ACS Synthetic Biology

pubs.acs.org/synthbio
Technical Note

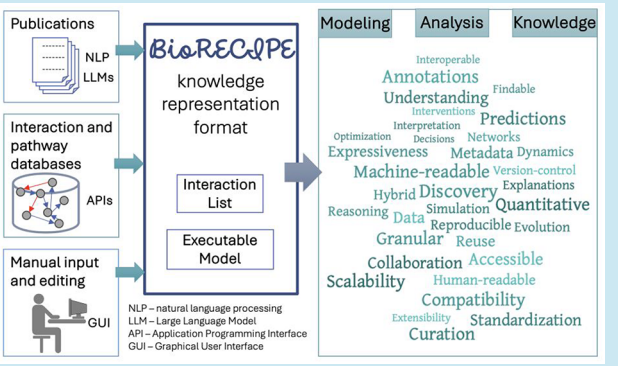
The BioRECIPE Knowledge Representation Format

Published as part of ACS Synthetic Biology virtual special issue "IWBDA 2023".

Emilee Holtzapple, Gaoxiang Zhou, Haomiao Luo, Difei Tang, Niloofar Arazkhani, Casey Hansen, Cheryl A. Telmer, and Natasa Miskov-Zivanov*

Cite This: ACS Synth. Biol. 2024, 13, 2621−2624
Read Online

ABSTRACT: The BioRECIPE (Biological system Representation for Evaluation, Curation, Interoperability, Preserving, and Execution) knowledge representation format was introduced to standardize and facilitate human−machine interaction while creating, verifying, evaluating, curating, and expanding executable models of intra- and intercellular signaling. This format allows a human user to easily preview and modify any model component, while it is at the same time readable by machines and can be processed by a suite of model development and analysis tools. The BioRECIPE format is compatible with multiple representation formats, natural language processing tools, modeling tools, and databases that are used by the systems and synthetic biology communities.



KEYWORDS: *modeling, representation format, FAIR principles, signaling pathways, networks, automation*

INTRODUCTION

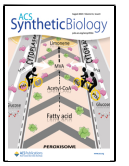
Systems biology and synthetic biology benefit from collaborations between biologists, computer scientists and engineers, therefore an easily readable, standardized representation of the complex events of cell signaling and gene regulatory networks (GRNs) (Figure 1A) is needed for sharing of information. A standardized format creates consistency, accuracy, and reproducibility, and makes models findable, accessible, interoperable, and reusable (FAIR principles).[1]

One common standard representation format is the Systems Biology Markup Language (SBML),[2] which is based on XML and is therefore machine-readable. SBML uses modules to represent various components within a biological system such as reactions, species, and compartments, and it supports analysis via ordinary differential equations, stochastic simulation algorithm,[3] or the reaction rule-based approach (e.g., BioNetGen).[4] SBML format also allows for a significant amount of user annotation providing standardized representation of starting conditions, context, metadata, and literature sources. Cell Markup Language (CellML)[5] is another XML-based standard representation format and a modeling framework used for executable models. These two formats differ in scope- CellML is ideal for detailed models of molecular interactions as it requires kinetic parameters for each interaction, while SBML is more suitable for modeling cell signaling pathways and networks. However, both SBML and CellML are not easily interpreted by life scientists without previous exposure to XML.

The Biological Pathway Exchange (BioPAX)[6] provides another standardized format to represent molecular interactions in a signaling pathway. BioPAX supports three levels of representation, with each level offering increased complexity and detail. A simpler representation format, the Biological Expression Language (BEL),[7] represents causal, correlative, and associative relationships between biological entities as a triplet statement. Similarly, the INDRA database[8] represents causal relationships between entities in biomedical literature as statements with more detail than BEL. Some interaction and pathway databases may use their own representation format, for example, the KEGG Markup Language (KGML)[9] was created for storing and standardizing models in the KEGG database.

Model representation formats rely on pre-existing ontologies to standardize individual biological entities and represent biological models. OBO (Open Biological and Biomedical Ontologies) is widely used to represent structured ontologies and controlled vocabularies, including the Gene Ontology (GO) Resource.[10] It is human-readable and allows for the definition of classes, properties, and relationships between

Received: February 12, 2024
Revised: May 22, 2024
Accepted: July 9, 2024
Published: July 25, 2024



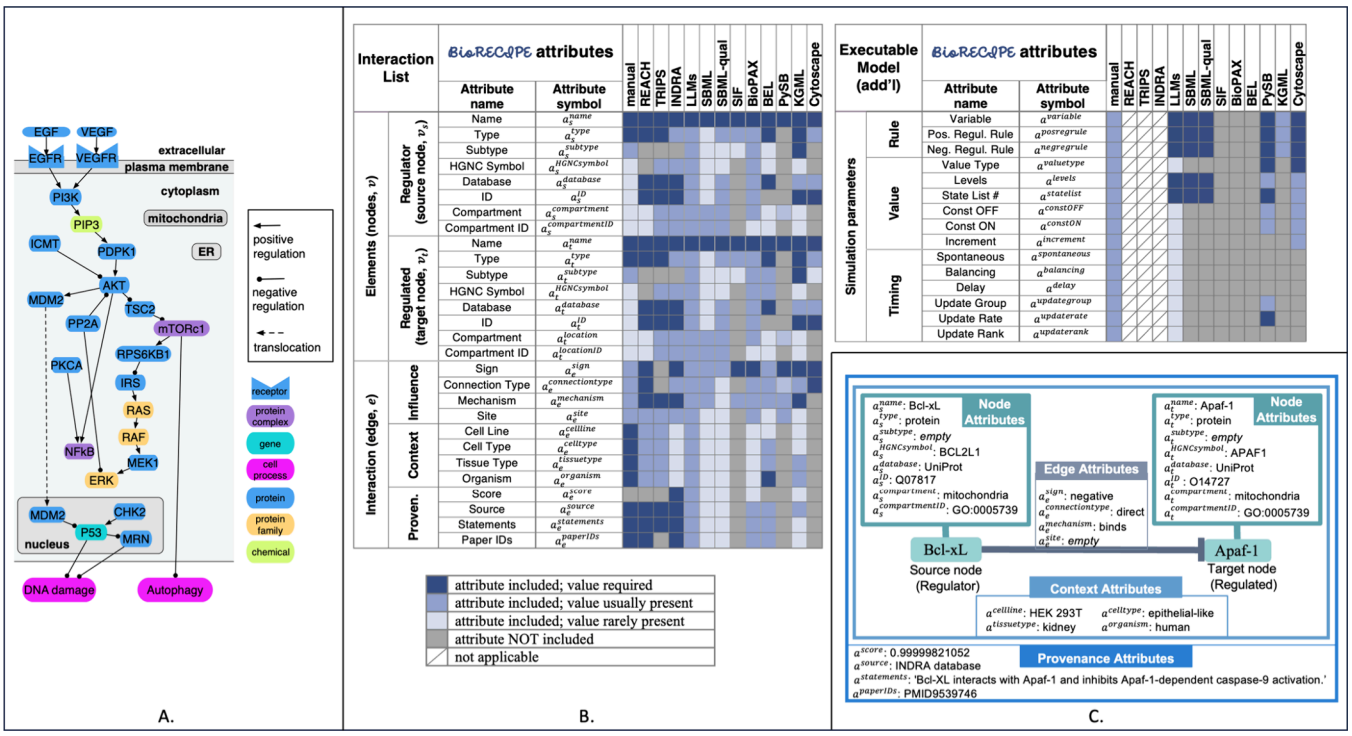


Figure 1. Examples: (A) Pathways, cell compartments, element types, and interactions that can be represented with BioRECIPE. (B) The list of all attributes used by BioRECIPE in Interaction List and Executable Model formats and a summary of whether these attributes are included and required by other formats and tools. (C) An example interaction and interaction attributes that are included in the BioRECIPE's Interaction List format.

BioRECIPE feature	Description
Standardization	The documentation is available, and definitions provided to standardize the representation of biological entities and events, ensuring consistency and interoperability across different tools and studies.
Expressiveness	Captures a wide range of biological phenomena, from molecular interactions to cellular processes.
Scalability	Represents systems of varying complexity, from simple pathways to whole-cell models.
Extensibility	Additional attributes, new types of data and relationships can be added as our understanding of biological systems evolves.
Machine-readable	Can be processed by computer programs, facilitating automated analysis and simulation.
Human-readable	The tabular spreadsheet format and the underlying graph structure are readable, visualizable, and understandable by humans.
Support for quantitative data	Incorporates a range of quantitative parameters essential for dynamic simulations.
Annotations and metadata	It allows for rich annotations and metadata, linking model attributes to biological databases and literature, which is crucial for model validation and reuse.
Compatibility	Interoperable with a range of tools and platforms, enabling seamless data exchange and integration.
Facilitates collaboration	Provides a common format for computational modelers and biologists, enabling collaboration and sharing of models.
Version control and evolution	The documentation and translators are open access and available on ReadtheDocs and GitHub enabling version control and updating.

BioRECIPE use		Description
MODELING	**Hybrid models**	Elements within the same model can have different types of update functions, e.g., Boolean, discrete, qualitative or weighted sums.
	Granular computing	Complex entities that emerge through data abstraction and the process of extracting knowledge from information or data is referred to as "information granules." These entities are clustered based on factors such as similarity, functional or physical proximity, indistinguishability, and coherence
	Networks	Representation of models that have an underlying network (graph) structure, including both undirected or directed graphs.
	Dynamics	Simulation parameters are included in the format and are used to setup scenarios and initialize and run simulations.
ANALYSIS	**Simulation**	It is compatible with the DiSH simulator which is a versatile simulator with deterministic and stochastic simulation schemes.
	Explanations	As it is readable by both machines and humans, and it includes the metadata, the models and interactions can be explored both manually and automatically. The information about interactions and metadata is available in a standardized format, and the compatibility with the simulator and a suite of other tools that provide insights into the behavior of all model elements under a range of scenarios, altogether (a) improve *understanding* of the model and the modeled system and allow for (b) *explanations* and *interpretations* of observed behaviors, (c) *predictions* of future behaviors, as well as (d) explorations of *interventions* informing *decisions*, (e) leading to *discovery* and (f) guiding and *optimizing* future experiments.
	Understanding	
	Interpretation	
	Predictions	
	Interventions	
	Discovery	
	Decisions	
	Reasoning	
	Optimization	
KNOWLEDGE	**Curation**	Compatibility with tools that filter information from literature, verify and validate models enables automated curation.
	Reproducibility	Detailed annotation and metadata in models enable reproducibility of both experimental data and model results.
	Findability	All related files are provided on GitHub, making the code and the examples easy to find by both humans and computers. All components of the format are defined and described in detail in the documentation on ReadtheDocs.
	Accessibility	All relevant files are open access.
	Interoperability	The format and the included translators facilitate interoperability of models with a range of tools or workflows for analysis, storage, and processing.
	Reusability	The detailed documentation enables reuse of published data through standardized models and replication of model results within same or different settings.

Figure 2. Description of BioRECIPE features and types of models that can be represented with BioRECIPE, model analysis that can be conducted on these models, and the descriptions of how BioRECIPE satisfies the FAIR principles.

terms, making it suitable for standardizing biological models. While not strictly reserved for biological modeling, Web Ontology Language (OWL) is a more powerful and expressive language for creating ontologies. It uses formal logic to create semantic networks and is particularly useful for capturing

https://doi.org/10.1021/acssynbio.4c00096
ACS Synth. Biol. 2024, 13, 2621−2624

complex relationships and reasoning in biological models in BioPAX format.

While powerful and flexible for machines, existing formats are difficult for humans to interact with. There is still a need for a standardized format that may be used for both static and executable models, is both human and machine-readable, can incorporate diverse annotations as well as data, and has translators for seamless conversion into other commonly used representation formats. To address this need, here we present the BioRECIPE representation format, which is interoperable with existing interaction, pathway, and model representation formats, and while it utilizes standardized ontologies, it is not dependent on any one ontology.

RESULTS

BioRECIPE is a tabular representation format, typically written in a spreadsheet file type, which is simple for humans to use, facilitates knowledge sharing and aligns with FAIR principles (Figure 2). Spreadsheets have been used in life science domains as inputs for plasmid annotation packages,[11] metabolic network modeling,[12] and in synthetic biology.[13] The BioRECIPE format can be used by both computational modelers and biology experts to create and modify Interaction List and Executable Model files. It also has a formal structure that can be read, created, updated, and output by computer programs. The detailed BioRECIPE documentation is available as ReadtheDocs pages,[14] with instructions for creating Interaction Lists and Executable Models in this format.

In BioRECIPE, interactions are represented using the *event-based* Interaction List spreadsheet format, where each biological event is assigned one row, and the columns correspond to attributes of the event participants and the interaction between them. The attributes are selected and organized to allow for detailed curation and an extendable representation of interactions. An example biological interaction, represented as a directed signed edge between two nodes, including node, edge, context, and provenance attributes is illustrated in Figure 1C. More examples are included in the Supporting Information, ReadtheDocs documentation,[14] and GitHub repository.[15]

The BioRECIPE format also provides representation of the static graph structure of models and attributes necessary to study the dynamics, usually through simulations. These executable models are represented in the BioRECIPE format using the *element-based* approach where each element in a model is assigned a row in the model spreadsheet, combining multiple interactions in which the element is the target of the influence. Element update functions are written using a simple notation in the BioRECIPE format, which supports discrete functions (min, max, weighted sum, Boolean), spontaneous increase and decrease of elements, different element update rates and types of regulation (positive, negative, highest level, weighted) and can be used under deterministic or stochastic simulation approaches.

Interactions and models written in the BioRECIPE format can be used by modelers, curators, and tool developers, and with a range of different tools that filter and classify interactions and automatically assemble and analyze models. The tools either use BioRECIPE directly or by translating interactions and models to and from other formats (Supporting Information). Interactions can be converted to the BioRECIPE format from the output of natural language processing tools or from interaction and pathway databases. Models can be converted from other representation formats and model databases. The BioRECIPE GitHub repository includes translators with instructions how to run them, examples of input and output files, and interactive Jupyter notebooks to guide the use of the BioRECIPE format and translators.[15]

CONCLUSIONS

The BioRECIPE representation format is a valuable tool for systems and synthetic biology that enables comprehensive model curation by both humans and machines. Cellular signaling pathways and GRNs use the same components, in different combinations that are context specific, and therefore, interaction details are crucial to modeling accuracy. Automated readers are improving, however often fail to capture these details. BioRECIPE allows for all key element and interaction attributes to be included, as well as attributes for simulation, making it compatible with many existing interaction, pathway, and model databases, and with a range of tools for extraction of interaction information, model curation, simulation, and analysis. This interoperability ensures that researchers can seamlessly integrate BioRECIPE into their existing workflows. Future directions include additional functionality to translate between other existing formats (e.g., SBOL[16]), or integration of the BioRECIPE representation format as input to model curation and storage platforms (e.g., CellCollective,[17] NDEx[18]). These platforms play a crucial role in managing and disseminating computational models, and the integration with BioRECIPE could further streamline the process of model sharing and collaboration within the scientific community.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.4c00096.

> Links to BioRECIPE resources (ReadtheDocs, GitHub folders, Jupyter notebooks) and example containing attribute values extracted from sentences and a summary of standard database tools, formats, and translators for BioRECIPE format translation and conversion (PDF)

AUTHOR INFORMATION

Corresponding Author

Natasa Miskov-Zivanov − *Electrical and Computer Engineering Department, Computational and Systems Biology Department, and Bioengineering Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States;* orcid.org/0000-0003-1110-3403; Email: nmzivanov@pitt.edu

Authors

Emilee Holtzapple − *Electrical and Computer Engineering Department and Computational and Systems Biology Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States;* orcid.org/0000-0003-3907-1597

Gaoxiang Zhou − *Electrical and Computer Engineering Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States;* orcid.org/0009-0003-1161-2272

https://doi.org/10.1021/acssynbio.4c00096
ACS Synth. Biol. 2024, 13, 2621−2624

Haomiao Luo − *Electrical and Computer Engineering Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States*
Difei Tang − *Electrical and Computer Engineering Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States*
Niloofar Arazkhani − *Electrical and Computer Engineering Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States*
Casey Hansen − *Bioengineering Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States*
Cheryl A. Telmer − *Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.4c00096

Author Contributions

Co-first authors: E.H. and G.Z. Study conception and design: E.H., G.Z., C.A.T., and N.M.Z. Data collection and case studies: E.H., G.Z., H.L., D.T., N.A., C.H., C.A.T., and N.M.Z. GitHub repository, translator development and testing, Jupyter notebooks, and ReadtheDocs: E.H., G.Z., H.L., D.T., N.A., C.A.T., and N.M.Z. Draft manuscript preparation: E.H., H.L., D.T., G.Z., N.A., C.H., C.A.T., and N.M.Z. Manuscript revisions: G.Z., C.A.T., and N.M.Z. All authors reviewed the results and approved the final version of the manuscript.

Funding

This work was funded in part by a DARPA Big Mechanism award AIMCancer (W911NF-17-1-0135) and in part by the NSF EAGER award CCF-2324742.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. W.; da Silva Santos, L. B.; Bourne, P. E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
(2) Keating, S. M.; Waltemath, D.; König, M.; Zhang, F.; Dräger, A.; Chaouiya, C.; Bergmann, F. T.; Finney, A.; Gillespie, C. S.; Helikar, T.; et al. SBML Level 3: an extensible format for the exchange and reuse of biological models. *Molecular Systems Biology* **2020**, *16* (8), No. e9110.
(3) Gillespie, D. T. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **2007**, *58*, 35−55.
(4) Harris, L. A.; Hogg, J. S.; Tapia, J.-J.; Sekar, J. A. P.; Gupta, S.; Korsunsky, I.; Arora, A.; Barua, D.; Sheehan, R. P.; Faeder, J. R. BioNetGen 2.2: advances in rule-based modeling. *Bioinformatics* **2016**, *32* (21), 3366−3368.
(5) Clerx, M.; Cooling, M. T.; Cooper, J.; Garny, A.; Moyle, K.; Nickerson, D. P.; Nielsen, P. M. F.; Sorby, H. CellML 2.0. *J. Integr Bioinform* **2020**, *17*, 2−3.
(6) Demir, E.; Cary, M. P.; Paley, S.; Fukuda, K.; Lemer, C.; Vastrik, I.; Wu, G.; D'Eustachio, P.; Schaefer, C.; Luciano, J.; et al. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* **2010**, *28* (9), 935−942.
(7) Slater, T. Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today* **2014**, *19* (2), 193−198.
(8) Grefenstette, J.; Brown, S.; Rosenfeld, R.; DePasse, J.; Stone, N.; Cooley, P.; Wheaton, W.; Fyshe, A.; Galloway, D.; Sriram, A. FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health* **2013**, *13*, 940 DOI: 10.1186/1471-2458-13-940.
(9) KEGG Markup Language. *The Kyoto Encylopedia of Genes and Genomes*, 2023. https://www.kegg.jp/kegg/xml/docs/.
(10) The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **2017**, *45* (D1), D331−D338.
(11) Zieliński, T.; Hay, J.; Romanowski, A.; Nenninger, A.; McCormick, A.; Millar, A. J. SynBio2Easy—a biologist-friendly tool for batch operations on SBOL designs with Excel inputs. *Synthetic Biology* **2022**, *7* (1), No. ysac002, DOI: 10.1093/synbio/ysac002.
(12) Lubitz, T.; Hahn, J.; Bergmann, F. T.; Noor, E.; Klipp, E.; Liebermeister, W. SBtab: a flexible table format for data exchange in systems biology. *Bioinformatics* **2016**, *32* (16), 2559−2561.
(13) Mante, J.; Abam, J.; Samineni, S. P.; Pötzsch, I. M.; Beal, J.; Myers, C. J. Excel−SBOL Converter: Creating SBOL from Excel Templates and Vice Versa. *ACS Synth. Biol.* **2023**, *12* (1), 340−346.
(14) BioRECIPE documentation. *The MeLoDy Lab*, 2023. https://melody-biorecipe.readthedocs.io/en/latest/ (accessed May 22, 2024).
(15) BioRECIPE. *Github*, 2024. https://github.com/pitt-miskov-zivanov-lab/BioRECIPE/tree/main (accessed May 22, 2024).
(16) McLaughlin, J. A.; Beal, J.; Mısırlı, G.; Grünberg, R.; Bartley, B. A.; Scott-Brown, J.; Vaidyanathan, P.; Fontanarrosa, P.; Oberortner, E.; Wipat, A.; et al. The Synthetic Biology Open Language (SBOL) Version 3: Simplified Data Exchange for Bioengineering. *Front Bioeng Biotechnol* **2020**, *8*, 1009.
(17) Helikar, T.; Kowal, B.; McClenathan, S.; Bruckner, M.; Rowley, T.; Madrahimov, A.; Wicks, B.; Shrestha, M.; Limbu, K.; Rogers, J. A. The Cell Collective: Toward an open and collaborative approach to systems biology. *BMC Systems Biology* **2012**, *6* (1), 96.
(18) Pratt, D.; Chen, J.; Welker, D.; Rivas, R.; Pillich, R.; Rynkov, V.; Ono, K.; Miello, C.; Hicks, L.; Szalma, S.; et al. NDEx, the Network Data Exchange. *Cell Syst* **2015**, *1* (4), 302−305.