Studying Biocuration Workflows

Gully Burns
Martin Krallinger
Kevin Cohen
Cathy Wu
Lynette Hirschman

Our Motivations

We ask the naïve question:

How is the process of Biocuration organized?

What are the similarities (and differences) in the workflow across different biological databases?

What are the main bottlenecks in the biocuration task?

Where might text-mining tools be most effective?

Gather information directly from curators

Why Workflows?

Capturing workflows as a set of informally-defined flow diagrams is a starting point for formalizing our understanding of the curation process

- Make different curation processes comparable through use of consistent reporting format
- Attempt to detect shared or distinct modules or approaches used by different groups
- Identify best practices, emphasize reproducibility
- possibly a good format for standardized documentation
- Facilitate training of new curators
- *Also, no-one's done it before (but Elixer?)*

Why Workflows?

Schrieber, G., *et al.* (2000), "Knowledge Engineering and Management. The CommonKADS Methodology", Cambridge, MA: MIT Press.

Our preliminary study

- We contacted a relatively small number of biocuration teams and constructed flow diagrams to depict their workflows.

- No formalization, entirely heuristic
- Quite slow,
 - interviewing biocurators was more efficient than attempting building workflows from written descriptions
- Seek to identify (through the use of a questionnaire):
 - specific points in the workflow with largest impacts on efficiency
 - possible ‘attack points’ for text-mining applications
 - common tasks shared between multiple systems
- Hope to provide useful catalogue resource for community.

Workflows included

- **Model Organism Databases**
 - The Arabidopsis Information Resource (TAIR)
 - Saccharomyces Genome Database (SGD)
- **Protein Databases**
 - BioGRID
 - Gallus Reactome
 - MINT
 - Protein Information Resource
- **Other**
 - Comparative Toxicogenomics Database (CTD)

We also gathered preliminary surveys from other groups, but were not able to factor them into the workflow study in time.

Example: Saccharomyces Genome Database (SGD)

This particular workflow has 3 stages

1. Automatic Scripts and Screeners

- MEDLINE queries and scripts
- Screeners performing triage on papers

2. 'LitGuide' process

- Categorization of papers according to types of curation required. Add annotations to articles

3. Detailed curation of Phenotype

- Note that we do not describe here other aspects (such as 'high priority curation', *etc.*)

Example: Saccharomyces Genome Database (SGD)

Example: Saccharomyces Genome Database (SGD)

- SGD automatic scripts + screeners
- Allows representation of concurrency ('forks'), decisions ('branches'), iterative steps and end-points

Example: Saccharomyces Genome Database (SGD)

Initial Assumptions

We assumed that literature-based workflows would have roughly the same structure

1. Article selection
2. Entity normalization / controlled vocabulary mapping
3. Relation extraction
4. Adding qualifying annotations

'Results'

Biocuration workflow catalogue

Model Organism Databases	2
The Arabidopsis Information Resource (TAIR)	2
Mouse Genome Informatics	4
Saccharomyces Genome Database	5
Protein Databases	8
BioGRID	8
Gallus Reactome	9
PPI	12
UniProtKB / SwissProt	14
Other	18
Comparative Toxicogenomics Database (CTD)	18

This document is a first attempt to develop an intuitive knowledge engineering methodology biocurators to describe their workflows in a generic, human readable format. This format may then be transcribed into a computational form. The basic idea is to encourage biocurators to describe a semi-formalized narrative where the inputs, process and outputs of each step of the workflow are explicitly described. This approach is intuitively based on the way that scientific protocols are described.

We use the outlining methodology of a word-processing document to provide the necessary indexing between parts of the workflow.

The crucial element to understand about this formulation that the workflow is (conceptually) a **bipartite graph** (a graph consisting of two types of nodes where the nodes of one type only connect to nodes of the other type), made up of **activities** and **data-objects**. These are described graphically in the following way:

Within the graphical workflows, if we were to include every input and output object of every activity, the graphs would become cluttered, overly busy and difficult to read. Therefore, the first time a data object is either created or used, it is represented explicitly. Unless a different object is created within the workflow (not if the current object is only modified, annotated or edited), we only represent the first object explicitly.

We use the framework provided by Activity Diagrams in the Universal Modeling Language to structure these workflows (see http://en.wikipedia.org/wiki/Activity_diagram for a preliminary introduction).

We have attempted to capture a (somewhat) representative example set of biocuration workflows as UML Activity Diagrams. These are not formal models, but provide a standard graphical format to help us understand commonalities between the biocuration tasks facing different groups.

- 20 page PDF 'workflow catalog'
- Difficult to interpret to provide new insight to text mining process.
- Personally found it to be a very valuable exercise
- Is it worth reiterating over process to build on this?

Workflow Study Summary

Database	Triage source	Entities	Relations	Qualifiers
TAIR	Medline	GO, PO, phenotype desc	Allele > [PO, GO, dev. stage]	???
SGD	Medline	allele, expt. type, mutant type, CHEBI, reporter	Allele > [all entities]	Yes
BioGRID	Medline + High throughput data	Protein names	PPI	'erroneous' codes
Gallus	Papers + Diagrams, Medline, OMIM	Based on Reactome's forms and tools	Based on Reactome's forms and tools	???
PPI / MINT	Medline under different approaches	UniProt/Genbank, PSI-MI	Interactor relations	???
PIR	UniProtKB / TrEMBL + BLAST etc.	UniProtKB / SwissProt records, family names, Protein Ontology	Feature lines, protein mods, interactions, etc.	???
CTD	Medline	Chemicals + genes	Chemical-Gene interactions	???

Acknowledgements

- **TAIR**
 - Eva Huala
 - Sue Rhee
- **SGD**
 - Eurie Hong
 - Rob Nash
- **CTD**
 - Allan Davis
- **PIR**
 - Cecilia Arighi
- **Reactome**
 - Peter D'Eustachio

‘Knowledge Engineering’

Knowledge Engineering is an engineering discipline that involves integrating knowledge into computer systems in order to solve complex problems normally requiring a high level of human expertise.

Feigenbaum, E., and P. McCorduck. (1983).