ssbio: A Python Framework for Structural Systems Biology

Nathan Mih[a], Elizabeth Brunk[b], Ke Chen[b], Edward Catoiu[b], Anand Sastry[b], Erol Kavvas[b], Jonathan M. Monk[b], Zhen Zhang[b], Bernhard O. Palsson[b]

* Correspondence should be addressed to: B.O.P. (palsson@ucsd.edu)

[a] Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, CA 92093

[b] Department of Bioengineering, University of California, San Diego, CA 92093

Abstract

Summary

Working with protein structures at the genome-scale has been challenging in a variety of ways. Here, we present *ssbio*, a Python package that provides a framework to easily work with structural information in the context of genome-scale network reconstructions, which can contain thousands of individual proteins. The *ssbio* package provides an automated pipeline to construct high quality genome-scale models with protein structures (GEM-PROs), wrappers to popular third-party programs to compute associated protein properties, and methods to visualize and annotate structures directly in Jupyter notebooks, thus lowering the barrier of linking 3D structural data with established systems workflows.

Availability and Implementation

ssbio is implemented in Python and available to download under the MIT license at http://github.com/SBRG/ssbio. Documentation and Jupyter notebook tutorials are available at http://ssbio.readthedocs.io/en/latest/. Interactive notebooks can be launched using Binder at https://mybinder.org/v2/gh/SBRG/ssbio/master?filepath=Binder.ipynb.

Contact

nmih@ucsd.edu

Supplementary Information

Supplementary data are available at *Bioinformatics* online.

Introduction

Merging the disciplines of structural and systems biology remains promising in a variety of ways, but differences in the fields present a learning curve for those looking toward this integration within their own research. Beltrao et al. stated it best, that “apparently structural biology and systems biology look like two different universes” (Beltrao et al. 2007). A great number of software tools exist within the structural bioinformatics community (Biasini et al. 2010; Grünberg et al. 2007; Gu & Bourne 2009; Hamelryck & Manderick 2003; O’Donoghue et al. 2015), and with recent advances in structure determination techniques, the number of experimental structures in the Protein Data Bank (PDB) continues to steadily rise (Mizianty et al. 2014). The challenges of integrating external data and software tools into systems analyses have been detailed (Ghosh et al. 2011), and structural information is no exception to the norm. At the systems-level, curated network models such as genome-scale metabolic models (GEMs) provide a context for molecular interactions in a functional cell (O’Brien et al. 2015). Recently, GEMs integrated with protein structures (GEM-PROs) have extended these models to explicitly utilize 3D structural data alongside modeling methods to substantiate a number of hypotheses, as we explain below.

Here, we present *ssbio*, a Python package designed with the goal of lowering the learning curve associated with efforts in structural systems biology. *ssbio* directly integrates with and builds upon the COBRApy toolkit (Ebrahim et al. 2013) allowing for seamless integration with existing GEMs. The core functionality of *ssbio* is additionally extended by hooks to many popular third-party structural bioinformatics algorithms, such as DSSP, MSMS, SCRATCH, I-TASSER, and others (see Supplementary Tables S1 and S2 for a full list) (Cheng et al. 2005; Kabsch & Sander 1983; Roy et al. 2010; Sanner et al. 1996).

Functionality

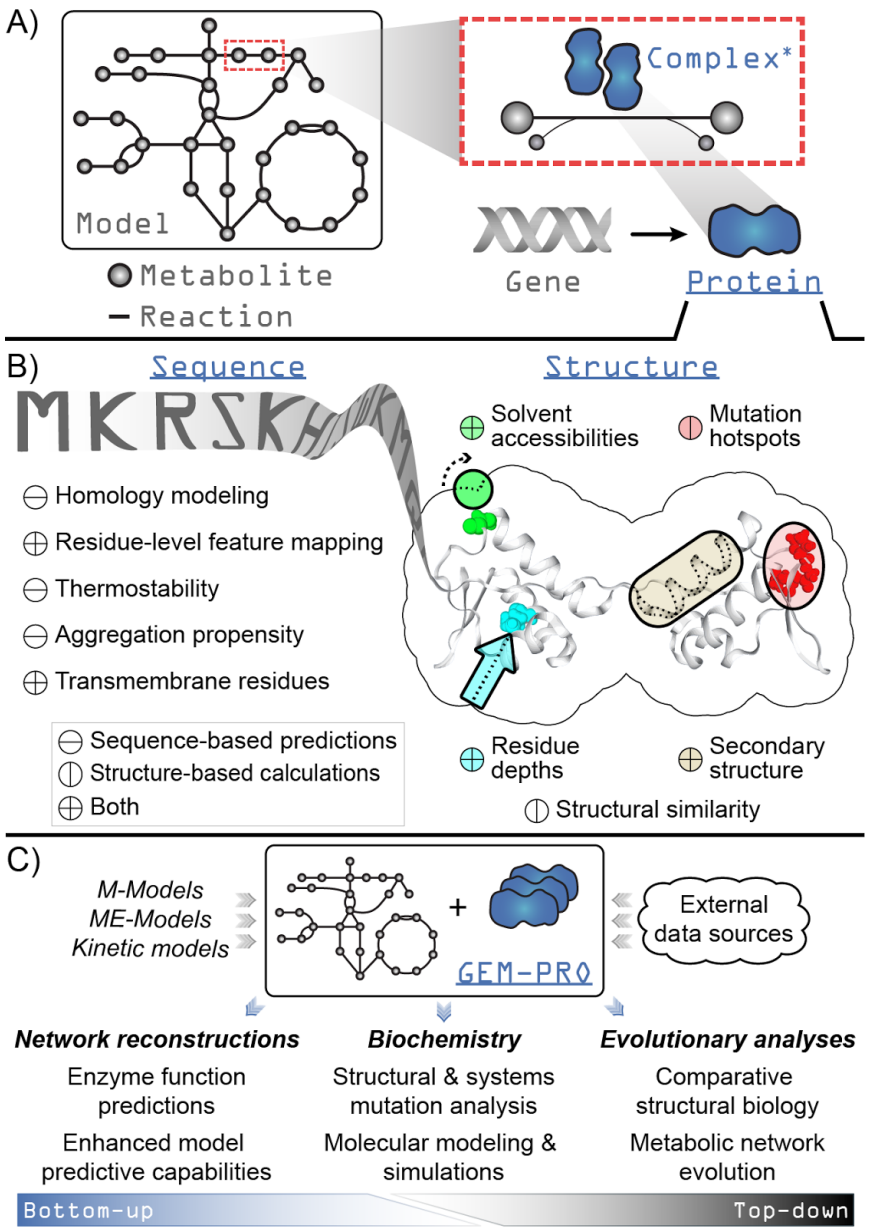

Fig. 1. Overview of the design and functionality of *ssbio.* Underlined fixed-width text in blue indicates added functionality to COBRApy for a genome-scale model loaded using *ssbio*. A) A simplified schematic showing the addition of a `Protein` to the core objects of COBRApy (fixed-width text in gray). A gene is directly associated with a protein, which can act as a monomeric enzyme or form an active complex with itself or other proteins (the asterisk denotes that methods for complexes are currently under development). B) Summary of functions available for computing properties on a protein sequence or structure. C) Uses of a GEM-PRO, from the bottom-up and the top-down. Once all protein sequences and structures are mapped to a genome-scale model, the resulting GEM-PRO has uses in multiple areas of study, as noted in the main text.

Protein class

ssbio adds a `Protein` class as an attribute to a COBRApy `Gene` and is representative of the gene's translated polypeptide chain (Fig. 1A). A `Protein` holds related amino acid sequences and structures, and a single representative sequence and structure can be set from these. This simplifies network analyses by enabling the properties of all or a subset of proteins to be computed and subsequently queried for. For details on these properties, as well as installation and execution instructions for the third-party software used to compute them,

please see the documentation. Additionally, proteins with multiple structures available in the PDB can be subjected to QC/QA based on set cutoffs such as sequence coverage and X-ray resolution. Proteins with no structures available can be prepared for homology modeling through platforms such as I-TASSER (Roy et al. 2010). Biopython representations of sequences (`SeqRecord` objects) and structures (`Structure` objects) are utilized to allow access to analysis functions available for their respective objects (Fig. 1B) (Cock et al. 2009). Finally, all information contained in a `Protein` (or in the context of a network model, multiple proteins) can be saved and shared as a JavaScript Object Notation (JSON) file.

GEM-PRO pipeline

The objectives of the GEM-PRO pipeline have previously been detailed (Brunk et al. 2016). A GEM-PRO directly integrates structural information within a curated GEM (Fig. 1C), and streamlines identifier mapping, representative object selection, and property calculation for a set of proteins. The pipeline provided in *ssbio* functions with an input of a GEM, but alternatively works with a list of gene identifiers or their protein sequences if network information is unavailable.

The added context of manually curated network interactions to protein structures enables different scales of analyses. For instance, from the top-down, global non-variant properties of protein structures such as the distribution of fold types can be compared within or between organisms (Brunk et al. 2016; Monk et al. 2017; Zhang et al. 2009). From the bottom-up, structural properties predicted from sequence or calculated from structure can be utilized to guide a metabolic reconstruction (Broddrick et al. 2016) or to enhance model predictive capabilities (Chang et al. 2010, 2013; Chen et al. 2017; Mih et al. 2016). Looking forward, applications to multi-strain modelling techniques (Bosi et al. 2016; Monk et al. 2016; Ong et al. 2014) would allow strain-specific changes to be investigated at the molecular level, potentially explaining phenotypic differences or strain adaptations to certain environments.

Scientific analysis environment

We provide a number of Jupyter notebook tutorials to demonstrate analyses at different scales (i.e. for a single protein sequence or structure, set of proteins, or network model). These notebooks can be launched in a virtual environment through the Binder project (https://mybinder.org/), with most third-party software pre-installed so users can immediately run through tutorials and experiment with them. Certain data can be represented as Pandas DataFrames (McKinney 2012), enabling quick data manipulation and graphical visualization. These notebooks are further extended by visualization tools such as NGLview for interacting with and annotating 3D structures (Nguyen et al. 2017; Rose & Hildebrand 2015), and Escher for constructing and viewing biological pathways (King et al. 2015) (Supplementary Figure S1). Module organization and directory organization for cached files is further described in the Supplementary Text.

Conclusion

ssbio provides a Python framework for systems biologists to start thinking about detailed molecular interactions and how they impact their models, and enables structural biologists to scale up and apply their expertise to multiple enzymes working together in a system. Towards a vision of whole-cell *in silico* models, structural information provides invaluable molecular-level details, and integration remains crucial.

Funding

This work was supported by the Novo Nordisk Foundation Center for Biosustainability [NNF10CC1016517 to N.M., K.C., E.C., and A.S.]; the Swiss National Science Foundation [p2elp2_148961 to E.B.]; and the National Institute of General Medical Sciences of the National Institutes of Health [U01-GM102098 to B.O.P., 1-U01-AI124316-01 to J.M.M. and E.K.].

Acknowledgements

We would like to thank Dr. Zachary King, Patrick Phaneuf, Marta Matos, and Colton Lloyd for valuable discussions in software development, and Dr. Laurence Yang, Yara Seif, JC Lachance, and Jared Broddrick for insight into desired functionalities of the package. We would also like to thank Marc Abrams for proofreading of the manuscript.

References

Beltrao, P., Kiel, C., & Serrano, L. (2007). ‘Structures in systems biology’, *Current opinion in structural biology*, 17/3: 378–84. Elsevier.

Biasini, M., Mariani, V., Haas, J., Scheuber, S., Schenk, A. D., Schwede, T., & Philippsen, A. (2010). ‘OpenStructure: a flexible software framework for computational structural biology’, *Bioinformatics* , 26/20: 2626–8. DOI: 10.1093/bioinformatics/btq481

Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V., & Palsson, B. Ø. (2016). ‘Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity’, *Proceedings of the National Academy of Sciences of the United States of America*, 113/26: E3801–9. DOI: 10.1073/pnas.1523199113

Broddrick, J. T., Rubin, B. E., Welkie, D. G., Du, N., Mih, N., Diamond, S., Lee, J. J., et al. (2016). ‘Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis’, *Proceedings of the National Academy of Sciences of the United States of America*, 113/51: E8344–53. DOI: 10.1073/pnas.1613446113

Brunk, E., Mih, N., Monk, J., Zhang, Z., O’Brien, E. J., Bliven, S. E., Chen, K., et al. (2016). ‘Systems biology of the structural proteome’, *BMC systems biology*, 10/1: 26. bmcsystbiol.biomedcentral.com. DOI: 10.1186/s12918-016-0271-6

Chang, R. L., Andrews, K., Kim, D., Li, Z., Godzik, A., & Palsson, B. O. (2013). ‘Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli’, *Science*, 340/6137: 1220–3. American Association for the Advancement of Science.

Chang, R. L., Xie, L., Xie, L., Bourne, P. E., & Palsson, B. Ø. (2010). ‘Drug off-target effects predicted using structural analysis in the context of a metabolic network model’, *PLoS computational biology*, 6/9: e1000938. Public Library of Science.

Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). ‘SCRATCH: a protein structure and structural feature prediction server’, *Nucleic acids research*, 33/Web Server issue: W72–6. DOI: 10.1093/nar/gki396

Chen, K., Gao, Y., Mih, N., O’Brien, E. J., Yang, L., & Palsson, B. O. (2017). ‘Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation’, *Proceedings of the National Academy of Sciences*, 114/43: 11548–53. DOI: 10.1073/pnas.1705524114

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., et al. (2009). ‘Biopython: freely available Python tools for computational molecular biology and bioinformatics’, *Bioinformatics* , 25/11: 1422–3. Oxford Univ Press.

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). ‘COBRApy: COnstraints-Based Reconstruction and Analysis for Python’, *BMC systems biology*, 7: 74. DOI: 10.1186/1752-0509-7-74

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y., & Kitano, H. (2011). ‘Software for systems biology: from tools to integrated platforms’, *Nature reviews. Genetics*, 12/12: 821–32. DOI: 10.1038/nrg3096

Grünberg, R., Nilges, M., & Leckner, J. (2007). ‘Biskit--a software platform for structural bioinformatics’, *Bioinformatics* , 23/6: 769–70. DOI: 10.1093/bioinformatics/btl655

Gu, J., & Bourne, P. E. (2009). *Structural Bioinformatics*. John Wiley & Sons.

Hamelryck, T., & Manderick, B. (2003). ‘PDB file parser and structure class implemented in Python’, *Bioinformatics* , 19/17: 2308–10.

Kabsch, W., & Sander, C. (1983). ‘DSSP: definition of secondary structure of proteins given a set of 3D coordinates’, *Biopolymers*, 22: 2577–637.

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., & Palsson, B. O. (2015). ‘Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways’, *PLoS computational biology*, 11/8: e1004321. DOI: 10.1371/journal.pcbi.1004321

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. ‘ O’Reilly Media, Inc.’

Mih, N., Brunk, E., Bordbar, A., & Palsson, B. O. (2016). ‘A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism’, *PLoS computational biology*, 12/7: e1005039. DOI: 10.1371/journal.pcbi.1005039

Mizianty, M. J., Fan, X., Yan, J., Chalmers, E., Woloschuk, C., Joachimiak, A., & Kurgan, L. (2014). ‘Covering complete proteomes with X-ray structures: a current snapshot’, *Acta crystallographica. Section D, Biological crystallography*, 70/Pt 11: 2781–93. DOI: 10.1107/S1399004714019427

Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., Herrgård, M. J., et al. (2016). ‘Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes’, *Cell systems*, 3/3: 238–51.e12. DOI: 10.1016/j.cels.2016.08.013

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., et al. (2017). ‘iML1515, a knowledgebase that computes Escherichia coli traits’, *Nature biotechnology*, 35/10: 904–8. DOI: 10.1038/nbt.3956

Nguyen, H., Case, D. A., & Rose, A. S. (2017). ‘NGLview - Interactive molecular graphics for Jupyter notebooks’, *Bioinformatics* . DOI: 10.1093/bioinformatics/btx789

O’Brien, E. J., Monk, J. M., & Palsson, B. O. (2015). ‘Using Genome-scale Models to Predict Biological Capabilities’, *Cell*, 161/5: 971–87. DOI: 10.1016/j.cell.2015.05.019

O’Donoghue, S. I., Sabir, K. S., Kalemanov, M., Stolte, C., Wellmann, B., Ho, V., Roos, M., et al. (2015). ‘Aquaria: simplifying discovery and insight from protein structures’, *Nature methods*, 12/2: 98–9. DOI: 10.1038/nmeth.3258

Ong, W. K., Vu, T. T., Lovendahl, K. N., Llull, J. M., Serres, M. H., Romine, M. F., & Reed, J. L. (2014). ‘Comparisons of Shewanella strains based on genome annotations, modeling, and experiments’, *BMC systems biology*, 8: 31. DOI: 10.1186/1752-0509-8-31

Rose, A. S., & Hildebrand, P. W. (2015). ‘NGL Viewer: a web application for molecular visualization’, *Nucleic acids research*, 43/W1: W576–9. DOI: 10.1093/nar/gkv402

Roy, A., Kucukural, A., & Zhang, Y. (2010). ‘I-TASSER: a unified platform for automated protein structure and function prediction’, *Nature protocols*, 5/4: 725–38. Nature Publishing Group.

Sanner, M. F., Olson, A. J., & Spehner, J.-C. (1996). ‘Reduced surface: an efficient way to compute molecular surfaces’, *Biopolymers*, 38/3: 305–20. Wiley Online Library.

Zhang, Y., Thiele, I., Weekes, D., Li, Z., Jaroszewski, L., Ginalski, K., Deacon, A. M., et al. (2009). ‘Three-dimensional structural view of the central metabolic network of Thermotoga maritima’, *Science*, 325/5947: 1544–9. American Association for the Advancement of Science.