CALIFORNIA INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTING AND MATHEMATICAL SCIENCES

On behalf of the COMBINE Coordinators
Michael Hucka, Ph.D.
Beckman Institute 139-74
California Institute of Technology
Pasadena, CA 91125

December 5, 2014

National Institutes of Health
9000 Rockville Pike
Bethesda, Maryland 20892

Notice Number: NOT-ES-15-002

Dear NIH,

We are the coordinators of COMBINE (Computational Modeling in Biology Network; http://co.mbine.org), an initiative to coordinate the development of many popular standards and formats for computational modeling in biology. We are grateful for the opportunity that NIH is providing to respond to the *Request for Information (RFI) Making Data Usable – A Framework for Community-Based Data and Metadata Standards Efforts for NIH-relevant Research*. We would like to take this opportunity to summarize COMBINE's efforts and relate some of our experiences in the development of community-based standards in biology.

The history and goals of COMBINE

The availability of appropriate data formats and process descriptions is an essential enabler for reproducible science. Researchers must be able to build on each other's work to develop a deeper understanding of biological phenomena, but this task is greatly impeded if they do not use common languages to describe their work. In the past two decades, this has led to the development of several formats and minimum information guidelines to facilitate the exchange of data and models. However, the existence of uncoordinated standards risks creating silos that induce new interoperability problems.

COMBINE was formed in 2009 by the groups involved in developing file formats and other standards in systems biology, including SBML [12, 13], SBGN [17], BioPAX [7], CellML [5, 10], SED-ML [20], SBOL [8], NeuroML [9], and others. The initial impetus was the realization that many individuals were involved in multiple efforts, traveling to separate international workshops year after year and performing many of the same organizational tasks multiple times for each standards community. Eventually, two "super meetings" were held involving many of the groups, and slowly we realized that not only could there be cost savings in co-locating meetings: the various efforts could also benefit from common infrastructure, operating procedures, and potentially a common voice to seek support.

COMBINE's aim is thus to act as a coordinator, facilitator, and resource for different standardization efforts whose domains of use cover related areas of the computational biology space. We hope that COMBINE can help the federated projects develop standards that are more interoperable and with less overlap than if the efforts proceeded separately. An important point about COMBINE, however, is that it *does not dictate what individual standardization efforts should do*. Actions are entirely up to the leaders and members of the communities involved in the individual efforts. COMBINE does offer examples of what has worked in terms of community organization approaches, as well as some common infrastructure for such things as cataloguing standards specifications, but the degree of participation is up to the groups behind the efforts.

Groups involved in COMBINE today

COMBINE today includes the efforts listed in the following table. All are open community efforts, with freely available specifications, open community participation, etc. They cover a range of topics: raw data standards, model format standards, graphical notation standards, ontologies, and minimum information guidelines.

Category	Name	COMBINE page or other reference
COMBINE representation standards	BioPAX (*Biological Pathways Exchange*)	http://co.mbine.org/standards/biopax
	CellML	http://co.mbine.org/standards/cellml
	SBGN (*Systems Biology Graphical Notation*)	http://co.mbine.org/standards/sbgn
	SBML (*Systems Biology Markup Language*)	http://co.mbine.org/standards/sbml
	SBOL (*Synthetic Biology Open Language*)	http://co.mbine.org/standards/sbol
	SED-ML (*Simulation Experiment Description Markup Language*)	http://co.mbine.org/standards/sed-ml
Associated standardization efforts	BioModels.net Qualifiers	http://co.mbine.org/standards/qualifiers
	COMBINE Archive	http://co.mbine.org/standards/omex
	MIASE (*Minimum Information About a Simulation Experiment*)	http://co.mbine.org/standards/miase
	MIRIAM (*Minimal Information Required In the Annotation of Models*)	http://co.mbine.org/standards/miriam
	KiSAO (*Kinetic Simulation Algorithm Ontology*)	http://co.mbine.org/standards/kisao
Related standardization efforts	BioSharing	[19]
	CNO (*Computational Neuroscience Ontology*)	[15]
	FieldML (*Field Markup Language*)	[3]
	GPML (*GenMAPP Pathway Markup Language*)	[1]
	MAMO (*Mathematical Modeling Ontology*)	[22]
	NeuroML	[9]
	NuML (*Numerical Markup Language*)	[6]
	PSI-MI (*Proteomics Standards Initiative*)	[11]
	SpineML (*Spiking Neural Markup Language*)	[18]
	TEDDY (*TErminology for the Description of DYnamics*)	[4]

The differences in the categories are as follows:

- The *COMBINE representation standards* meet a number of basic criteria which include the following: (i) represent information in biology, (ii) possess democratically-elected editorial boards, (iii) possess full specifications of version 1.0 or higher, (iv) have API library implementations supporting the standard, and (v) have continued development supported by a unified group of identifiable people.
- The *Associated standardization efforts* are either in a more fledgling state of development, or are not standard formats per se but rather tools or services that facilitate the use or interoperability of the COMBINE representation standards.
- The *Related standardization efforts* are other efforts that are either candidate COMBINE standards in early stages of development, or else are mature efforts in their own right that have their own substantial communities and, while not part of COMBINE, are efforts that the COMBINE community is involved in.

A standardization community that wants to become part of COMBINE begins as a *Related standardization effort*. The developers of the standard should join the COMBINE announcement mailing lists and, especially, participate in the annual COMBINE meetings discussed below. If the effort meets the basic criteria to be considered a full-fledged standard, the *COMBINE Coordinators* will accept the effort as a COMBINE standard.

Activities Performed by COMBINE

COMBINE, as an organization, currently performs the following activities:

- *Organize meetings*: COMBINE organizes open meetings where interested people can gather for face-to-face discussions and work on the standards. The primary meetings are the annual COMBINE Forum and the annual HARMONY (HAckathon on Resources for MOdeliNg in biologY) workshop, held approximately six months apart. The joint meetings help the different standardization efforts work together; they also make financial sense by reducing the overall number of meetings, travel, and money spent on hosting meetings. (However, COMBINE does not currently have any funding of its own, and the meetings must be organized by groups that volunteer to host them.) The leaders of the various standards also endeavor to write meeting reports that summarize the outcomes of the meetings [e.g., 16, 21].

- *Help coordinate standards development*: Thanks in large part to the meetings that COMBINE organizes, the discussion forums it provides, and the involvement of many of the same people in multiple standardization efforts, COMBINE helps coordinate the activities of the different efforts. This reduces duplication of effort, user confusion, and non-interoperability among the efforts.

- *Identify missing standards and initiate efforts to develop them*: COMBINE's meta-community is in an ideal position to identify what is missing from the current constellation of standards in computational systems biology. This has already yielded benefits: we have recently developed the COMBINE archive, a format that fills the need for a simple, consistent way of bundling multiple files related to a modeling project [http://co.mbine.org/standards/omex; 2]; and we have also begun to identify missing minimal requirements for common annotations across the spectrum of data used in biological modeling, such as parameter identifiability (tentatively called the Minimal Information for Model Inference and Parametrisation—MIMIP) and mathematical classification (the Mathematical Modelling Ontology—MAMO).

- *Provide a specification infrastructure*: COMBINE provides a consistent framework for cataloguing the definitions of COMBINE standards. This framework includes a consistent, hierarchical identifier scheme for identifying standard specifications; a URI scheme for locating specifications and standards [using Identifiers.org to provide permanent, resolvable URIs for standards; 14]; and a web page structure for the description of each standard.

- *Develop common procedures*: Many standardization efforts are started by academics who have little experience with community organization. Effective organization is something that takes time and experience to learn. In COMBINE, we are documenting our experiences and collecting them into a collection of examples, recommendations and best practices (e.g., http://co.mbine.org/Documents/criteria). We hope to provide would-be standards developers with a set of off-the-shelf "standard operating procedures" for different situations and goals.

- *Organize tutorials*: Educating biologists about available standards and compatible software tools is another important activity pursued by COMBINE. We organize tutorials at the primary COMBINE meetings as well as at international conferences, in particular the annual *International Conference on Systems Biology* (ICSB).

- *Maintain collective online forums/groups*: COMBINE maintains mailing lists and online discussion forums (http://co.mbine.org/comm). A discussion list cover the topic of general interest for all COMBINE members, while dedicated lists cover specific issues such as the COMBINE archive, metadata, etc. General announces are done via social media (e.g., Twitter feed *@combine_coord*).

An additional activity that we hope to undertake soon is fund-raising. This will require COMBINE to become a legal entity that can accept funding. Once this is in place, we hope to be able to fund the meetings and online infrastructure, and perhaps also seek funding for further standards development.

Experiences and perspectives in community-based standards development

NEEDS WORK.

References

[1] GPML pathway format. http://developers.pathvisio.org/wiki/EverythingGpml, 2014.

[2] Frank T. Bergmann, Richard Adams, Stuart Moodie, Jonathan Cooper, Mihai Glont, Martin Golebiewski, Michael Hucka, Camille Laibe, Andrew K. Miller, David P. Nickerson, Brett G. Olivier, Nicolas Rodriguez, Herbert M. Sauro, Martin Scharm, Stian Soiland-Reyes, Dagmar Waltemath, Florent Yvon, and Nicolas Le Novère. One file to share them all: Using the COMBINE archive and the OMEX format to share all information about a modeling project. arXiv:1407.4992 [cs.DL], 2014.

[3] G. Richard Christie, Poul M. F. Nielsen, Shane A. Blackett, Chris P. Bradley, and Peter J. Hunter. FieldML: concepts and implementation. *Philosophical Transactions of the Royal Society of London A*, 367(1895):1869–1884, 2009.

[4] Mélanie Courtot, Nick Juty, Christian Knüpfer, Dagmar Waltemath, Anna Zhukova, Andreas Dräger, Michel Dumontier, Andrew Finney, Martin Golebiewski, Janna Hastings, Stefan Hoops, Sarah Keating, Douglas B. Kell, Samuel Kerrien, James Lawson, Allyson Lister, James Lu, Rainer Machne, Pedro Mendes, Matthew Pocock, Nicolas Rodriguez, Alice Villéger, Darren J. Wilkinson, Sarala Wimalaratne, Camille Laibe, Michael Hucka, and Nicolas Le Novère. Controlled vocabularies and semantics in systems biology. *Molecular Systems Biology*, 7(1):543, September 2011.

[5] Autumn A Cuellar, Catherine M Lloyd, Poul F Nielsen, David P Bullivant, David P Nickerson, and Peter J Hunter. An overview of CellML 1.1, a biological model description language. *Simulation*, 79(12):740–747, 2003.

[6] Joseph O. Dada, Irena. Spasic, Norman W. Paton, and Pedro Mendes. SBRML: a markup language for associating systems biology data with models. *Bioinformatics*, 26(7):932–938, 2010.

[7] Emek Demir, Michael P. Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming N. Wu, Peter D'Eustachio, Carl Schaefer, Joanne Luciano, Frank Schacherer, Irma Martinez-Flores, Zhenzjun J. Hu, Veronica Jimenez-Jacinto, Geeta Joshi-Tope, Kumaran Kandasamy, Alejandra C. Lopez-Fuentes, Huaiyu Y. Mi, Elgar Pichler, Igor Rodchenkov, Andrea Splendiani, Sasha Tkachev, Jeremy Zucker, Gopal Gopinath, Harsha Rajasimha, Ranjani Ramakrishnan, Imran Shah, Mustafa Syed, Nadia Anwar, Özgün O. Babur, Michael Blinov, Erik Brauner, Dan Corwin, Sylva Donaldson, Frank Gibbons, Robert Goldberg, Peter Hornbeck, Augustin Luna, Peter Murray-Rust, Eric Neumann, Oliver Reubenacker, Matthias Samwald, Martijn van Iersel, Sarala Wimalaratne, Keith Allen, Burk Braun, Michelle Whirl-Carrillo, Kei-Hoi H. Cheung, Kam Dahlquist, Andrew Finney, Marc Gillespie, Elizabeth Glass, Li Gong, Robin Haw, Michael Honig, Olivier Hubaut, David Kane, Shiva Krupa, Martina Kutmon, Julie Leonard, Debbie Marks, David Merberg, Victoria Petri, Alex Pico, Dean Ravenscroft, Liya Y. Ren, Nigam Shah, Margot Sunshine, Rebecca Tang, Ryan Whaley, Stan Letovksy, Kenneth H. Buetow, Andrey Rzhetsky, Vincent Schachter, Bruno S. Sobral, Ugur Dogrusoz, Shannon McWeeney, Mirit Aladjem, Ewan Birney, Julio Collado-Vides, Susumu Goto, Michael Hucka, Nicolas Le Novère, Natalia Maltsev, Akhilesh Pandey, Paul Thomas, Edgar Wingender, Peter D. Karp, Chris Sander, and Gary D. Bader. The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942, 2010.

[8] Michal Galdzicki, Kevin P. Clancy, Ernst Oberortner, Matthew Pocock, Jacqueline Y. Quinn, Cesar A. Rodriguez, Nicholas Roehner, Mandy L. Wilson, Laura Adam, J. Christopher Anderson, Bryan A. Bartley, Jacob Beal, Deepak Chandran, Joanna Chen, Douglas Densmore, Drew Endy, Raik Grünberg, Jennifer Hallinan, Nathan J. Hillson, Jeffrey D. Johnson, Allan Kuchinsky, Matthew Lux, Goksel Misirli, Jean Peccoud, Hector A. Plahar, Evren Sirin, B. Stan, Guy-Bart, Alan Villalobos, Anil Wipat, John H. Gennari, Chris J. Myers, and Herbert M. Sauro. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nature Biotechnology*, 32(6):545, 2014.

[9] Padraig Gleeson, Sharon Crook, Robert C. Cannon, Michael L. Hines, Guy O. Billings, Matteo Farinella, Thomas M. Morse, Andrew P. Davison, Subhasis Ray, Upinder S. Bhalla, Simon R. Barnes, Yoana D. Dimitrova,

and R. Angus Silver. NeuroML: a language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS Computational Biology*, 6(6):e1000815, 2010.

[10] Warren J. Hedley, Melanie R. Nelson, D. P. Bullivant, and Poul F. Nielson. A short introduction to CellML. *Philosophical Transactions of the Royal Society of London A*, 359:1073–1089, 2001.

[11] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, R. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, YX X. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, SGN G. N. Grant, C. Sander, P. Bork, WM M. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, L. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler. The HUPO PSI's Molecular Interaction format—a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183, 2004.

[12] Michael Hucka, Frank T. Bergmann, Stephan Hoops, Sarah M. Keating, Sven Sahle, James C. Schaff, and Lucian P. Smith. The Systems Biology Markup Language (SBML): Language specification for Level 3 Version 1 Core. Available via the World Wide Web at http://sbml.org/Documents/Specifications, 2010.

[13] Michael Hucka, Andrew Finney, Herbert M. Sauro, Hamid Bolouri, John C. Doyle, Hiroaki Kitano, Adam P. Arkin, Benjamin J. Bornstein, Dennis Bray, Athel Cornish-Bowden, Autumn A. Cuellar, Sergey Dronov, Ernst-Dieter Gilles, Martin Ginkel, Victoria Gor, Igor I. Goryanin, Warren J. Hedley, T. Charles Hodgman, Jan-Hendrik. Hofmeyr, Peter J. Hunter, Nick S. Juty, J. L. Kasberger, Andreas Kremling, Ursula Kummer, Nicolas Le Novère, Leslie M. Loew, Daniel Lucio, Pedro Mendes, Eric D. Mjolsness, Y. Nakayama, Melanie R. Nelson, Poul F. Nielsen, Takeshi Sakurada, James C. Schaff, Bruce E. Shapiro, Thomas S. Shimizu, Hugh D. Spence, Jörg Stelling, Koichi Takahashi, Masaru Tomita, John Wagner, and J. Wang. The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[14] Nick Juty, Nicolas Le Novère, and Camille Laibe. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Research*, 40(Database issue):D580–D586, Jan 2012.

[15] Yann Le Franc, Andrew P. Davison, Padraig Gleeson, Fahim T. Imam, Birgit Kriener, Stephen D. Larson, Subhasis Ray, Lars Schwabe, Sean Hill, and Erik De Schutter. Computational Neuroscience Ontology: a new tool to provide semantic meaning to your models. *BMC Neuroscience*, 13(Suppl 1):P149, 2012.

[16] Nicolas Le Novère, Michael Hucka, Nadia Anwar, Gary D Bader, Emek Demir, Stuart Moodie, and Anatoly Sorokin. Meeting report from the first meetings of the Computational Modeling in Biology Network (COMBINE). *Standards in Genomic Sciences*, 5(2):230, 2011.

[17] Nicolas Le Novère, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, Emek Demir, Katja Wegner, Mirit I Aladjem, Sarala M Wimalaratne, et al. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741, 2009.

[18] Paul Richmond, Alex Cope, Kevin Gurney, and David J Allerton. From model specification to simulation of biologically constrained networks of spiking neurons. *Neuroinformatics*, 12(2):307–323, 2014.

[19] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, et al. Toward interoperable bioscience data. *Nature Genetics*, 44(2):121–126, 2012.

[20] Dagmar Waltemath, Richard Adams, Frank T Bergmann, Michael Hucka, Fedor Kolpakov, Andrew K Miller, Ion I Moraru, David Nickerson, Sven Sahle, Jacky L Snoep, et al. Reproducible computational biology experiments with SED-ML—the Simulation Experiment Description Markup Language. *BMC Systems Biology*, 5(1):198, 2011.

[21] Dagmar Waltemath, Frank T Bergmann, Claudine Chaouiya, Tobias Czauderna, Padraig Gleeson, Carole Goble, Martin Golebiewski, Mickael Hucka, Nick Juty, Olga Krebs, et al. Meeting report from the fourth

meeting of the Computational Modeling in Biology Network (COMBINE). *Standards in Genomic Sciences*, 9(3), 2014.

[22] Anna Zhukova, Dagmar Waltemath, Maciej J. Swat, Jon Olav Vik, Nicolas Rodriguez, and Nicolas Le Novère. Mathematical Modelling Ontology (MAMO). https://sourceforge.net/p/mamo-ontology/wiki/Home/, 2014.