Biocurator Study: Question 3

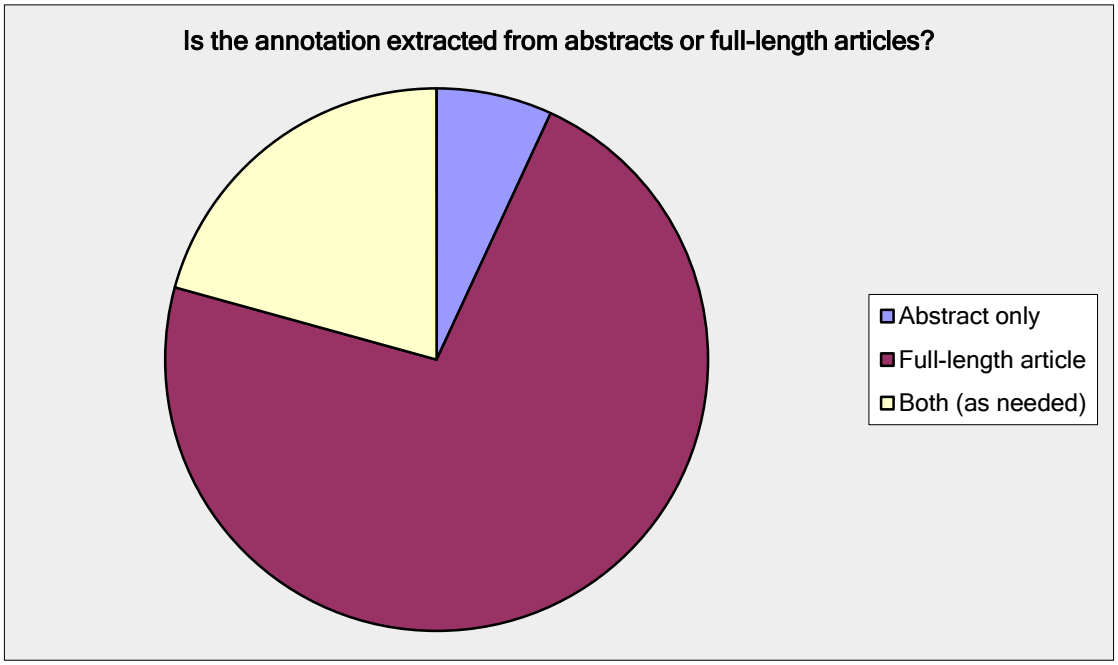
What are the criteria for selecting articles for curation? (Select ALL that apply)

Answer Options	Response Percent	Response Count
PubMed keyword search for a topic	48.3%	14
PubMed search for a specific bio-entity, i.e., protein/gene	58.6%	17
PubMed search based on a specific target model organism/taxonomy	75.9%	22
Literature search using text mining: if yes, please specify what tool(s)	34.5%	10
Exhaustive curation of specific journal(s); if yes, please provide top journals	17.2%	5
Curation based on citations derived from other databases, e.g. UniProt, GOA.	17.2%	5
References from an article that has been/ is being curated	41.4%	12
Provided by a user community (e.g., requests by users)	41.4%	12
Other (please specify)		14
	answered question	29
	skipped question	1

Other (please specify)
This is for the text mining tool; IHOP is used when trying to find literature about the interacting partners of a given protein.
Some existing tools have been tried about, but we have also developed our own text mining tools
We use Quosa to retrieve and screen articles (based on model organism terms) and assign specific journals to curators for review.
Method 1: Pubsearch and Textpresso used to associate abstracts and full text articles to a curated list of gene names, articles chosen for curation based on association to gene of interest. Method 2: Manual scanning of abstracts by curators to identify high priority articles (newly characterized genes)
We used an in house developed text-mining tool
Texpresso
New sequences at GenBank, for a specific organism AND attached to a pubmed article
Text mining provided by YYY tool developed at ZZZ
After a PubMed search for a XXX gene name and either the word W1 or W2, the hits are filtered against ZZZ's
Inhouse tool
Literature search using text mining - Textpresso
Journal of Biological Chemistry; Biochemistry; Biochemical Journal; Journal of Biochemistry
Some articles undergo 'first pass' curation during which data types/actions of importance are flagged. Papers with such flags are then prioritised for 'full curation'.

Biocurator Study: Question 4

Is the annotation extracted from abstracts or full-length articles?		
Answer Options	**Response Percent**	**Response Count**
Abstract only	6.9%	2
Full-length article	72.4%	21
Both (as needed)	20.7%	6
	answered question	29
	skipped question	1



Biocurator Study: Question 5

Do you curate additional data (select ALL that apply)

Answer Options	Response Percent	Response Count
Figures	95.8%	23
Tables	100.0%	24
If so, please describe the extra information that these sections provide and how this information is recorded in the database	22	22
	answered question	24
	skipped question	6

If so, please describe the extra information that these sections provide and how this information is recorded in the database

Many of the sequence data, such as accessions, positional and feature information are found in Figures or/and tables.
Our curation creates associations between proteins and functions backed up by literature references. We make no distinction as to where in the reference (text, figure, table) we found the evidence.
Captions often contain critical information necessary for gene indexing and other curation tasks.
Often the map positions, flanking markers, LOD scores, phenotype data etc etc are in the tables and figures as well as the genes and QTLs
Association of gene names to standard identifiers or sequence records (recorded as link between genomic sequence record and gene name), gene expression data, phenotype (we don't generally reinterpret the images themselves but do extract information from the text of figure legends and tables)
For figures, essentially the legends which sometimes provides interesting information. As well, data are displayed in tables, e.g. proteins identified by high throughput proteomics, enzyme kinetic parameters. More and more data are now provided in the supplementary material sections
Used only to support information extracted. That is we generally require that information is supported by data presented in figures and tables.
Figures and tables are not additional data for curating PPI. It is where the relevant information describing the interaction is reported. Even supplementary materials can fit the scope.
Interaction evidence is derived primarily from figures and tables
Any relevant data that we can capture.
We manually associate genes to papers. Often the correct identifiers/names are in tables; tables might also have data for GO that is not specifically mentioned in text. Same, but less so for figures.
Genetic map; orthologs in other organisms also characterized in literature
Figures and Tables provide additional functonal information that is not mentioned directly in the text.
We capture all interactions in the paper, figures, tables, and supplemental data
Values for kinetic data
Concentrations, buffer information, sometimes the kinetics data is written only there
Figures and Tables are often used to provide additional annotations, particularly for the Biological Process ontology. Not so much for Molecular Function and Cellular Component.
They often provide numbers that are added to CC and or FT lines in the database.
We look at the figures and tables to be sure that we agree with the author's conclusions. Occasionally author's overinterpret their results. We only annotate what there is actual evidence for.
I read the entire paper; figures, tables, etc. and use my understanding of the biology to choose the GO terms and evidence type appropriate.
They provide phenotypic data.
Specific genotypes are sometimes given in the figures/figure legends which is not found elsewhere. Data in tables is often not mentioned or described in full in the text. So, novel data in the figures and tables is curated in the same way as data in the full text - i.e., manually.

Biocurator Study: Question 6

How do you link a bio-entity from an abstract/full-length article to a database record? (What kind of attributes or contextual information do you use, e.g., name, symbol, organism source, molecular weight, and sequence length)

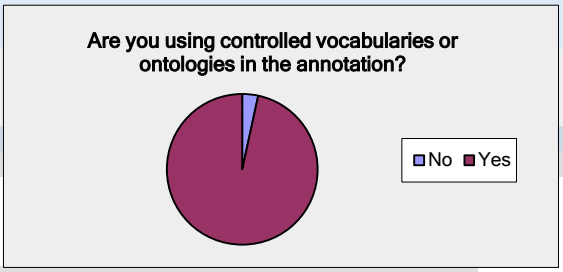
Answer Options	Response Count
	28
answered question	28
skipped question	2

Response Text
First, see if there is any database identifier that can be directly linked to the database, or sequence (sometimes run peptide mapping if a subsequence is shown). If not, then protein/gene name plus the organism is the second option. For splice variants the sequence length or some information about a sequence feature helps to determine the correct database entry ID.
Name, symbol, organism
We use all of this information as needed, plus additional information from articles cited by the current one, to make this link.
Usually linking is organism-specific and it is based on name and symbol information.
Symbol, ORG marker, journal ID (which links to PMID)
QTLs-flanking markers, statistics, strains and populations, traits, method; genes-GO, pathway ontology, disease, phenotype ontology, strains-origin, availability, derivation, characteristics, phenotypes etc.
First pass automated association based on string matching the entity name to the abstract or full text. For some categories (gene symbols) we manually verify the link correctness based on other words appearing in the abstract, sometimes also author names associated to a gene symbol
Usually by name, incuding symbol and synonyms. The locus name and the orf name are also very important to identify a newly characterized gene. Of course, the organism source is essential.
Primarily name, and organism source, as we focus on curation gene specific information from the XXX literature.
We use the Swiss-prot protein identifiers as main source.
Gene/protein names and symbols; if names are ambiguous or not found we use contextual information in text, identifiers or sequence information given in paper
In many cases we can use the gene or protein name or synonym. Occasionally we need to use protein and nucleotide sequence to map a genes describe in a paper to a gene on the genome.
Organism, gene names, identifiers
Symbol or name but these are often wrong! confirm identity by reading the article and/or sequence blasting
Public database accessions (if mentioned), otherwise gene/protein names, symbols, synonyms
I think the link is through the systematic name (YFL039C, e.g.), but the curator interface can accept all standard names (ACT1, e.g.), as long as there is not a nomenclature conflict; these are rare, and the interface asks for clarification.
Gene symbol/name, species
Name, organism
Varies depending on type of bio-entity
Full text
Name, gene symbol, and organism source.
OLN, gene and or protein name, organism, in some cases strain, PDB records,
See http://www.geneontology.org/GO.format.annotation.shtml for details
Name, symbol, Genbank ID
All annotation is ultimately linked to a gene object at ZZZ; SO: gene object gets GO terms assigned; the evidence supporting use of those terms is linked the the journal article.
Using the figure legend in addition to the paper ID.
Primarily via its current symbol (ie. the current, ZZZ-approved abbreviated form of a fullname). If the paper uses a non-current synonym or fullname for the entity, this is also recorded, but the link is still made via the current symbol.
Symbol

Biocurator Study: Question 7

Are you using controlled vocabularies or ontologies in the annotation?

Answer Options	Response Percent	Response Count
No	3.4%	1
Yes	96.6%	28
If so, please list the ones you use.		29
answered question		29
skipped question		1



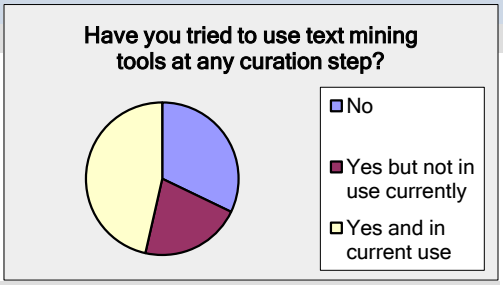
If so, please list the ones you use.

Gene ontology, Sequence ontology, Psi-MOD
GO
GO molecular function, biological process, cellular component; NCBI taxonomy for species names; ChEBI for names and family relationships of small molecules
Controlled vocabularies from major publicly available databases (e.g. BioCyc, UniProt, Entrez Gene, ChEBI and BRENDA). Also, some experiments are being made with Gene Regulation Ontology and Gene Ontology.
GO, EMAP, MA, MPATH, MP, DOID, EHDA, EHDAA
MP, PW, GO and we use MeSH for disease
Gene Ontology, Plant Ontology (covers anatomy and developmental stage)
Swiss-Prot keywords, in-house vocabulary lists, GO terms.
Gene ontology and phenotype annotations are based on ontologies in the case of GO and controlled vocabularies in the case of phenotypes.
PSI_MI-2.5
ZZZ evidence codes
GO; ORGXXX phenotype ontology (internal ontology- not in OBO), derived from the ORGXXX anatomy (that one is in OBO).
GO; Phenotype (our own); Controlled vocabularies to annotate strains
Plant ontology
Gene Ontology
Experimental systems are listed at web site . Genetic interactions use an old version of the ORGZZZ phenotype ontology.
Gene ontology, MeSH diseease terms, mammalian phenotype ontology
ChEBI, NCBI taxonomy, Gene Ontology, SBO
Inhouse (incorporates UMLS, MeSH, MedDRA, GO)
GO, SBO, internal standards, chebi, NCBI taxonomy
Gene Ontology
UniProtKB controlled vocabulary
Gene Ontology
Nomen guidelines, human orthology
Gene Ontology ; Anatomy (mouse adult and embryonic); cell type (CL_); Protein Ontology (PRO)
GO, PATO, various anatomical ontologies, cell type, chebi
GO, SO, ORGYYY anatomy (YYbt), YY development (YYdv), several miscellaneous CVs (e.g. publication type, allele class) contained within something we call 'YYcv'.
TO; GO; PO
GO; SO; ORG lifestage; ORG phenotype; ORG anatomy

Biocurator Study: Question 8

Have you tried to use text mining tools at any curation step?

Answer Options	Response Percent	Response Count
No	32.1%	9
Yes but not in use currently	21.4%	6
Yes and in current use	46.4%	13
Please specify any tools in use and those that have been tried		18
	answered question	28
	skipped question	2



Please specify any tools in use and those that have been tried

iHOP to mine for papers involving protein-protein interactions; RLIMSP to extract information about protein phosphorylation; Use rudimentary internal tool that finds and highlights exact matches for GO terms
XXX tool
ZZZ workbench (controlled vocabulary and rule based annotation)
Quosa, ProMiner, OBA, Protégé.
Textpresso
Pubsearch, textpresso
We informally use Textpresso to support our curation efforts. That is, when we are looking for specific information on strain backgrounds, mutant alleles or are just trying to identify a subset of information based on keywords or categories we use our in-house version of Textpresso.
Textpresso
Textpresso
Developing a text mining tool TTT in collaboration with researchers at UUU
I use Papers 9.1, which I find helps me speed up download of pdfs. It's also good for finding groups of similar articles, such as those with the word "complementation." I also use a couple of the search and highlighting features of Adobe Acrobat.
At the company where I had my previous curation position, we tried ReelTwo text mining software that featured machine learning to fine tune the searches. It seemed good at selecting abstracts featuring interacting proteins.
Inhouse tool
Textpresso SVMs for document classification
Currently in development
ZZZ, which was a European project we helped develop
We have a tool that parses abstracts that we have in the system and highlights GO term text matches. The abstracts are then "ranked" with number of matches to help us decide what papers to look at first. It is very crude.
Textpresso; custom scripts

Biocurator Study: Question 9

What would be (or is currently) the main usage for text mining tools? Please select ALL that apply:

Answer Options	No usage	Slight usage	Moderate usage	Heavy Usage	Rating Average	Response Count
Selection or prioritization of relevant articles for curation	4	4	5	14	3.07	27
Linkage of biological entities to existing resources (e.g., mapping protein	6	3	10	8	2.74	27
Selection of correct terms from a vocabulary/ontology (e.g., GO curation)	2	7	10	8	2.89	27
Extraction of relations among entities (e.g., identification of interacting	7	6	9	4	2.38	26
Generation of a human-readable summary of information	8	3	8	7	2.54	26
Identification of underlying evidence in the text for an annotation	6	2	10	9	2.81	27
Other (please elaborate)						5
					answered question	29
					skipped question	1

Other (please elaborate)

Submit a protein/gene name of interest in a given web interface, get ranking of articles based on the information content which could be based on categories (possibly customized: papers with disease terms, functional terms, post-translational modifications, PPIs). Be able to select papers (based on a category or combination of categories of interest), and see the summary table with relevant sentences, or extracted information along with PMIDs, and controlled vocabulary terms derived. Be able to save table in a format that could be used within the curation editor.
In particular, it is very important when working in metabolic reconstruction to provide the biologist with not only a comprehensible summary of the information, but also a navigable bibliographic network of all references that support the evidence that a given set of reactions is part of that metabolism. Gene coding, uptakes, external conditions, physiological states and so on, help constructing the "big picture", complementing/raising doubts about database contents.
We would love the following functions: 1) ranking of articles based on the 4 ontologies we use so that we could prioritize by the paper with information on more than one ontology; 2) a tool that would identify all new papers published about a gene and use the existing ontology annotations for that gene against the new paper to determine if there is new information in it - i.e., more granular than current annotations, a different branch of the ontology, better evidence than currently in the database etc
Note - I read very quickly, type poorly. So evaluating a text-mining annotation for accuracy and then saying yes/no for data entry would really speed things up if the text-mining output were tab-delimited.
MARKUP!!! Customizable color-coding of entities in pdfs so I can scan them quickly. Automated download would also help.

Biocurator Study: Question 10

How would you use text mining tools in curation? Please prioritize ALL that apply

Answer Options	Would not use	Would use occasionally	Would use moderately	Would use frequently	Would use all the time	Rating Average	Response Count
Batch processing (e.g., document ranking, or highlighting 'interesting' terms)	2	1	9	6	10	3.75	28
Interactive curation	2	1	8	8	6	3.60	25
Web-based	6	0	7	6	4	3.09	23
Please explain how these would fit into the curation workflow							18
						answered question	29
						skipped question	1

Please explain how these would fit into the curation workflow

Submit a protein/gene name of interest in a given web interface, ranking of articles based on the information content which could be based on categories (possibly customized: papers with disease terms, functional terms, post-translational modification, PPI). Be able to select papers , then have some summary table with
Text mining would be executed prior to human curation.
Batch processing is used for learning from manual relevance assessment and thus, providing biologists automatic assessments without requiring further information from them (explicitly!). Biologists decide which controlled vocabularies they want to use on their particular problems and are presented with a manual curation environment for annotation correction and vocabulary refinement. Every annotation is linked to the vocabulary source and, whenever possible, to Gene Ontology. The final results are presented in a comprehensible report that the biologist may use to perform reconstruction-oriented data integration.
Run batch processes on sets of journal articles, produce summary reports and mark-up articles online for curator review and evaluation. We view text mining as tools for curation, not replacement for human curators.
Step 1 - document ranking to prioritize curation efforts‰ Step 2 - curation of the document, preferably from a list of computationally suggested ontology terms, biological entities and experimental methods based on the document text. Our offsite curators depend on web-based tools to do their curation work and could be expanded to a community curation model
To be fully operational and well accepted by the curators, text mining tools should be fully integrated in the annotation platform. Batch procedures are useful for database updates and for defining annotation priorities.
We would be interested in text mining based tools that could aid in the identification of relevant literature, as well as mark-up of full-text literature. This might simply include the identification of all relevant genes/proteins in the article, or tools that help categorize papers based on our literature guide topics. At the more extreme end we would love to have tools that would allow us to more rapidly identify relevant sections of text that support potential gene ontology annotations, the curation of phenotypes (alleles, backgrounds, observations etc.), interactions (both protein and genetic). One other usage from the other direction would be text based tools that would allow use to link from an annotation back to the relevant passages in a paper that support that annotation.
Batch processing- automated updating of papers lists as new papers are published; ranking of most relevant documents to prioritize those most likely to contain interaction data‰ Interactive curation- e.g. gene names highlighted in paper could be clicked to reveal info form external sources to confirm identity; curation could be done on text of paper into the database directly‰ Web-based: ideally all tools would be web based.
We would use tools that help curation, i.e., help finding terms and evidence. We have not yet thought about prioritization because we manage to annotate most of the papers published every year (it's only about 200), but we'd like to go through our 'backlog' (4-5000 articles) and it would make sense to prioritize that if possible.
It would help if words are highlighted and then suggested GO terms appear. Of course really streamlined. Then we would need to create a tool that adds the term very easily to the database. A human readable short summary automatically created would help, even if this needs minimal editing at the end. The whole text mining should go along with more uniform writing of papers, so certain data is in certain 'fields'.
Articles mined for gene name/function and GO annotation. Currently we don't do GO annotation at ZZZ due to only having 1.25 curators.
I would like my text mining tools to highlight my keywords or categories of keywords in the pdf in customizable colors so I can quickly scan them.

The batch processing would be a first line operation in the workflow to select abstracts/articles for curation. It seems like a web-based text mining tool would not be fast enough for batch processing of thousands of abstracts. I'm not sure what you mean by "interactive curation".
I'm not clear on exactly what you mean by web-based, but we use Textpresso and SVMs to identify the most highly relevant documents for curation and additionally use Textpresso to retrieve individual sentences within documents that help to curate specific biological facts.
I always use PubMed searches to find articles, so this text mining tools would probably be used after a few searches to make sure I'm not missing anything. I wouldn't trust the tools not to miss something, so I'd still use PubMed searches.
I would be happy to give new tools a tryout and see whether they would make my curation tasks easier.
Not just for me: all papers we take into the system are first "indexed" to genes. This is how they appear on reports (new papers for gene x, etc.). A tool that would accurately identify the ORGXXX genes used in the paper and index them would be useful.
Tools would present curator with information; Tools would populate the database; Tools would alert curators to the presence of information

Biocurator Study: Question 11

What are the main obstacles in using text mining tools? Please rank in order of importance

Answer Options	1 (least important)	2	3	4	5 (most important)	Rating Average	Response Count
Tool not available for task of interest	5	2	4	2	9	3.36	22
Tool doesn't achieve sufficient recall (too many misses)	6	2	4	5	7	3.21	24
Tool doesn't achieve sufficient precision (too many false positives)	4	4	4	4	7	3.26	23
Tool doesn't have user-friendly interface	7	3	5	2	4	2.67	21
Tool is hard to integrate into existing workflow	2	3	8	4	7	3.46	24
Tool only works on abstracts (not full length articles)	1	1	6	3	13	4.08	24
Tool only works on ASCII or XML/HTML (not on other forms, e.g., pdf, Word)	2	2	8	5	5	3.41	22
Other (please elaborate)							6
						answered question	27
						skipped question	3

Other (please elaborate)

It is hard to find a tool that supports annotation not just for gene/protein mentions but also for compounds, organisms and other entities of interest in real-world scenarios.
We have not done much work to determine whether text mining tools would help our work. (Much of our curation is also with gene models.) We are a very small team (2 FT developers and 2 FT curators) and we have been very selective about deploying new software. We do have an ORGZZZ implementation of Textpresso at CalTech, but all the work was done on their end. We'd be happy to get in other collaborations of the sort!
Not sure, have not used any, though somewhat familiar with textpresso. But not used textpresso for curation so far.
Missed information and inaccuracy - since one would need to read the paper anyhow, and if most curators are like me, they read quickly with an eye for where needed information is lurking.
There was a tool mentioned at the Biocurator meeting that sounded really good, but I wasn't able to install it--too hard.
I haven't tried text mining in a few years so am not up-to-date on current tools

Biocurator Study:Question 12

What are the main bottlenecks in your current curation process and where could text mining tools help?

Answer Options	Response Count
	26
answered question	**26**
skipped question	**4**

Response Text

The main bottleneck is the time consuming step of extracting information about the different protein forms described in a paper and their associated attributes. A combination of text mining tools where for example, detection of protein-protein interaction combined with phosphorylation information (or other post translational modification) and highlighting of functional, and disease terms (with possibility to save all these info) will really improve the pace of our work.
Full length article text mining
The main bottleneck relevant to text mining is reliably finding papers that describe functional characterization of the human form of a protein, with isoforms (if relevant) clearly distinguished.
In genome-scale metabolic reconstruction, the major bottlenecks lay in integrating data from public databases and verifying its consistency (implicitly this implies checking out data references), as well as, getting additional data from literature.
Indexing articles that have been selected for curation during primary and secondary triage.
We curate by gene so the curators go through all ORGZZZ papers ever published on that gene - sometimes over 1000 and eliminate papers by title, then abstract and finally curate whole paper - would love to have useful ones identified automatically; secondly updating is difficult and would love the tool to automatically find new papers on previously curated genes and determine if the data is new or more specific.
Time required to find relevant text passages, map to correct ontology terms and find the relevant experimental method within the article being curated.
Finding papers of newly characterized proteins, extraction of high throughput data from tables or additional material, database update with new published information
Currently, bottlenecks include rapid identification of information supporting GO and phenotype annotations and the ability to convert the published findings into the appropriate annotations. So for example, text mining tools that could help us identify the relevant passages would help, as would a tool that could take this information and provide a short list of possible controlled vocabulary terms. Sometimes identifying the details is time consuming and text mining tools that could highlight such details (strain background, alleles, drug concentrations, and other experimental conditions) would be useful.
Protein identification is the major hurdle but available tools are not so effective
Sifting through papers which contain no interaction data (only 10-20% even after texpresso selection)
We curate strains and that's very difficult because the information can be spread among different sections of a paper.
It often takes too long to extract information, especially for strains in the ORGYYY literature. One paper with lots of data takes too long to curate - big backlog!

Assignment of correct symbols to loci and alleles. GO annotation is not done at all due to lack of time although we are working with external group to help. Text mining output in text-form (readable by Word/EXCEL/human would be very helpful.
Finding the appropriate literature to annotate when there is no standardized nomenclature for the species we annotate.
Basically, I have to process every pdf individually, and, if I save changes, I can't reverse them. Also, the program I use now has a lot of nice features, but it's buggy and can be slow; I often have to try to download pdf's multiple times before I'm successful, and I have to quit every now and then to repair the database or rebuild the keyword list.
The main bottleneck is selection of appropriate abstracts/papers to curate. A second bottleneck would be finding the exact phrases and sentences to support annotations.
Relevant information stored in full text. Information is scattered in the whole text (some data in material/methods and related information in results or discussion of the paper). Most of the kinetic data are stored in tables or figures.
Manual curation of text mining results
Curation efficiency - too much information and not enough time to curate it all. Any improvement to efficiency that text mining can provide is a big help.
The bottlenecks I face are time and choice of entries, text mining isn't going to help me prioritize these problems.
Identification of relevant papers, and identification of relevant terms from those papers.
Indexing: identification of ORGXXX genes that the paper is about.Main bottle neck after that: too few curators, too mnay papers.
Lack of integration between the tools. Even if an NLP tool worked with sufficient accuracy, how would I integrate with my current tools?
Too many papers on ORGYYY with too few curators! Probably try to curate too much data from each paper. Too many different tools (some of which are old and not user friendly) need to be used during curation. Text mining tools could potentially focus our curation efforts on the most important/valuable papers, and potentially speed up the 'full curation' process.
Timed batch searches

Biocurator Study: Question 13

What would you identify as the most pressing issue(s) for the text mining community to address that would help your curation process; please provide specifics in the context of your application needs.

Answer Options	Response Count
	25
answered question	25
skipped question	10

Response Text
I believe the most pressing need is to have a single interface that would allow the user to select a set of text mining tools to used, and get an integrated and user-friendly result. It would be great to have the option to enter a batch of PMIDs, to start from a protein/gene name, or even an identifier. Also, the need to use full length article; in our database, we are very interested in looking into post translational modifications and the position where these occur. This information is not always displayed in an abstract, and our current text mining tool cannot extract it.
Full length article text mining. For pathway curation text mining of results & figure legends would likely distinguish the new information provided in the article from prior knowledge.
User-friendly interfaces and development 'standards', i.e., some sort of development commitment that allowed common use of existing tools in real-world scenarios without time-consuming software refactoring. More efforts directed to real-world scenarios. Right now it is not so important to have algorithms with great precision and recall, but, in my opinion, it is necessary to start addressing issues that are relevant for biologists. For example, the identification of numeric parameters (e.g. experiment setup) is of most importance for most works. The ability to provide, at least, some contribution in this area may actually attract significant attention from biologists. The fact that identification would not be comprehensive at first could be accepted if carefully explained and user-friendly tools could be provided for additional curation.
Providing tools that support text mining of full length articles in the formats most publishers use for online dissemination (PDF and/or HTML). Or providing good, reliable utilities that convert articles from PDF/HTML to plain text for text mining applications.
Much of the work in textmining is not turned into a viable stand alone or web-based tool, what we really would like to be able to do is work with the text mining community to create/adapt tools for our needs
We need a tool that will computationally suggest gene ontology annotations based on article full text and will allow a curator to view the underlying text that the annotation is based on and either modify, accept or reject the suggested annotation. This would need to be integrated in a convenient interface allowing browsing of additional text passages and additional ontology terms as needed to complete the curation task.
Developing tools able to treat full-text articles in pdf format. Most of the curators are reading full-text papers and pre-annotated pdf documents for biological entities, species and other specific information, e.g. post-translational modifications, mutations - will be of tremendous help
Being able to process text and convert the multiple ways information can be expressed by scientists into a standard controlled vocabulary that could then be more quickly and accurately be entered into the database.
In the protein-protein interaction field, precision and recall should increase. Tools should be easy to incorporate in any curation pipeline.
Named entity recognition especially species-specificity- it is difficult to imagine this being overcome when it is often difficult for curators to decide which entity the author refers to; problem will more likely be solved by author providing identifier mark up.

I am not very familiar with text mining tools, but from my perspective curators are worried about low recall and error rates. Maybe this already exists but the tool should have at least two parts: the high confidence statement be processed (without to much curator reviews), and present the curator with the statements where the tool cannot make the judgment. This way, you use the expertise of the curator only in cases where that's necessary.
Work with journals to press for more uniform data; 'minable' fields, then create tool to extract this info at least into a first pass that then can be evaluated by a curator.
Really boring stuff such as outputs that can be edited by a human prior to loading into a DB. I am going to be reading the paper anyhow to be sure the meat was found, etc, but don't type and look up GO terms, etc quickly.
Ability to text mine full documents. I would like to see improved accuracy but I am not sure this will ever reach the level of human interpreation.
I don't need fancy artificial intelligence stuff like finding relationships. I just need an easy-to-use, easy-to-install program that downloads, scans, and marks up pdfs en masse.
The most pressing issue is accurate title/abstract/paper selection with minimal false positives. If software could select abstracts for a list of 1000 genes using a list of user-defined words/terms, a lot of time would be saved in the curation process.
PubMed search using full text paper and not only abstracts
Linguistic rather then simple pattern matching tools
Pressing journals for use of standardized nomenclature wherever possible, or marking up articles for known entities, such as gene and protein name. Assembling large, species-wide training sets for curation of various data types.
The ability to scan whole articles including figures and tables is essential, preferable in a pdf format but I could probably adapt to HTML. Note that I sometimes use older articles that are only available as non-searchable pdfs (at least by acrobat).
We often have to create new GO terms, and I don't think any text mining tool can help with that.
I have doubts that any text mining tool would be able to understand and correctly pick GO terms and correct evidence codes; you need someone who can think.
Ontology mark-up, particularly of figure legends and tables.
1. Identification and ranking of 'most important/valuable' papers, according to our own current (esoteric?) criteri; 2. Identification and automated curation of basic genetic entities, such as genes, so a paper can at least be indexed according to the genes it mentions.
Access to high quality xml or html of full text articles for every published paper