J.M., Coronado, G.D., Livaudais, J., Angeles-Llerenas, A., Ortega-Olvera, C., Romieu, I., Lazcano-Ponce, E., Torres-Mejía, G.: Alcohol and risk of breast cancer in Mexican women. Cancer Causes Control **21**, 863–870 (2010)
2. Cock, P.J.A., Grüning, B.A., Paszkiewicz, K., Pritchard, L.: Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. Peer J. **1**, e167 (2013)
3. Hucka, M.: Systems biology markup language (SBML). In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (eds.) Encyclopedia of Systems Biology SE – 1091, pp. 2057–2063. Springer, New York (2013)
4. Thavappiragasam, M., Lushbough, C.M., Gnimpieba, E.Z.: Heuristic parallelizable algorithm for similarity based biosystems comparison. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB 2014, pp. 782–789 (2014)
5. Thavappiragasam, M., Lushbough, C., Gnimpieba, E.: SBMLChecker, a Semantic approach for SBML model reliability evaluation, 2–5 (2014). worldcomp-proceedings.com
6. Thavappiragasam, M., Lushbough, C.M., Gnimpieba, E.Z.: Automatic biosystems comparison using semantic and name similarity. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB 2014, pp. 790–796 (2014)
7. Dräger, A., Rodriguez, N., Dumousseau, M., Dörr, A., Wrzodek, C., Le Novère, N., Zell, A., Hucka, M.: JSBML: a flexible Java library for working with SBML. Bioinformatics **27**, 2167–2168 (2011)