Graduate School of Library and Information Science
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Understanding biocurators

Attributes and roles of model organism database curators

W. John MacMullen
University of Illinois, 501 E. Daniel St, MC-493, Champaign IL 61820-6211 USA wjohn@uiuc.edu

ABSTRACT

Objective:
Biocurators are subject-matter experts who curate knowledge from the biomedical literature and other sources to enrich the content of model organism databases and other biomedical information resources. This project describes biocurators' educational backgrounds and biological expertise, organisms with which they have laboratory- and Gene Ontology annotation experience, and details about their work tasks and roles.

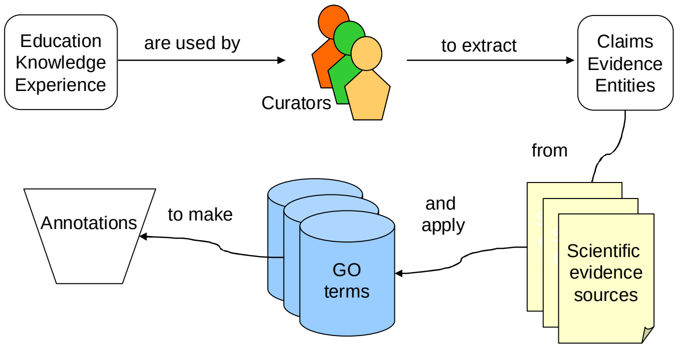
Methods:
Contextual data about educational backgrounds (degree levels and subjects), subject-matter expertise (specializations and experience), and work roles was collected from 31 biocurators as a part of two larger studies of Gene Ontology annotation variation. A brief self-report questionnaire was used to obtain curators' background information. Individual semi-structured 30-minute interviews were conducted with 15 curators, and a 60-minute focus group was conducted with 12 biocurators, some from the same cohort. The interviews and focus group explored the tasks, workflows, and practice environments of the curators. The data were analyzed with descriptive statistics for the questionnaire data and content analysis for the interview and focus group data.

Results:
Most (90%) biocurators studied held Ph.D degrees, in such subject areas as genetics (28%), biochemistry (10%), and molecular biology (10%), and had extensive laboratory experience. The years of GO annotation experience biocurators reported ranged from a few months to several years. Biocurators' tasks include Gene Ontology annotation, phenotype characterization, linking to other information resources, and supplementary indexing using specialized controlled vocabularies to provide end-users with access points that are tied to biological entities (e.g., genes) rather than scientific articles, and are more granular and specialized than topical (MeSH) indexing. Biocurators also participate in interface design and end-user education and support.

Conclusions:
In addition to being users of library services, biocurators are both peers of, and potential collaborators for, librarians in the health and biomedical sciences. Librarians serving biomedical research populations should be aware of the attributes and roles of biocurators, whose roles are so similar to their own.

Acknowledgments

- This work benefited from discussions with curators from multiple model organism databases at the 2005 and 2006 Gene Ontology Annotation Camps held at Stanford University. The Stanford Department of Genetics and the *Saccharomyces* Genome Database (SGD) provided travel support for this project.
- W.J.M. was supported in part by an Individual Biomedical Informatics Fellowship (F37 LM009194) from the National Library of Medicine, and an unrestricted research gift from Microsoft Research to the Annotation of Structured Data research group in the School of Information and Library Science at the University of North Carolina at Chapel Hill.


Education
Knowledge
Experience
are used by
Curators
to extract
Claims
Evidence
Entities
from
Scientific evidence sources
and apply
GO terms
to make
Annotations

Gene Ontology annotation process

Curators' Biological expertise (N=23)

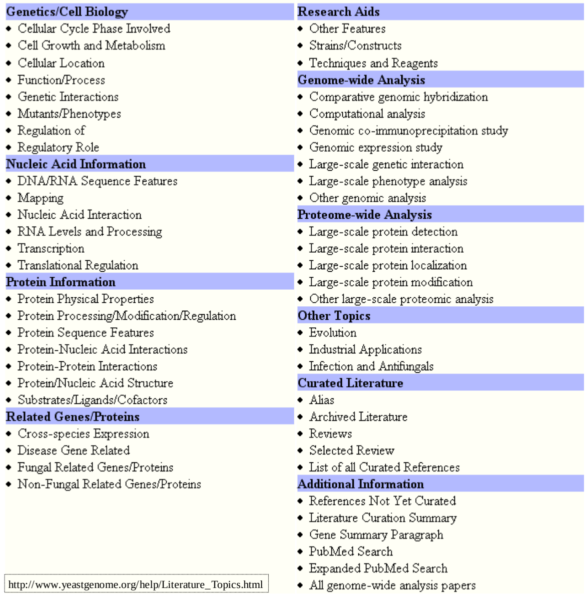
Biological Expertise (self-reported, first responses)	#	%
Development	4	17.4
Plant developmental biology	2	8.7
Cell biology	1	4.3
Cell cycle	1	4.3
Cell wall construction	1	4.3
DNA-protein interactions	1	4.3
DNA methylation	1	4.3
Embryonic development	1	4.3
Enzymology	1	4.3
Genome evolution	1	4.3
Immunology	1	4.3
Inflammation	1	4.3
Protein regulation and degradation	1	4.3
Proteomics	1	4.3
Transposition	1	4.3
Regulation of gene expression	1	4.3
Virology	1	4.3
"none"	1	4.3
blank	1	4.3

Model Organisms

Model organisms are biological organisms which have high research utility due to certain features, such as relative simplicity, small genome size, or functional similarity to aspects of human biology. Within biomedical research they are valued for their use as surrogates for human gene expression analysis. Model organism databases (MODs) provide rich collections of professionally curated information about specific model organisms. The ten MODs studied in this project were:

- DictyBase, for the mold *Dictyostelium discoideum*
- Flybase, for the fruitfly *Drosophila melanogaster*
- GeneDB from Sanger Institute for the fungus *Schizosaccharomyces pombe*
- Gramene, for the rice *Oryza sativa*
- MGD, the Mouse Genome Database, for *Mus musculus*
- RGD, the Rat Genome Database, for *Rattus norvegicus*
- SGD, the Saccharomyces Genome Database, for *Saccharomyces cerevisiae* (yeast)
- TAIR, the Arabidopsis Information Resource, for *Arabidopsis thaliana* (mustard plant)
- Wormbase, for the roundworm *Caenorhabditis elegans*
- Zfin, the Zebrafish Information Network, for *Danio rerio*

Yeast genome database supplementary indexing vocabulary


Genetics/Cell Biology
- Cellular Cycle Phase Involved
- Cell Growth and Metabolism
- Cellular Location
- Function/Process
- Genetic Interactions
- Mutants/Phenotypes
- Regulation of
- Regulatory Role

Nucleic Acid Information
- DNA/RNA Sequence Features
- Mapping
- Nucleic Acid Interaction
- RNA Levels and Processing
- Transcription
- Translational Regulation

Protein Information
- Protein Physical Properties
- Protein Processing/Modification/Regulation
- Protein Sequence Features
- Protein-Nucleic Acid Interactions
- Protein-Protein Interactions
- Protein/Nucleic Acid Structure
- Substrates/Ligands/Cofactors

Related Genes/Proteins
- Cross-species Expression
- Disease Gene Related
- Fungal Related Genes/Proteins
- Non-Fungal Related Genes/Proteins

http://www.yeastgenome.org/help/Literature_Topics.html

Research Aids
- Other Features
- Strains/Constructs
- Techniques and Reagents

Genome-wide Analysis
- Comparative genomic hybridization
- Computational analysis
- Genomic co-immunoprecipitation study
- Genomic expression study
- Large-scale genetic interaction
- Large-scale phenotype analysis
- Other genomic analysis

Proteome-wide Analysis
- Large-scale protein detection
- Large-scale protein interaction
- Large-scale protein localization
- Large-scale protein modification
- Other large-scale proteomic analysis

Other Topics
- Evolution
- Industrial Applications
- Infection and Antifungals

Curated Literature
- Alias
- Archived Literature
- Reviews
- Selected Review
- List of all Curated References

Additional Information
- References Not Yet Curated
- Literature Curation Summary
- Gene Summary Paragraph
- PubMed Search
- Expanded PubMed Search
- All genome-wide analysis papers

Curators' Ph.D degrees (N=28)

Degrees	#	%
Genetics	8	28.6
Biochemistry	3	10.7
Molecular biology	3	10.7
Biology	2	7.1
Molecular, Cellular, and Development Biology	2	7.1
Veterinary Sciences	2	7.1
Bacteriology	1	3.6
Biophysics	1	3.6
Botany and molecular biology	1	3.6
Microbiology, molecular biology and genetics	1	3.6
Plant biochemistry	1	3.6
Plant biochemistry and molecular biology	1	3.6
Plant genetics	1	3.6
Virology	1	3.6

Poster available at: http://macmullen.com/conferences/MLA