Published in final edited form as:
OMICS. 2006 ; 10(2): 199–204. doi:10.1089/omi.2006.10.199.

Development of FuGO: An Ontology for Functional Genomics Investigations

Patricia L. Whetzel[1], **Ryan R. Brinkman**[2], **Helen C. Causton**[3], **Liju Fan**[4], **Dawn Field**[5], **Jennifer Fostel**[6], **Gilberto Fragoso**[7], **Tanya Gray**[5], **Mervi Heiskanen**[7], **Tina Hernandez-Boussard**[8], **Norman Morrison**[9,10], **Helen Parkinson**[11], **Philippe Rocca-Serra**[11], **Susanna-Assunta Sansone**[11], **Daniel Schober**[11], **Barry Smith**[12], **Robert Stevens**[9], **Christian J. Stoeckert Jr.**[1], **Chris Taylor**[11], **Joe White**[13], **Andrew Wood**[10], and **FuGO Working Group**[14]

[1]Center for Bioinformatics and Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania. [2]Terry Fox Laboratory, British Columbia Cancer Research Center, Vancouver, BC, Canada. [3]MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College, Hammersmith Hospital Campus, London, United Kingdom. [4]Ontology Workshop LLC, Columbia, Maryland. [5]Molecular Evolution and Bioinformatics Section, Oxford Centre for Ecology and Hydrology, Oxford, United Kingdom. [6]NIEHS, Research Triangle Park, North Carolina. [7]NCICB, NCI Center for Bioinformatics, Rockville, Maryland. [8]Department of Genetics, Stanford University Medical Center, Stanford, California. [9]Department of Computer Science, University of Manchester, Manchester, United Kingdom. [10]NERC Environmental Bioinformatics Centre, Oxford Centre for Ecology and Hydrology, Oxford, United Kingdom. [11]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom. [12]Department of Philosophy, Center of Excellence in Bioinformatics and Life Sciences, and National Center for Biomedical Ontology, University at Buffalo, Buffalo, New York, and Institute for Formal Ontology and Medical Information Science, Saarbrücken, Germany. [13]Dana-Farber Cancer Institute, Boston, Massachusetts.

Abstract

The development of the Functional Genomics Investigation Ontology (FuGO) is a collaborative, international effort that will provide a resource for annotating functional genomics investigations, including the study design, protocols and instrumentation used, the data generated and the types of analysis performed on the data. FuGO will contain both terms that are universal to all functional genomics investigations and those that are domain specific. In this way, the ontology will serve as the "semantic glue" to provide a common understanding of data from across these disparate data sources. In addition, FuGO will reference out to existing mature ontologies to avoid the need to duplicate these resources, and will do so in such a way as to enable their ease of use in annotation. This project is in the early stages of development; the paper will describe efforts to initiate the project, the scope and organization of the project, the work accomplished to date, and the challenges encountered, as well as future plans.

Address reprint requests to: Dr. Patricia L. Whetzel, Center for Bioinformatics and Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, whetzel@pcbi.upenn.edu.

[14]See FuGO webpage for a complete list of contributing groups and their members: ⟨http://fugo.sf.net⟩.

This paper is part of the special issue of OMICS on data standards.

INTRODUCTION

Functional Genomics is an umbrella term for a wide range of biologically and technologically defined domains, each with its own linguistic peculiarities and commonalities. The application of high-throughput functional genomics technologies such as microarrays and modern mass spectrometry to a wide variety of biological conditions has facilitated the investigation of new kinds of biological questions. The use of these technologies has also resulted in the generation of massive amounts of data and metadata (in essence, metadata is information about the data). Their extension to ever more diverse organisms, experimental conditions, and study designs means an increase in the likely variety and complexity of metadata. In order for these kinds of investigations to be understandable to and potentially reproducible by other scientists, unambiguous descriptions of the data and metadata are required.

Providing unambiguous descriptions is complicated by the fact that natural languages are not precise tools, as words often have several meanings. Dictionaries do not really help, because they will offer all the possible meanings of a word (or phrase). What is needed is a set of words (terms) that are assigned a single "official" meaning. This is a "controlled vocabulary." But we can go further: as well as ensuring that a controlled vocabulary contains all the terms needed to describe an investigation, each with a single, clear definition, we can group together those words that relate to similar things. Imagine a music collection: to organize that collection, one could arrange it so that the artists' names were in alphabetical order, or alternatively one could put similar kinds of music together (dance, rock, classical and so on). Such categories could be (repeatedly) subdivided; classical music could be baroque, romantic or modern, for example. This is how an ontology is structured: the well-defined terms that we might find in a controlled vocabulary are grouped into a hierarchy via the parent-child relation. More general parent terms are positions in the hierarchy above more specific child terms. Each child has a single parent, but any given parent will characteristically have multiple children, called "siblings." The terms in the hierarchy then represent classes or types in reality, with more general parent terms representing more inclusive classes, and more specific child terms representing their subclasses or subtypes. Relationships other than similarity between classes can also be captured, such as expressing that one thing is "part of" another. This classification is useful to allow people or computers to find the official term for a thing (an alphabetical list would not be much help); the ontology also allows "generic" queries to be written (i.e., searching on the parent of a group of terms, for example "equipment" rather than "chromatography column"). Using terms drawn from ontologies in the description of an investigation also makes the description "machine-readable," which means that computers can retrieve (from databases), analyse and compare descriptions and data, and in many cases actually complete an analysis unaided, which is crucial given the continual increase in data volumes.

One example of a mature ontology is the Gene Ontology (Gene Ontology Consortium, 2000), which aims to support the consistent description of gene products. It actually consists of three ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions. The appropriate use of GO terms by databases facilitates uniform queries across them. The three ontologies are structured so that one can query them at different levels; for example, finding database records that mention gene products involved in signal transduction, or that refer specifically to receptor tyrosine kinases. This structuring also allows properties to be assigned to a gene product at different levels, depending on how much is known about that gene product.

A closer parallel to the Functional Genomics Investigation Ontology (FuGO) in terms of focus is the MGED Ontology (Whetzel et al., 2006). The MGED Ontology focuses on experimental methods, as does FuGO, whereas the Gene Ontology concentrates on the entities studied in biological experiments. The MGED Ontology has been successfully used to annotate

microarray data for some time. It was developed to provide a source of terms to support consistent description of microarray-based experiments, and also enables the querying of data simply and consistently across databases.

The purpose of FuGO is to support the consistent description of functional genomics investigations, regardless of the particular biological or technological domain. A tightly knit collaboration between the transcriptomics, proteomics and metabolomics standards initiatives (Ball and Brazma, *this issue*; Fiehn et al., *this issue*; Taylor et al., *this issue*; Sansone et al., *this issue*) is at the root of this collaborative project, but the aim is to be fully open and inclusive. The scope of FuGO includes investigations (their rationale and to an extent their structure), input and output materials, protocols and instrumentation, the data generated and the types of analyses performed. FuGO will contain both highly general terms referring to kinds of entities common across various biological and technological domains and terms for domain-specific subtypes. The key point is that the terms in the ontology are organized (grouped and interrelated) as described above to support intelligent, potentially domain-independent searching (of databases or the literature), automated reasoning across data sets and automated error-correction, in addition to providing a resource of general utility to the scientific community (e.g., the ontology can serve as an encyclopedia of experimental methods).

PROJECT ROADMAP

The first stage of the FuGO project involved the identification of communities with an interest in collaborating on the development of a set of descriptors to annotate functional genomics investigations. A complete listing of all the currently involved communities is given in Table 1.

The next stage was to define the scope and organization of the project, to ensure effective and transparent communication. The FuGO project is organized into three groups; the coordination committee, the advisory board, and the working group. The coordination committee includes individuals that represent functional genomics communities interested in this project; it currently contains representatives from the groups listed in Table 1 and remains open to members' representative of other interested communities. The purpose of this committee is to address the interests of all involved communities, to discuss general organizational matters such as the project web site (⟨http://fugo.sf.net⟩), to coordinate presentations at meetings and to manage the design, evolution and extension of FuGO.

The FuGO advisory board (FAB) consists of an invited group of experts on functional genomics and ontology development and management. The purpose of this group is to provide advice on "ground rules" for the ontology's development (e.g., the use/avoidance of particular relations). High-level design decisions, such as the overall structure of the ontology, will also be addressed by this body. The FAB includes leaders of the National Center for BioMedical Ontology (Rubin et al., *this issue*) and of the Ontogenesis Network: (⟨www.ontonet.org⟩).

The FuGO working group contains all members of the above two groups, plus participants from all of the involved communities. The FuGO working group is responsible for developing the general structure of the ontology, engineering extension mechanisms for specific biological and technological domains and proposing terms for the ontology (both universal and community-specific). Documentation of the scope and organization of FuGO and a listing of the members of both the advisory board and the community coordinators together with their roles and responsibilities can be found on the FuGO web site.

The next phase of the project focused on the review of use cases from the involved communities. Use cases were collected by members of the coordination committee and then reviewed by the FuGO working group to identify terms to be included in the ontology (and to educate the

members of the working group about the needs of each community). Additional terms have also been identified by examination of previous work by the various involved communities. For example, the MGED Society has been active in the standards efforts and includes groups that are now contributing to the development of FuGO. In particular, the MGED RSBI working group has identified a core set of basic terms with which to describe an investigation that can encompass any specific biological application; they have developed a concept map based on these use cases (Sansone et al., *this issue*). Members of the MGED Ontology working group are also contributing to the development of FuGO in part by reviewing the MGED Ontology (MO) to identify types that are common to functional genomics investigations across all domains, as well as those that are specific to microarray technology. The lessons learned in developing the MO will speed the development of FuGO and ensure that the resulting ontology is reusable and not tied to any specific object model. The Human Proteome Organization's Proteomics Standards Initiative (HUPO PSI) (Taylor et al., *this issue*) generates controlled vocabularies to support its modular, instrument-oriented standard data formats; these vocabularies will be integrated into FuGO. The PSI does not attempt to generate investigation-level descriptors in isolation; but does take part in the collaborative development of those terms in FuGO, and ultimately will be a consumer of them. The Metabolomics Society's Ontology Working Group (Fiehn et al., *this issue*) is in the process of reviewing all existing relevant resources, including terms from the PSI and existing and emerging work from within metabolomics; they too will generate new vocabulary terms as required and contribute those to FuGO. Overall, the corralling together of these diverse groups will reduce the chance that FuGO will contain duplicate (equivalent) terms for the same types as these are addressed by different communities (such commonalities can be found at many levels).

The next step of the project will be to construct the ontology by gathering the terms needed by each of the communities, then identifying the associated relations needed to construct the ontology. In cases where mature ontologies already exist, for example, ontologies for anatomical terms, FuGO will reference these ontologies as a source of annotation terms to avoid duplication, but in such a way as to enable their ease of use in annotation. The FuGO project will follow ontology best practices as far as is practical (Smith et al., 2005) in consultation with the FuGO advisory board. FuGO is being developed in Protégé using the OWL plugin (Noy et al., 2003; Knublauch et al., 2004).

Policies for maintaining the ontology will also be needed, for example, on the correction or deprecation of included terms or relations and on the accession process for new terms. These policies will need to take into account the provenance of the term, whether the term is universal or specific to a given biological or technological domain, which of the communities is responsible for the final approval of the definition of the term, and the location or parentage of the term in the ontology. In addition, a policy regarding the identification of which individuals will be able to edit the ontology is needed; for example, will one individual per community edit the file or will there be just one master ontology editor? Other policy needs are likely to arise as the building of the ontology progresses.

CHALLENGES

The challenges associated with this project are both technological and societal. Technological challenges include identifying a mechanism to develop and maintain the ontology so that it meets the needs of all the communities involved. Each community may wish to maintain its own ontological area, which implies a need for modularization. Terms may well be needed by several communities, but not all; the ontological equivalent of "public," "private," or "protected" are not yet available. Furthermore, when delivering the (full) ontology to various communities for use in experimental descriptions, community-based (partial, and potentially overlapping) views will need to be generated. In developing OWL ontologies, these issues

remain challenging. Development of new applications may be required to meet the needs of those developing and maintaining an ontology whose content is sourced from such a heterogeneous set of contributors.

The societal challenges relate to the coordination of the different communities; for example, avoiding duplication of effort in identifying and defining terms and managing the development of the ontology within the timeframe needed by different communities. Other challenges include the lack of central funding for the FuGO project. This has been overcome by investment from the involved communities and the use of freely available resources where possible. Currently, work on this project occurs by way of weekly conference calls, which have been subsidized by the involved communities, and discussions on the FuGO project mailing list (⟨https://lists.sourceforge.net/lists/listinfo/fugo-devel⟩) hosted by SourceForge.net, which provides free hosting to open source software development projects. Additional work has occurred at FuGO ontology workshops, the first of which was run in February 2006. A report of the workshop is posted online (⟨http://fugo.sf.net⟩).

CONCLUSION

The development of the Functional Genomics Investigation Ontology is a collaborative, international effort, which will provide a resource for annotating functional genomics investigations. FuGO will contain terms that are both universal across functional genomics investigations and those that are community-specific. In addition, FuGO will reference existing mature ontologies to avoid duplication, but in such a way as to enable their ease of use in annotation. More information on how to become involved with this project is available at the FuGO web site (⟨http://fugo.sf.net⟩).

Acknowledgments

The FuGO project would like to thank the National Cancer Institute and the MGED Society for hosting web and phone conferences. The contributing authors are supported in part by the Intramural Research Program of the NIH and NIEHS (contract 273-02-C-0027), NIH/NIBIB (grant EB-5034), NHGRI and NIBIB (P41 HG003619-01), EU Network of Excellence NuGO (NoE 503630), BBSRC (grant BB/D524283/1), NIH (grant UO1GM61374), and NCRI, EMBL, and NIH Roadmap for Medical Research (Grant 1 U 54 HG004028).

REFERENCES

Ball CA, Brazma A. MGED standards: work in progress. OMICS. 2006(this issue)

Bruskiewich R, Davenport G, Hazekamp T, et al. Generation Challenge Programme (GCP): standards for crop data. OMICS. 2006(this issue)

Fiehn O, Kristal B, Van Ommen B, et al. Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. OMICS. 2006(this issue)

Field D, Morrison N, Selengut J, et al. eGenomics: cataloging our complete genome collection. OMICS. 2006(this issue)

Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–29. [PubMed: 10802651]

Knublauch, H.; Musen, MA.; Rector, AL. Editing description logic ontologies with the Protégé OWL Plugin. A technical discussion for logicians at the International Workshop on Description Logics. Canada: Whistler; 2004.

Morrison N, Cochran G, Faruque N, et al. Concept of sample in OMICS technology. OMICS. 2006(this issue)

Noy, NF.; Crubezy, M.; Fergerson, RW., et al. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment; AMIA Annu Symp Proc 2003; 2003. p. 953

Rubin DL, Lewis SE, Mungall CJ, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS. 2006(this issue)

Sansone SA, Rocca-Serra P, Tong W, et al. A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. OMICS. 2006(this issue)

Smith B, Ceusters W, Klagges B, et al. Relations in biomedical ontologies. Genome Biol 2005;6:R46. [PubMed: 15892874]

Spidlen J, Gentleman RC, Haaland PD, et al. Data standards for flow cytometry. OMICS. 2006(this issue)

Taylor CE, Hermjakob H, Julian RK Jr, et al. The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). OMICS. 2006(this issue)

Whetzel PL, Parkinson H, Causton HC, et al. The MGED Ontology; a resource for semantics-based description of microarray experiments. Bioinformatics 2006;22:866–873. [PubMed: 16428806]

TABLE 1

Communities Collaborating in the Development of the Functional Genomics Investigation Ontology

Community	Organization	URL	Reference
Crop sciences	Generation Challenge Programme	⟨www.generationcp.org⟩	Bruskiewich et al. *this issue*
Environmental genomics	MGED RSBI	⟨www.mged.org/Workgroups/rsbi/rsbi.html⟩	Morrison et al. *this issue*
Flow cytometry	International Society for Analytical Cytology	⟨www.flowcyt.org⟩	Spidlen et al. *this issue*
Genomics	Genomic Standards Consortium (GSC)	⟨www.genomics.ceh.ac.uk/genomecatalogue⟩	Field et al. *this issue*
Metabol/nomics	mSI Ontology Working Group	⟨http://msi-workgroups.sourceforge.net⟩	Fiehn et al. *this issue*
Nutrigenomics	MGED RSBI	⟨www.mged.org/Workgroups/rsbi/rsbi.html⟩	Sansone et al. *this issue*
Polymorphism	N/A	N/A	N/A
Proteomics	HUPO-PSI	⟨http://psidev.sourceforge.net⟩	Taylor et al. *this issue*
Toxicogenomics	MGED RSBI	⟨www.mged.org/Workgroups/rsbi/rsbi.html⟩	Sansone et al. *this issue*
Transcriptomics	MGED Ontology Working Group	⟨http://mged.sourceforge.net/ontologies/⟩	Ball and Brazma, *this issue*

N/A, not applicable.