NATIONAL ACADEMY OF SCIENCES

Survey of Bioinformatics Standards[1]

Jorge L. Contreras[2]

I. INTRODUCTION

The rapid growth of bioinformatics research over the past two decades has led to a surge of interest in the development of interoperability and compatibility standards for bioinformatics applications. These range from standards for genome annotation and controlled vocabularies (ontologies) to data formats and search engine integration. A variety of organizations are involved in these standards-development activities, from large, established standards bodies such as the Institute for Electrical and Electronics Engineers (IEEE) and the Worldwide Web Consortium (W3C) to broad-based bioinformatics industry associations such as the Asia-Pacific Bioinformatics Network and the European Bioinformatics Institute (EBI) to narrowly-focused efforts such as the Proteomics Standards Initiative (PSI) and the Functional Genomics Investigation Ontology (FuGO) project.

We conducted an industry-wide study of bioinformatics standards-development activities and reviewed the policies and procedures adopted by each standards-development organization, particularly in the areas of intellectual property and antitrust procedures. We observed that the majority of bioinformatics standards-development efforts are relatively informal and unstructured, mostly deriving from academic laboratories and scientific collaborative efforts. In many cases, these organizations either lack written policies entirely, or adopt vague,

---

[1] The results reported in this paper were first presented at the Law and Informatics Symposium at Northern Kentucky University, March 2012, and will appear in Jorge L. Contreras, *Implementing Procedural Safeguards for the Development of Bioinformatics Interoperability Standards*, __ N.Ky. L. Rev. ___ (2012).

[2] Associate Professor of Law, American University – Washington College of Law. The author gratefully acknowledges research assistance by Chris Pepe.

aspirational statements regarding a desire that materials produced be "open" and publicly-available.

II. THE BIOINFORMATICS STANDARDS LANDSCAPE

Bioinformatics research utilizes a broad range of technologies, from experimental apparatus such as microarrays, to data analysis tools, to databases that store and allow the sharing of experimental and analytic data. Standards are required within each of these broad technology categories to enable data sharing and analysis and the interoperability of different experimental platforms. To-date, hundreds of standards relevant to bioinformatics applications have been developed in three broad categories: terminological artifacts, reporting requirements and exchange formats.[3] Below is a brief description of these categories and a summary of some of the more prominent standardization efforts being undertaken in each.

A. Terminological Artifacts

In order for different groups and applications to communicate about a wide array of organisms, experimental conditions and study designs, a consistent and unambiguous vocabulary is required. A "controlled vocabulary" offers a single set of terms with explicitly-defined meanings within a particular field. An "ontology" creates relationships among these terms, often in hierarchical form, such as the familiar taxonomic classification for biological entities (kingdom, phylum, class, etc.).

One of the earliest and most mature bioinformatics ontologies is the Gene Ontology (GO), which has produced "a structured, precisely defined, common, controlled vocabulary for describing the role of genes and gene products in any organism."[4] The Gene Ontology Consortium, which developed the Gene Ontology, began its work in 1998 as a joint project of research groups studying three different

[3] This classification system was developed by Biosharing.org, which catalogs and provides information regarding most of the standards discussed below. *See* www.biosharing.org/standards (last visited October 28, 2011). A list of the standards compiled by biosharing.org can be found at www.biosharing.org/standards_view.

[4] *See* Michael Ashburner et al., *Gene Ontology: Tool for the Unification of Biology*, NATURE GENETICS, May 2000, at 25.

organisms (the fruit fly, budding yeast, and mouse). As of 2006, the GO included more than 1.6 million annotated gene products.[5]

Numerous other ontology projects have arisen following the success of the GO.[6] These include the Systems Biology Ontology (SBO) originated by the European Bioinformatics Institute (EBI),[7] the Open Biological and Biomedical Ontologies (OBO Foundry)[8], and the Functional Genomics Investigation Ontology (FuGO) project.[9]

B. Reporting Requirements

The advent of microarray technology in the 1990s quickly led to the realization by the research community that standardized methods of reporting experimental data generated by microarray studies would be required. The Microarray Gene Expression Data (MGED) Society (now the Functional Genomics Data Society) was formed in 1999 by EBI to develop Minimum Information About a Microarray Experiment (MIAME), a checklist specifying the information about every microarray experiment that should be reported in order to enable its proper validation, reproduction and interpretation.[10]

Many other minimum information standards development efforts have followed the early success of MIAME.[11] Among others, these include specifications for minimum information in hybridization and immunohistochemistry experiments (MISFISHIE),[12] proteomics experiments (MIAPE),[13] molecular interactions

---

[5] See Gene Ontology Consortium, *The Gene Ontology (GO) Project in 2006*, 34 NUCLEIC ACID RESEARCH D322, D323 (2006).
[6] A more comprehensive list of ontology projects can be found in Chris F. Taylor, *Standards for Reporting Bioscience Data: A Forward Look*, 12 DRUG DISCOVERY TODAY 527, 529 (2007).
[7] *See* Eur. Molecular Biology Lab.-Eur. Bioinformatics Inst., SYSTEMS BIOLOGY ONTOLOGY, http://www.ebi.ac.uk/sbo/main/ (last visited October 28, 2011).
[8] *See* Barry Smith, *The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration*, 25 NATURE BIOTECHNOLOGY 1251 (2007).
[9] *See* Patricia L. Whetzel, et al., *Development of FuGO: An Ontology for Functional Genomics Investigations*, 10 OMICS 199 (2006).
[10] *See* Catherine A. Ball & Alvis Brazma, *MGED Standards: Work in Progress*, 10 OMICS 138, 139 (2006).
[11] There are more comprehensive lists of minimum information standards projects available. Taylor, *supra* note 9, at 528; Lyle D. Burgoon, *Clearing the Standards Landscape: the Semantics of Terminology and their Impact on Toxicologenomics*, 99 TOXICOLOGICAL SCIENCES 403, 408-09 (2007).
[12] *See* Inst. for Sys. Biology, MISFISHIE, http://scgap.systemsbiology.net/standards/misfishie/ (last visited Oct. 28, 2011).

(MIMIx)[14] and (meta)genome sequences (MIGS/MIMS).[15] The Minimum Information for Biological and Biomedical Investigations (MIBBI) Project collects and publishes information about various minimum information standards.[16]

C. Exchange Formats

A recent compilation of molecular biology databases lists 1330 different data sources across the world.[17] In order for researchers to make use of data beyond their own laboratories, they require the ability to access and utilize data from disparate sources. The exchange of data among different software applications and databases has become commonplace in today's data-driven economy. Much of this exchange is accomplished using markup languages, schema that enable the annotation of digital text in a manner that can be interpreted by a computer. The most common markup language, Hypertext Markup Language (HTML), is the dominant language for encoding web pages. Markup languages work through the use of "tags" that delimit characteristics of the text that they designate. For example, in HTML, the tags <B> and </B> cause the text between the tags to be displayed in boldface type. Extensible Markup Language (XML), developed by the Worldwide Web consortium (W3C), is a flexible data format that enables users to create customized tags based on the specific types of data in which they are interested. Hundreds of XML-based languages exist today, many of which are optimized for bioinformatics applications.[18] Among the best-known XML-based bio-focused

---

[13] *See* Sandra Orchard & Henning Hermjakob, *The HUPO Proteomics Standards Initiative – Easing Communication and Minimizing Data Loss in a Changing World*, 9 BRIEFINGS IN BIOINFORMATICS 166, 167 (2007).
[14] *See id*. at 170-71.
[15] *See* Pelin Yilmaz, *The Genomic Standards Consortium: Bringing Standards to Life for Microbial Ecology*, 5 ISME J. 1565 (2011).
[16] See Chris F. Taylor, *Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: the MIBBI Project*, 26 NATURE BIOTECHNOLOGY 889 (2008).
[17] Michael Y. Galperin & Guy R. Cochrane, *The 2011 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection*, 39 NUCLEIC ACIDS RESEARCH D1 (2011).
[18] More comprehensive lists of bioinformatics markup languages can be found in Burgoon, *supra* note 14, at 409; and Luciano Milanesi, *Trends in Modeling Biomedical Complex Systems*, 10 BMC BIOINFORMATICS, Supp. 12 (2009).

languages are the Systems Biology Markup Language (SBML),[19] CellML for computational cell biology,[20] MAGE-ML, for the exchange of microarray data, [21] and BioPAX, for the exchange of biological pathway data.[22]

Despite the widespread adoption of XML as the preferred standard for data exchange in the biosciences, some commentators have criticized XML as too limited and imprecise for the robust exchange of scientific data.[23] An alternative data exchange standard is the Reference Data Format (RDF), also developed by W3C, which takes advantage of the so-called "semantic web" and offers developers greater freedom to define data relationships.[24] The Bio2RDF Project led by W3C has recently developed software and a website for accessing data from different public bioinformatics databases in RDF format.[25]

Closely related to markup languages are object models, which describe the relationships among computer programming "objects" of interest within a given discipline. Object models can thus act as blueprints for the development of specialized markup languages. Of particular interest in bioinformatics are the Microarray and Gene Expression Object Model (MAGE-OM),[26] the SysBio-OM developed by the National Institute of Environmental Health Sciences (NIEHS)[27] and the Functional Genomics Experiment Object Model (FuGE-OM).[28]

The object models and markup languages discussed above are essential for the exchange of data among disparate data sources and databases. Yet data

---

[19] *See* M. Hucka, *The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models*, 19 BIOINFORMATICS 524 (2003).
[20] *See* Catherine M. Lloyd et al., *CellML: its Future, Present and Past*, 85 PROGRESS IN BIOPHYSICS & MOLECULAR BIOLOGY 433 (2004).
[21] *See* Ball & Brazma, *supra* note 13, at 140-41.
[22] See Christoph Wierling et al., *Resources, Standards and Tools for Systems Biology*, 6 BRIEFINGS IN FUNCTIONAL GENOMICS & PROTEOMICS 240, 245-46 (2007).
[23] *See, e.g.*, John Quackenbush, *Standardizing the Standards*, MOLECULAR SYSTEMS BIOLOGY, Feb. 21, 2006, at 1, 1-2; Xiaoshu Wang, et al., *From XML to RDF: How Semantic Web Technologies will change the Design of 'omic' Standards*, 23 NATURE BIOTECHNOLOGY 1099 (2005).
[24] *See id*.
[25] *See* Francois Bellau, et al., *Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems*, 41 J. BIOMEDICAL INFORMATICS 706 (2008).
[26] Functional Genomics Data Society, Microarrary Gene Experiment – Object Model, http://www.mged.org/Workgroups/MAGE/mage.html (last visited Oct. 29, 2011).
[27] Lyle D. Burgoon, *Clearing the Standards Landscape: the Semantics of Terminology and their Impact on Toxicogenomics*, TOXICOLOGICAL SCIENCES, Oct. 2007, at 403, 409.
[28] FUNCTIONAL GENOMICS EXPERIMENT, Functional Genomics Experiment Object Model, http://fuge.sourceforge.net/index.php (last visited Oct. 29, 2011).

exchange tools alone are not enough, and commentators find that the proliferation of incompatible data sources has led to increasing duplication of effort, poor interoperability and loss of data.[29] To address these problems, an international consortium of database users and developers has begun work on a uniform set of defining attributes for biological databases called BioDBCore.[30]

III. SURVEY OF BIOINFORMATICS STANDARDS POLICIES

We reviewed the publicly-available policies and rules of a number of major bioinformatics standards initiatives. The results, summarized in *Table 1* below, indicate that few standardization efforts in bioinformatics address legal issues in detail, and a significant number omit any legal guidelines in their public documentation.

Table 1

Summary of Selected Bioinformatics SDO Policies

Standards-Development Organization (SDO)	Year Started	Notable Standards	Type	Antitrust Guidelines	Intellectual Property policy
Asia-Pacific Bioinformatics Network	1998	Minimum Information About a Bioinformatics Investigation	Reporting	None	None
BioPAX.org	2002	BioPAX	Exchange	None	Open source licenses that are academic and corporate friendly are used for all work created by the group
European Bioinformatics Inst. (EBI)	1992	Systems Biology Ontology (SBO)	Terminology	None	No restrictions on the use or redistribution of data. Some original data may be subject to patent, copyright, or other intellectual property rights, and users must ensure that their exploitation of the data does not infringe the rights of such third parties.

[29] *See* Pascale Gaudet, et al., *Towards BioDBCore: A Community-Defined Information Specification for Biological Databases*, 39 NUCLEIC ACIDS RESEARCH D7 (2011).
[30] *Id.*

Standards-Development Organization (SDO)	Year Started	Notable Standards	Type	Antitrust Guidelines	Intellectual Property policy
Functional Genomics Investigation Ontology (FuGO) project	2006	Functional Genomics Investigation Ontology (FuGO)	Terminology	None	None
Functional Genomics Data Socy. (FGED) [formerly Microarray Gene Expression Data (MGED) Socy.]	1999	MGED Ontology Minimum Information About a Microarray Experiment (MIAME) Minimum Information Specification For *In Situ* Hybridization and Immunohistochemistry Experiments (MISFISHIE) MAGE-ML	Terminology Reporting Reporting Exchange	None	No funding sources have IP rights over FGED's output; FGED currently owns no patents or copyrights, and advocates free use of data, tools, and publications
Genomic Standards Consortium	2005	Minimum Information about a MARKer gene Sequence standard (MIMARKS)	Reporting	None	GSC utilizes Common Public License Version 1.0 (CPL). The license generally provides that any contributor grants a royalty free license to use any copyrighted or patented materials that he/she contributes.
Intl. Socy. Biocuration	N/A	BioDBCore	Reporting	None	None
Gene Ontology Consortium	1998	Gene Ontology	Terminology	None	Unrestricted use if user acknowledges GO as source, displays the version number, and does not alter files
Metabolomics Standards Initiative (MSI)	2005	Core Information for Metabolomics Reporting (CIMR)	Reporting	None	Working group output is fully available to public
OBO Foundry	2001	Does not create its own, but endorses ontologies that meet its principles and standards	Terminology	None	OBO will not endorse ontology unless it allows for free use to all

Standards-Development Organization (SDO)	Year Started	Notable Standards	Type	Antitrust Guidelines	Intellectual Property policy
Proteomics Standards Initiative (PSI) [HUPO]	2002	Minimum Information about a Proteomics Experiment (MIAPE) Min. Information about a Molecular Interaction Experiment (MIMIx) Min. Information about a Protein Affinity Reagent (MIAPAR) Proteomics Standards Initiative-Molecular Interaction (PSI-MI) XML mzML	Reporting Reporting Reporting Exchange Exchange	None	PSI takes no position regarding validity or scope of any IP or other rights pertaining to the implementation or use of technology or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights.
SBML.org	2000	Systems Biology Markup Language	Exchange	None	The organization considers SBML to be free and open to the community, such that no one owns rights to it
University of Auckland	2004[31]	CellML	Exchange	None	Individuals may freely use, publish and redistribute CellML; write and sell applications which create, load, or write CellML-valid XML files; distribute or sell their own CellML-valid XML files; and transmit verbatim copies of the CellML format to any person without restriction
Worldwide Web Consortium (W3C)	1994	Bio2RDF	Exchange	None	Bio2RDF is released under GPL v. 2.0; W3C will not approve a recommendation if it is aware that essential patent claims are not available on royalty-free terms

As shown in *Table 1*, in many cases, the organizations responsible for bioinformatics standards development consist of academic research networks or scientific collaborations. Many either lack written policies entirely, or have adopted

[31] Work on CellML began in 2004. The University itself is much older.

vague, aspirational statements regarding a desire that materials produced be "open" and publicly-available. Nevertheless, the data collected in *Table 1* does reveal a number of recurring themes. These include a general preference that bioinformatics standards be "open" and "not restricted" by intellectual property rights. These preferences are not surprising, given the dominance of academic research groups in the development of bioinformatics standards to-date. It remains to be seen whether these preferences continue to dominate the field if commercial interests begin to play a greater role in the bioinformatics field, more generally.