Appendix A: Workflows

We captured the structure and variability of the biocuration process by modeling the workflows themselves. We interviewed biocurators individually and manually constructed flow diagrams based on Activity Diagrams from the Universal Modeling Language (UML; see http://www.uml.org/). UML is typically used by computer programmers to describe the design of their software systems and was chosen since our underlying intention is to treat biocuration as a mechanical process that could be automated with text-mining technology.

The basic structure of a UML Activity Diagram is a flow diagram, with symbols representing data objects, activities executed on these objects and 'flow' elements. These flow elements include branch points (typically involving a decision point determines which subsequent path the process will follow), forks (where the control flow splits into parallel processing paths), and loops (where the flow iterates over a set of objects). Figure 1 below shows one such workflow that describes one of three staged procedures for curating information from the literature for the *Saccharomyces* Genome Database (SGD). The complete preliminary study is available in Burns et al., BioCuration Workflow Catalogue, Nature Precedings.[1] Additional analysis of the workflows is also available in two briefings from the workshop: "Studying Biocuration Workflows" by Burns et al.[2], and "A Framework for BioCuration Workflows (part II)" by Krallinger,[3] both in Nature Precedings.

The workflow is represented as 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:

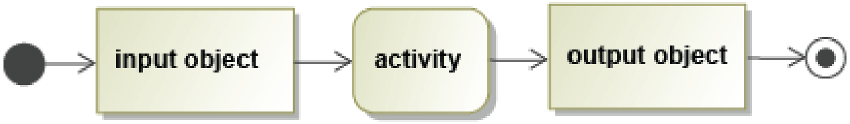

Figure A.1 Example UML graph

Figure A.2 illustrates part of the SGD curation workflow, as provided by the SGD team in 2009; the following text describes the inputs and outputs[4]. The text below describes the inputs and outputs.

---

[1] Burns, G. A. P. C., M. Krallinger, et al. (2009). Biocuration Workflow Catalogue - Text Mining for the Biocuration Workflow. 3rd International Biocurator Conference, Berlin. http://precedings.nature.com/documents/3250/version/1

[2] Burns,G.A.C., Krallinger,M. Cohen,K.B. Wu, C. and L. Hirschman(2009). Studying Biocuration Workflows. 3rd International Biocurator Conference, Berlin. http://precedings.nature.com/documents/3249/version/1

[3] M. Krallinger (2009). A Framework for BioCuration Workflows (part II). 3rd International Biocurator Conference, Berlin. http://precedings.nature.com/documents/3126/version/1

[4] This workflow has since been simplified; there is now a single script that looks for either the mention of 'cerevisiae' or 'yeast' in the abstract; all of these papers are then screened by a biocurator for relevance and associated with specific gene name(s) and given the status of 'not yet curated', as illustrated on the left side of figure A.2; personal communication, R. Nash.

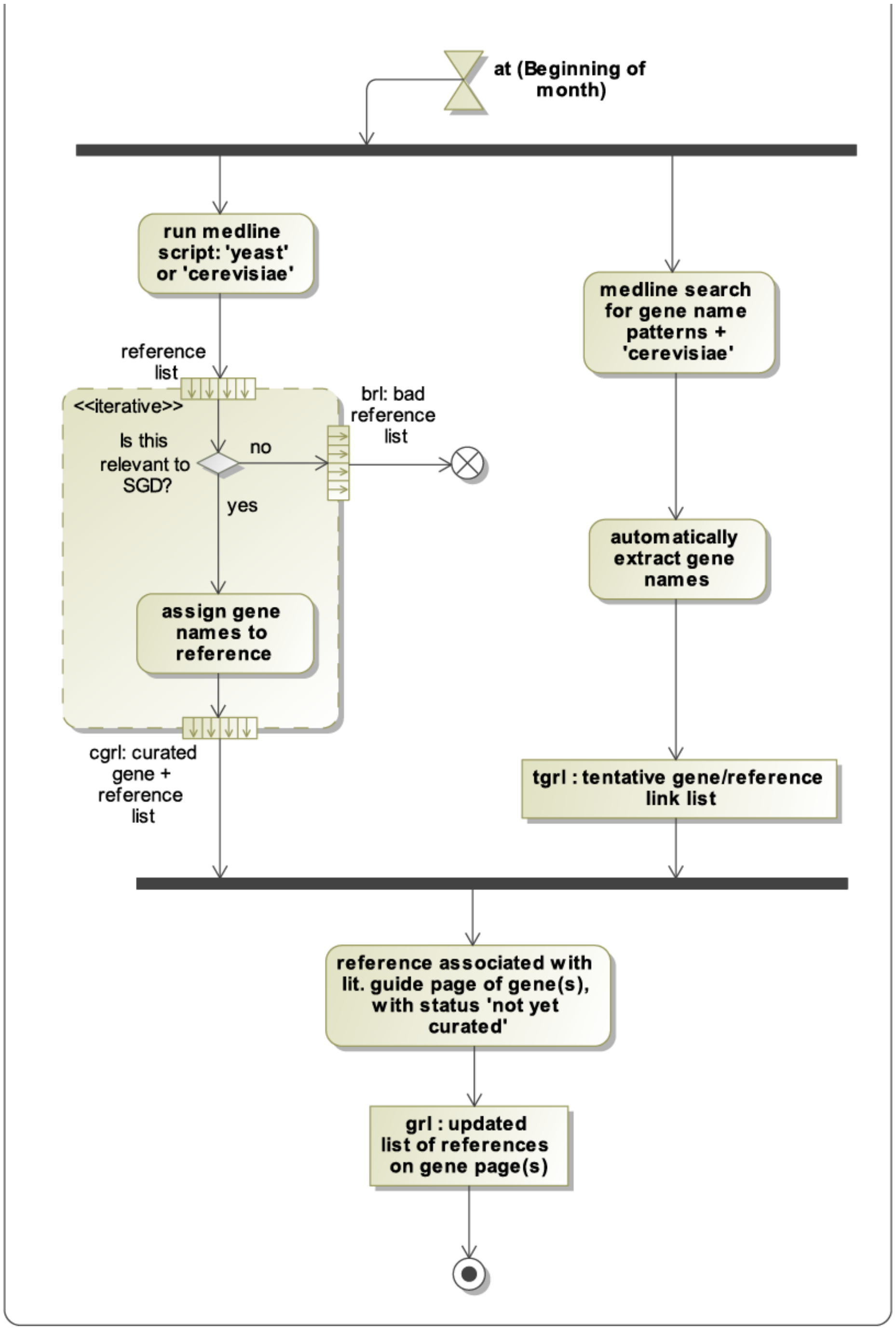
at (Beginning of month)
run medline script: 'yeast' or 'cerevisiae'
reference list
<<iterative>>
Is this relevant to SGD?
no
yes
brl: bad reference list
assign gene names to reference
cgrl: curated gene + reference list
medline search for gene name patterns + 'cerevisiae'
automatically extract gene names
tgrl : tentative gene/reference link list
reference associated with lit. guide page of gene(s), with status 'not yet curated'
grl : updated list of references on gene page(s)

Figure A.2: The flowchart for automated scripts that search Medline every month for relevant literature entries for SGD.

Description of the workflow in Figure A.2:

1. **Left column:**
   Activity: Run a MEDLINE script with the search term “yeast OR cerevisae”, have a biocurator check to see if the reference is relevant to SGD. If the reference is relevant,

keep it. If not, add it to bad reference list and go to the next reference on the list. For each reference on the list, assign gene names to the reference in the system and repeat until all references have been processed in this way.
Output: a list of all genes contained in all the references obtained with this script

2. **Right column:**
   Activity: Use available scripts from SGD to search MEDLINE for gene name patterns in conjunction with the term ‘cerevisae’. These patterns provide a method to extract gene names from the text to provide a tentative list of all gene names mentioned in the abstract.
   Output: a list of all genes contained in all the references obtained with this script
3. **Bottom:** lists of genes and references from (1) and (2)
   Activity: For each paper in the list, add each associated gene from the output of (1) and (2) to a global gene / reference list.
   Output: a list of all genes contained in the references obtained within this time period.

Typically, biocurators would describe multiple processes for different stages of their internal curation process. They would use internally-determined computational methods and data structures (such as priority lists, tagging schema and data forms) that were tailored to the specific needs of their own process. These different methods are highly variable, depending on the data structures and methods used in each curation process. For example, at Swiss-Prot, there exist priority lists for protein annotation. Therefore, queries are already more specific at the outset, since they will include gene/protein names together with the species. Text mining tools are then used afterwards to extract more specific information such as sequence variations/polymorphisms or post-translational modifications. Of course, understanding the workflow of each curated database is essential to develop effective text mining tools for that database.

There are many database-specific variations captured in these workflows. This raises the question of how to formalize the capture these workflows that would best serve the needs of the biocurators themselves. Comparing schema for different biocuration practices might allow biocuration teams to develop 'best practices' by examining other groups' methods in depth.

The results of this study were challenging to synthesize formally. We found that there were a number of similarities between Model Organism Databases (they all used a 'triage' step that involved searching the literature with tools like Medline before obtaining the full-text articles for closer scrutiny). This contrasted strongly with Pathway databases, such as the Gallus Reactome, which relied more heavily on the interpreted knowledge of biologists. In this case, models of reactions involved in a biological process were initially constructed with paper and pencil by experts before the literature was explicitly queried for justifying information.

Some recurrent tasks (that could possibly be automated with text mining tools) were seen in many workflow instances: (A) the need to identify species; (B) the 'triage' task of making a 'go / no-go' decision as whether (or not) to include a specific paper in the biocuration process for a specific database; (C) selecting appropriate terminology from large ontologies. More systematic, formal methods of representing curation workflows may yield more precise analytics, but they themselves would require the development of an appropriate ontology (such as the 'Ontology for Biomedical Investigation' or OBI) and an involved curation process to capture each workflow.