Molecular Systems Biology (2006) doi:10.1038/msb4100052

www.molecularsystemsbiology.com
Article number: 2006.0010

molecular systems biology

NEWS AND VIEWS

Standardizing the standards

John Quackenbush[1,2]

[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA and [2] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Molecular Systems Biology 21 February 2006; doi:10.1038/msb4100052

The nice thing about standards is that there are so many to choose from.

Andrew S Tannenbaum

One of the most daunting aspects of using genomic technologies—including microarray, proteomic, metabolomic, and other approaches—is the sheer quantity of data that they produce. With thousands of biologically relevant molecules surveyed across (increasingly) large numbers of samples, interpretation of the data requires the use of computational approaches. And while many researchers thought that storing the data could simply build on our experiences with genome sequencing, it quickly became apparent that if one was to make sense of the results from any analysis, there was a need to store much more complex ancillary data than would be necessary for genome sequence. In 1999, as microarrays were establishing themselves as a truly viable technology, the Microarray Gene Expression Data Society (MGED; http://www.mged.org) arranged to define the critical information necessary to effectively analyze a microarray experiment and to describe a means of encoding that information. Through a series of discussions between interested parties, public presentations, and working group meetings, what emerged were the Minimal Information About a Microarray Experiment (MIAME) (Brazma *et al*, 2001; Ball *et al*, 2002, 2004) and MAGE-ML (Spellman *et al*, 2002), an XML-based markup language used for describing a microarray experiment.

The early success of MIAME and its widespread adoption by scientific journals also exposed some of its weaknesses, including the need to develop domain-specific extensions of MIAME to capture information about the experimental design and sample characteristics necessary for interpreting data coming, for example, from toxicology experiments (MIAME-Tox; Sansone *et al*, 2005) and extensions to other domains such as *in situ* hybridizations (MISFISHIE, the Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments; http://scgap.systemsbiology.net/standards/misfishie). In fact, the MGED subgroup on Reporting Structure for Biological Investigations Working Groups (RSBI WGs; http://www.mged.org/Workgroups/rsbi/rsbi.html) is looking at ways to extend MIAME to a wide range of other areas.

The principles underlying MIAME, particularly the need to clearly describe an experiment and report the variables necessary for data analysis, have resonated beyond the microarray community. For example, the metabolomics/metabonomics community/communities (I am not going to decide which is right, and by not doing so invite the scorn of all rather than one group or the other) are moving toward standardization and reporting of metabolic analyses (Lindon *et al*, 2005) and practitioners of proteomics have at least two XML-based standards for reporting proteomics from which to choose, HUP-ML (Hermjakob *et al*, 2004) and AGML (Stanislaus *et al*, 2004), as well as guidance through the Minimum Information About a Proteomics Experiment (MIAPE) (Orchard *et al*, 2004).

A recent paper by Novère *et al* extends the reporting standards notion beyond the experimental world, to the description of quantitative models of biochemical systems and attempts to reconcile some of the various standards that have evolved. The Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) standard proposed by this group is an attempt to bring together CellML (Lloyd *et al*, 2004) and SBML (Finney and Hucka, 2003; Hucka *et al*, 2003) and to gain acceptance from databases that archive models to provide access to these in a standard machine-readable format. This is an ambitious, but important, goal as systems biology hopes to produce quantitative models of cells and cellular processes. However, unless these models, which can become quite complex, are easily testable and comparable, they will ultimately be of little value. This is an important first step in helping to establish modeling and the value it will bring to developing a predictive biology, but the ultimate impact will depend on how widely the standard is adopted and how many software tools are developed to facilitate its use.

The utility of XML-based standards for facilitating data analysis in the complex realms was recently highlighted in a publication by Keller *et al* (2005). Keller describes the trans-proteomic pipeline, a proteomics data analysis pipeline consisting of a variety of software tools, which use different open XML standards to describe the data and manage the workflow: mzXML (Pedrioli *et al*, 2004) for the raw mass spec data, pepXML (http://www.matrixscience.com/xmlns/schema/pepXML_v18) for the peptides identified from the raw data, and protXML (http://sashimi.sourceforge.net/schema_revision/protXML/protXML_v3.xsd). This pipeline serves as a converter from one format to the other, and an interpreter and integrator of the results. Whereas this may seem trivial to those of us who remember the early days of DNA sequencing, where much of what we did in analyzing data was to convert sequence formats from GenBank to FASTA to GCG to Intelligenetics and back in all iterations, what Keller's pipeline does is much more subtle—it strings together descriptions of very different domains in the analysis, linking the spectral data

in a seamless way to the peptides it identifies and the proteins those peptides comprise.

Although the proliferation of standards and their increasing use are quite encouraging, there are some potential drawbacks. One of the major problems, as noted by Wang *et al* in a recent Nature Biotechnology paper, is the evolution of incompatible standards. What these authors point out is that the flexibility of XML allows definition of various tags that describe the same concept in a manner that does not lend itself to an obvious cross-reference. Using AGML and HUP-ML, Wang *et al* describe how a 2D protein gel can be described in ways that obfuscate the fact that these are, indeed, both descriptions of the same object. Even in MAGE-ML, we have found that XML's flexibility can allow two conflicting but completely 'correct' descriptions of the same experiment.

To address this problem, Wang *et al* (2005) suggest the use of the semantic web and its reference-document format (RDF; http://www.w3.org/RDF). Unlike XML, which has an inherently hierarchical structure, in RDF 'everything is a resource that connects with other resources via properties.' The problem with XML, as Wang *et al* note, is that 'descriptions of semantic relationships between nested content holders' are missing—which really means that for related objects, it is difficult to capture their relationship in the existing XML formats. The irony of this is that RDFs are described using XML; however, it is a very abstract yet simple representation that allows relationships between objects to be presented as the properties of the resources.

The beauty of a description based on RDF is that it can then be put into a variety of other formats, including XML, Notation 3 (N3, http://www.w3.org/DesignIssues/Notation3.html, a compact alternative to RDF's XML), and Directed Labeled Graphs (DLG, a graphical representation of RDF where 'nodes' are resources and 'edges' are properties linking the resources). Does this reintroduce the problem? Well, not really. The higher-level abstraction of RDF provides a way to cross-reference the various instantiations of the standard and provides a means of disambiguating their potential conflicts.

Is this the solution we are all waiting for? Well, not really. As the authors point out, constructing useful RDF descriptions requires a standard ontology—standardized descriptions of objects, elements, and processes using controlled vocabularies. And although in the first instance, this might seem to be a solvable problem across all of the diverse experimental domains trying to develop standards, the proliferation of disparate medical ontologies within the singular practice of medicine suggests that standardizing ontologies will not be an easy task. Despite this, abstracting the problem to the level of ontologies rather than leaving it in the muck and mire of XML specifications makes some sense.

But what is the real solution to this problem? The answer is pretty simple: money. What is remarkable about all of these standards, including MIAME, is that they have largely been developed through grass-roots efforts by 'concerned stakeholders' who want to assure that the data they are generating and managing are useful. This is 'blue collar' science—it is hard, often thankless work, and nobody is going to win a Nobel prize for creating a standard for describing how a microarray was hybridized or how a sample was injected into a mass spec. And because it is not glamorous, hypothesis-driven research, funding to support developing these standards or better yet, bringing them together, has been limited and slow in coming. But this is something we should all be concerned about. After all, the work of any one of us builds on that of those who have preceded and using that prior knowledge effectively is one of the things that will help accelerate the overall rate of scientific discovery. I, for one, am thankful to those who are developing and implementing standards (my involvement in MGED notwithstanding) and supportive of efforts to fund their work. After all, a rose by any other name is still a rose; you just cannot find it in the database.

References

Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M, Sansone SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N (2004) Submission of microarray data to public repositories. *PLoS Biol* **2:** E317

Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, Cavalieri D, Gaasterland T, Hingamp P, Holstege F, Ringwald M, Spellman P, Stoeckert Jr CJ, Stewart JE, Taylor R, Brazma A, Quackenbush J (2002) The underlying principles of scientific publication. *Bioinformatics* **18:** 1409

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* **29:** 365–371

Finney A, Hucka M (2003) Systems biology markup language: level 2 and beyond. *Biochem Soc Trans* **31:** 1472–1473

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol* **22:** 177–183

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19:** 524–531

Keller A, Eng J, Zhang N, Li X-j, Aebersold R (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* doi:10.1038/msb4100024

Lindon JC, Nicholson JK, Holmes E, Keun HC, Craig A, Pearce JT, Bruce SJ, Hardy N, Sansone SA, Antti H, Jonsson P, Daykin C, Navarange M, Beger RD, Verheij ER, Amberg A, Baunsgaard D, Cantor GH, Lehman-McKeeman L, Earll M, Wold S, Johansson E, Haselden JN, Kramer K, Thomas C, Lindberg J, Schuppe-Koistinen I, Wilson ID, Reily MD, Robertson DG, Senn H, Krotzky A, Kochhar S, Powell J, van der Ouderaa F, Plumb R, Schaefer H, Spraul M (2005) Summary recommendations for standardization and reporting of metabolic analyses. *Nat Biotechnol* **23:** 833–838

Lloyd CM, Halstead MD, Nielsen PF (2004) CellML: it's future, present and past. *Prog Biophys Mol Biol* **85:** 433–450

Orchard S, Hermjakob H, Julian Jr RK, Runte K, Sherman D, Wojcik J, Zhu W, Apweiler R (2004) Common interchange standards for proteomics data: public availability of tools and schema. *Proteomics* **4:** 490–491

Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **22:** 1459–1466

Sansone S, Morrison N, Rocca-Serra P, Fostel J (2005) Standardization initiatives in the (eco)toxicogenomics domain: a review. *Comp Funct Genomics* **8:** 633–641

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert Jr CJ, Brazma A (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3** RESEARCH0046

Stanislaus R, Jiang LH, Swartz M, Arthur J, Almeida JS (2004) An XML standard for the dissemination of annotated 2D gel electrophoresis data complemented with mass spectrometry results. *BMC Bioinform* **5:** 9

Wang X, Gorlitsky R, Almeida JS (2005) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nat Biotechnol* **23:** 1099–1103