Quantifying literature citations, index terms, and Gene Ontology annotations in the *Saccharomyces* Genome Database to assess results-set clustering utility

W. John MacMullen
School of Information and Library Science, University of North Carolina at Chapel Hill 100 Manning Hall, CB#3360, Chapel Hill NC
27599-3360 macmw@ils.unc.edu

A set of 37,325 unique literature citations was identified from 120,078 literature-based annotations in the *Saccharomyces* Genome Database (SGD). The citations, gene products, and related Gene Ontology (GO) annotations were analyzed to quantify unique articles, journals, genes, and to rank by publication year, language, and GO term frequency. GO terms, MeSH indexing terms, MeSH Journal Descriptors, and SGD Literature Topics were quantified and analyzed to assess their potential utility for results set clustering. Results: Bradford's Law of Scattering was shown to hold for the citations, journals, gene products, and GO annotations. Only the MeSH terms and article title/abstract pairs had significant numbers of term co-occurrence. Multiple term types may be useful for faceted searching and clustered results set browsing if the strengths of each are leveraged.

Introduction

Life scientists use structured annotations to synthesize the vast amounts of knowledge available about biological organisms, and as a method for information overload reduction (MacMullen, 2005). 'Annotation' in this environment means linking an organism's primary data (e.g., gene sequences) to secondary sources, such as evidence of gene function extracted from experiments described in the scientific literature. Structure is applied through the use of standardized methods and vocabularies, such as the Gene Ontology (GO), to capture features of gene products, including function, process participation, and sub-cellular localization. The relatively recent availability of large-scale annotation of biological data in vertically-integrated, organism-specific databases provides interesting opportunities for information science research, because significant numbers of biological entities (such as genes and proteins) have been annotated in multiple organisms, drawing on tens of thousands of publications as evidence, and using common ontologies. The availability of this information in online public repositories enables both life- and information scientists to ask and answer questions that were, until recently, impractical, if not impossible.

This study explores the quantities and types of literature-oriented annotations in the *Saccharomyces* Genome Database (SGD), the data repository for the model organism *Saccharomyces cerevisiae*, commonly known as baker's yeast (Dwight, et al, 2004). First, descriptive statistics on the literature citations and Gene Ontology annotations are obtained and presented. Second, four types of indexing terms used to annotate the gene products in SGD are analyzed to assess their utility for results set clustering.

The significant growth of published scientific articles (now exceeding 16 million in the PubMed® database) often leads to extremely large results sets when scientists conduct searches in bibliographic databases and other repositories containing citation data, such as model organism databases. One general strategy for ameliorating information overload is minimizing the overall size of results sets through conventional information retrieval methods. These 'lossy' approaches assume that smaller is better, or that a single answer is desired, but as a result, many relevant articles could be overlooked. This study is interested in assessing the viability of clustering results sets by searcher-defined facets such as sub-domain or concept of interest. This concept is explored here by analyzing the similarity of four types of indexing terms used in annotations: GO terms, MeSH® indexing terms, MeSH Journal Descriptors, and SGD Literature Topics.

Methods

Selected data files from the SGD database were downloaded and processed as described below. The two files used for this study were from the Literature Curation directory of the SGD FTP site: the gene_literature.tab ('literature') and orf_geneontology.tab ('GO') files (Balakrishnan, et al., 2006). The literature file contains the citations to the scientific publications curated by SGD, the gene to which each publication is annotated, and the SGD-assigned literature topics (a high-level controlled vocabulary). The GO file contains data about the Gene Ontology (GO) terms annotated to each putative gene. The files are tab-delimited and contain multiple instances of non-unique values in several columns.

The two data files were imported into a purpose-built MySQL database. The literature file contained a total of 120,660 rows, and the GO file contained 5,806. The literature file is derived from the SGD production database, and concatenates individual components of article citations (e.g., authors, title) into a single field, which makes certain analyses (such as journal frequency counts) quite difficult. To acquire data that could be more easily analyzed, the unique PubMed IDs were extracted from the PMID column in the MySQL table and exported to a comma-delimited text file. This list of 37,325 citations was used as input for PubMed queries using the NCBI eFetch utility (NCBI, 2006). The PubMed results sets of full bibliographic citations were saved in MEDLINE structured text format, imported into the RefWorks bibliographic management software (RefWorks, 2006), and then exported as tab-delimited files. This was determined to be the most efficient method for transforming the MEDLINE-formatted files into a format that could be imported into a MySQL database. This process also enabled the acquisition of other relevant information in the full PubMed records, such as MeSH terms and publication language, neither of which are available in the SGD files.

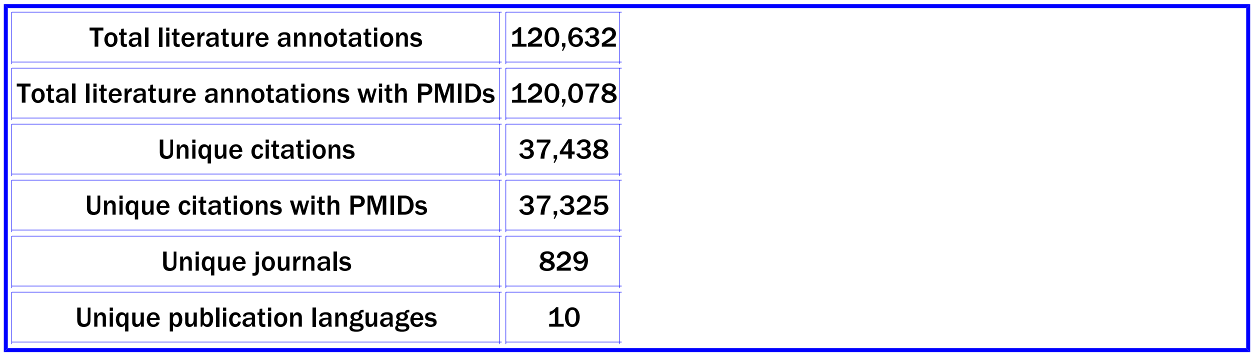
Frequency distributions of the journal titles, publication languages, and year of publication of the 37,325 unique citations, as described in the Results section, were obtained using MySQL queries. A sample of 10 rows from the set of unique citations was extracted by using the MySQL RAND() function as described in DuBois (2000:188) to provide one random PubMed ID and SGDID pair from each of the top 10 journals represented in the set of unique citations. This sample was used for a detailed examination of Gene Ontology (GO) terms, MeSH indexing terms, MeSH Journal Descriptors, and SGD Literature Topics associated with the genes underlying those annotations.

Results

Descriptive statistics of literature citations

This section and Table 1 provide descriptive statistics for the first SGD data file, gene_literature.tab, with additional data provided by the augmentation of the literature file by the manually-acquired citation information described above.

Table 1



Total literature annotations	120,632
Total literature annotations with PMIDs	120,078
Unique citations	37,438
Unique citations with PMIDs	37,325
Unique journals	829
Unique publication languages	10

The original SGD literature file had 554 citations (103 unique) lacking PMIDs (addressed later in this section), and 11 citations with invalid PMIDs. The valid PMIDs for those 11 citations were obtained from PubMed. It was further discovered that in 10 of the 11 cases, the same citations with correct PMIDs were already present in the data set. The 10 duplicate citations with incorrect PMIDs were removed, and the PMID of the one remaining citation was changed to the correct one, for a net decrease of 10.

SGD's approach to literature curation is cumulative, meaning it attempts to curate all yeast-related publications over time, and does not remove older publications when they are superceded by newer ones. Figure 1 shows the frequency distribution of the unique citations by publication year. Of the total 37,325 citations, 23,443(63%) were published within the past 10 years; 11,570 (31%) within the past 5 years. The numbers for 2006 include only those citations added to SGD prior to the release of the data on January 29, 2006.

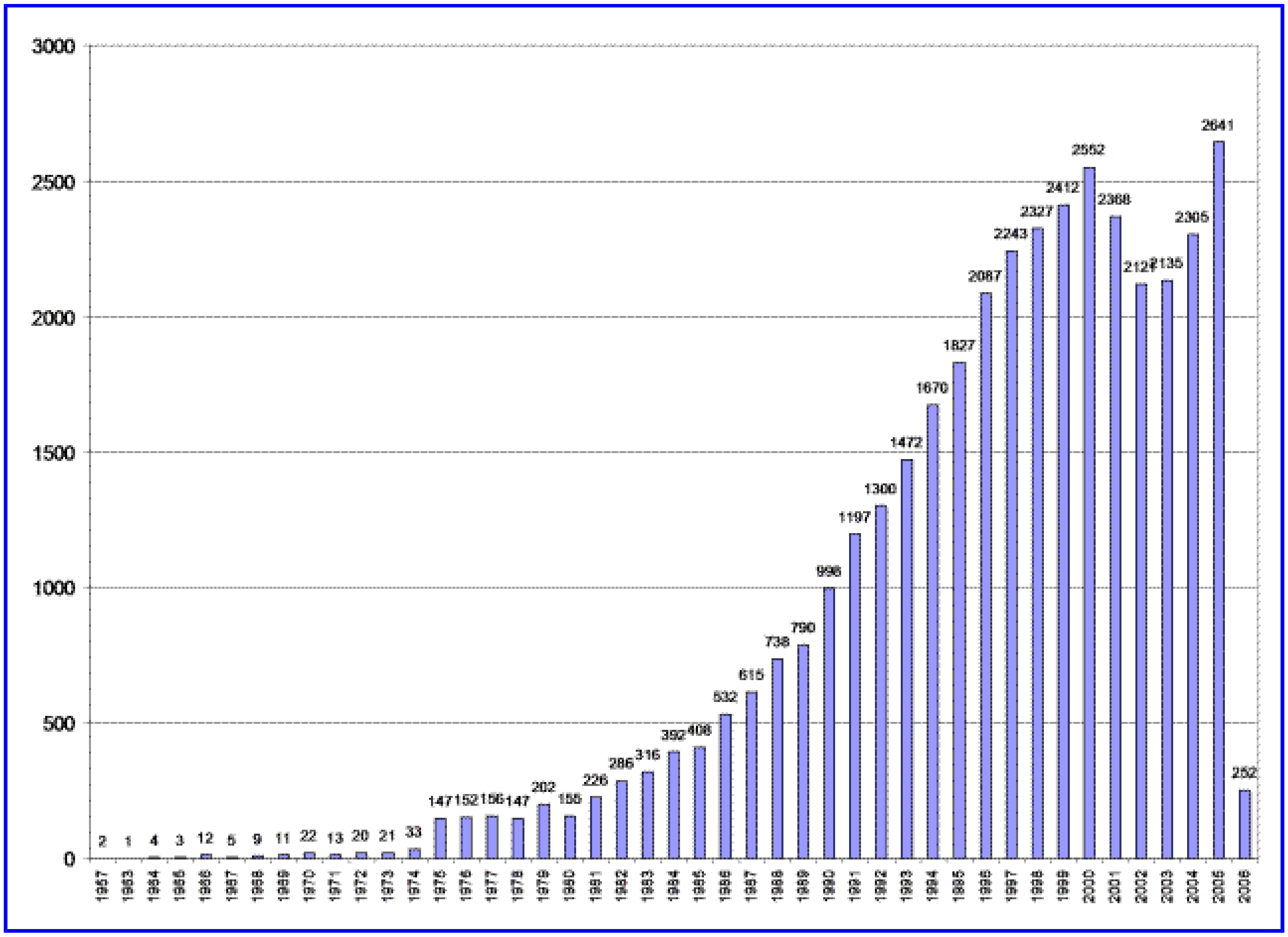


Figure 1. Frequency distribution of unique citations by publication year

Figure 2 provides the distribution of the unique citations by publication language. The truism that English is the language of science is borne out here, with only 1% of citations published in languages other than English.

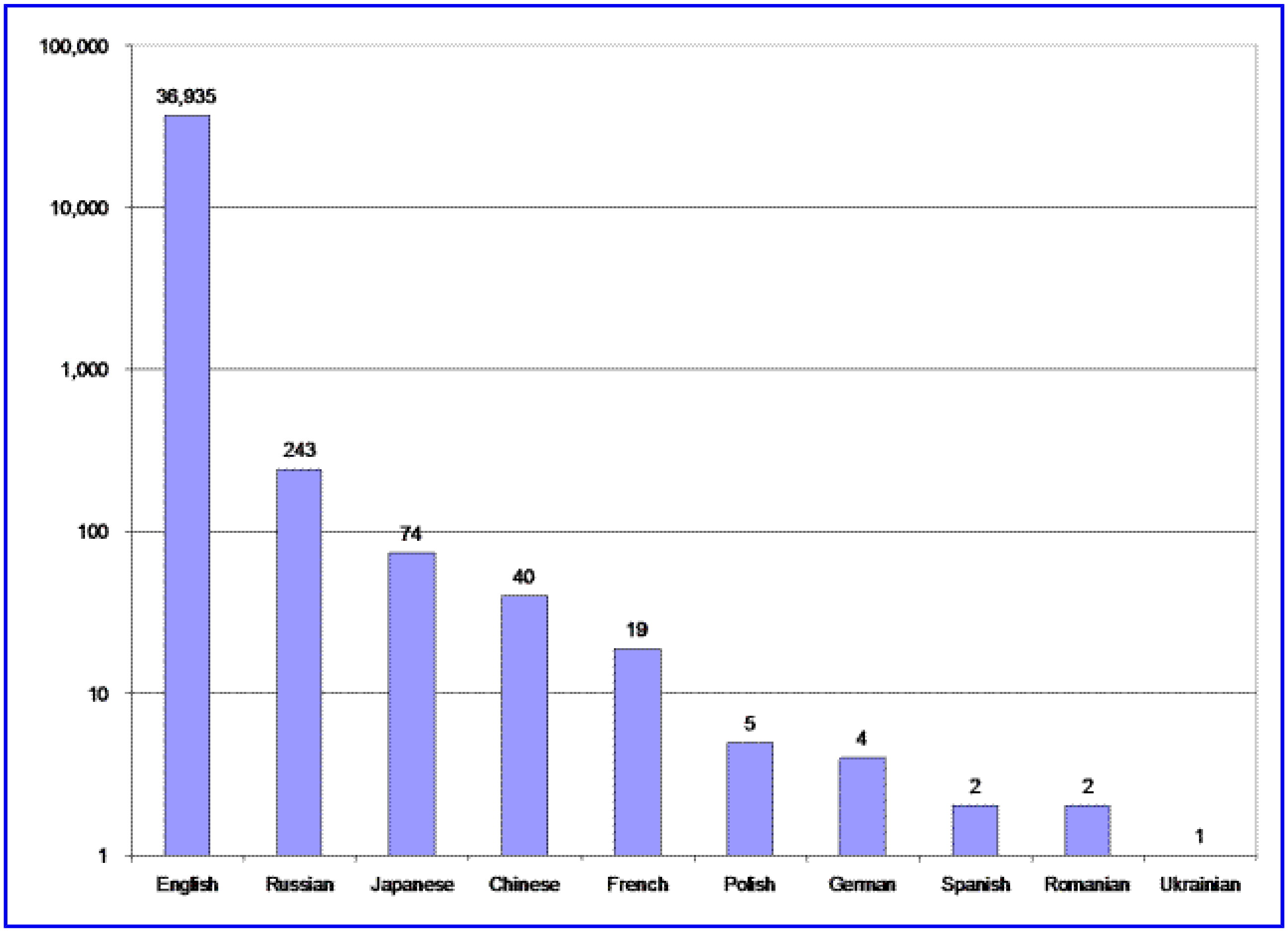


Figure 2. Frequency distribution of unique citations by publication language (log scale)

The 37,325 citations are drawn from 829 unique journals. Figure 3 shows the distribution of the citations by the top 25 journals, which account for a combined 70.7% of all citations.

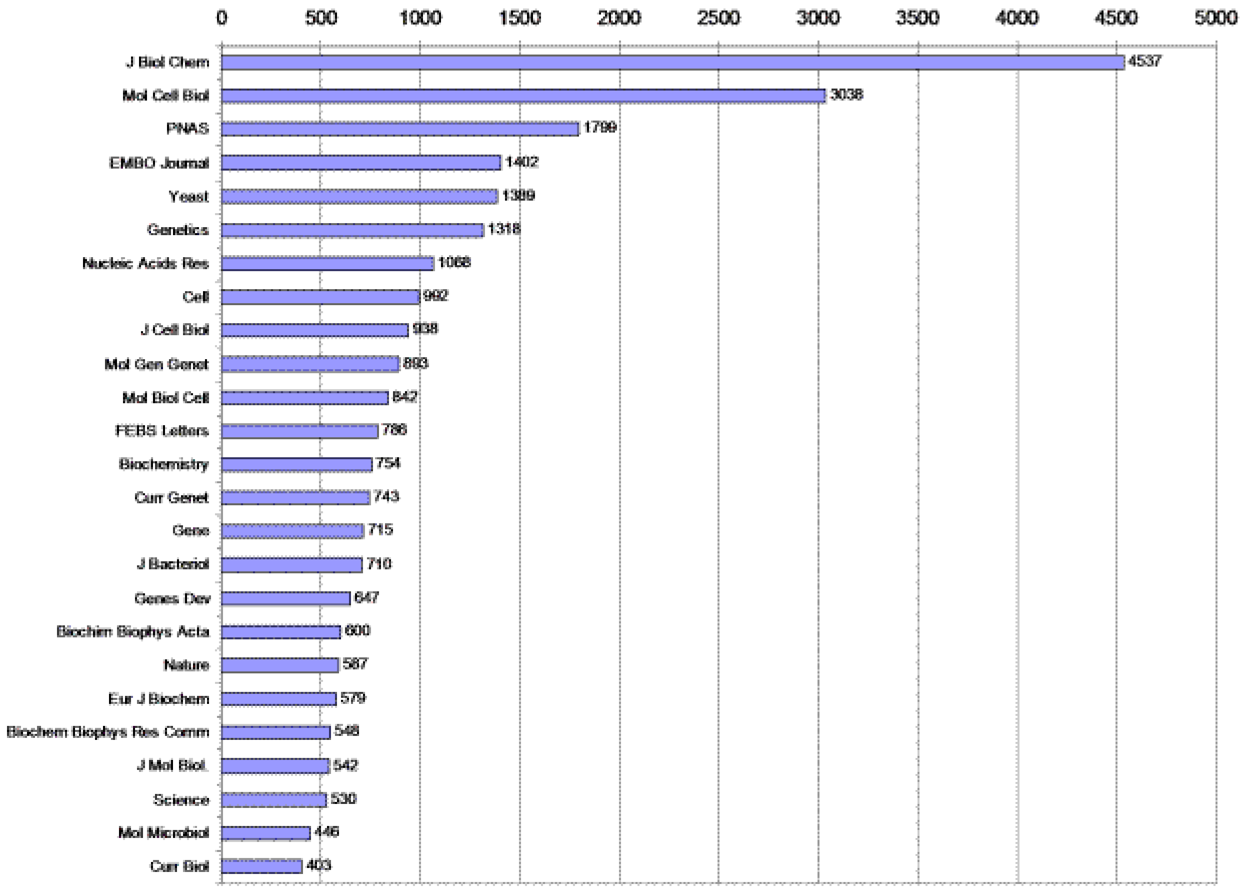


Figure 3. Frequency distribution of citations by top 25 journals

Table 2 illustrates that the distribution of journals and citations follow the variant of Pareto’s principle known as Bradford’s Law of Scattering (Bradford, 1948), which states that when a collection of citations or journals on a particular subject are divided in three equal groups, or ‘zones’, by percentage, the first third is composed of a small portion of the overall total, and the second and third grow exponentially larger. In this case, 2,434 of the 37,325 unique citations account for one third of all 120,078 annotations. Table 2 also shows that while 32.6% of citations are concentrated in 5 core journals, the overall literature is dispersed across a large number of journals.

Table 2

	Unique Citations				Unique Journals			
Zone	**Qty.**	**%**	**Cum.**	**%**	**Qty.**	**%**	**Cum.**	**%**
1	2,434	33.33	2,434	33.33	5	32.59	5	32.59
2	8,488	33.33	10,922	66.66	16	34.08	21	66.67
3	26,403	33.33	37,325	100.00	808	33.33	829	100.00

Figure 4 shows the frequency distribution of the top 10 most annotated citations. The top citation, ‘Global analysis of protein localization in budding yeast.’ (Huh, W.K., et al., 2003) is very recent, and accounts for 1.1% of the 120,632 annotations. The others in the top 10 account for a combined 2.6%, and all have been published within the past 10 years.

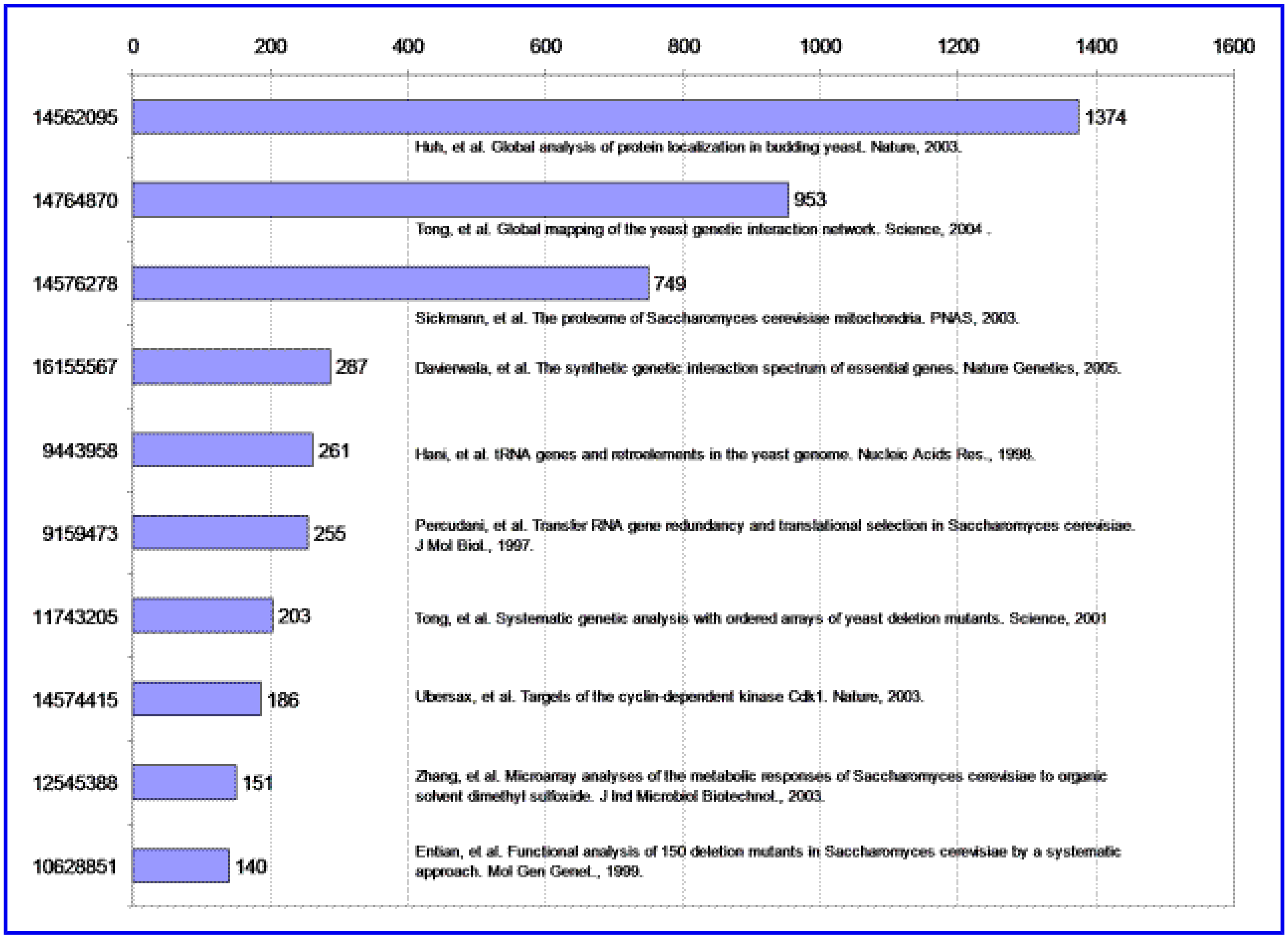


Figure 4. Frequency distribution of top 10 unique citations

A total of 103 unique citations in the original SGD literature file lacked PubMed IDs and thus were not included in the analysis above. These citations accounted for 554 annotations, with the top 5 unique representing 47% of that total. If these citations had PMIDs, they would not rank any higher than 56th in the main group of citations characterized above. Table 3 shows the frequency distribution and percent of the top 5. Four are book chapters, a type of document that is not indexed by PubMed, and the fifth was not found in PubMed, and is presumed to contain errors. A third potential reason citations may lack PMIDs is if the journals in which they appear are not indexed by MEDLINE.

Table 3

Citation	Freq	%
Hinnebusch, A. (1992). “General and Pathway-specific Regulatory Mechanisms…”. In *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, Cold Spring Harbor Laboratory Press.	78	14
Burge, C.B., et al. (1999). “Splicing of precursors to mRNAs by the spliceosomes.” In *The RNA World*, 2nd ed., Cold Spring Harbor Laboratory Press.	71	13
Mortimer, R.K. and Hawthorne, D.C. (1973). Genetic mapping in Saccharomyces…. Genetics 74:33-54 [incorrect citation; not found in PubMed]	40	7
Paltauf F, et al. (1992) “Regulation and compartmentalization of lipid synthesis in yeast.” In *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, Cold Spring Harbor Laboratory Press.	36	7
Wente SR, et al. (1997). “The nucleus and nucleocytoplasmic transport in Saccharomyces cerevisiae.” In *The Molecular and Cellular Biology of the Yeast Saccharomyces: Cell Cycle and Cell Biology*, Cold Spring Harbor Laboratory Press.	33	6
Total	258	47

The relationship of citations to genes is many-to-many: one paper can be about more than one gene, and one gene in SGD can have more than one citation annotated to it. Figure 5 shows the top 25 of 7,135 unique genes ranked by the number of articles annotated to them.

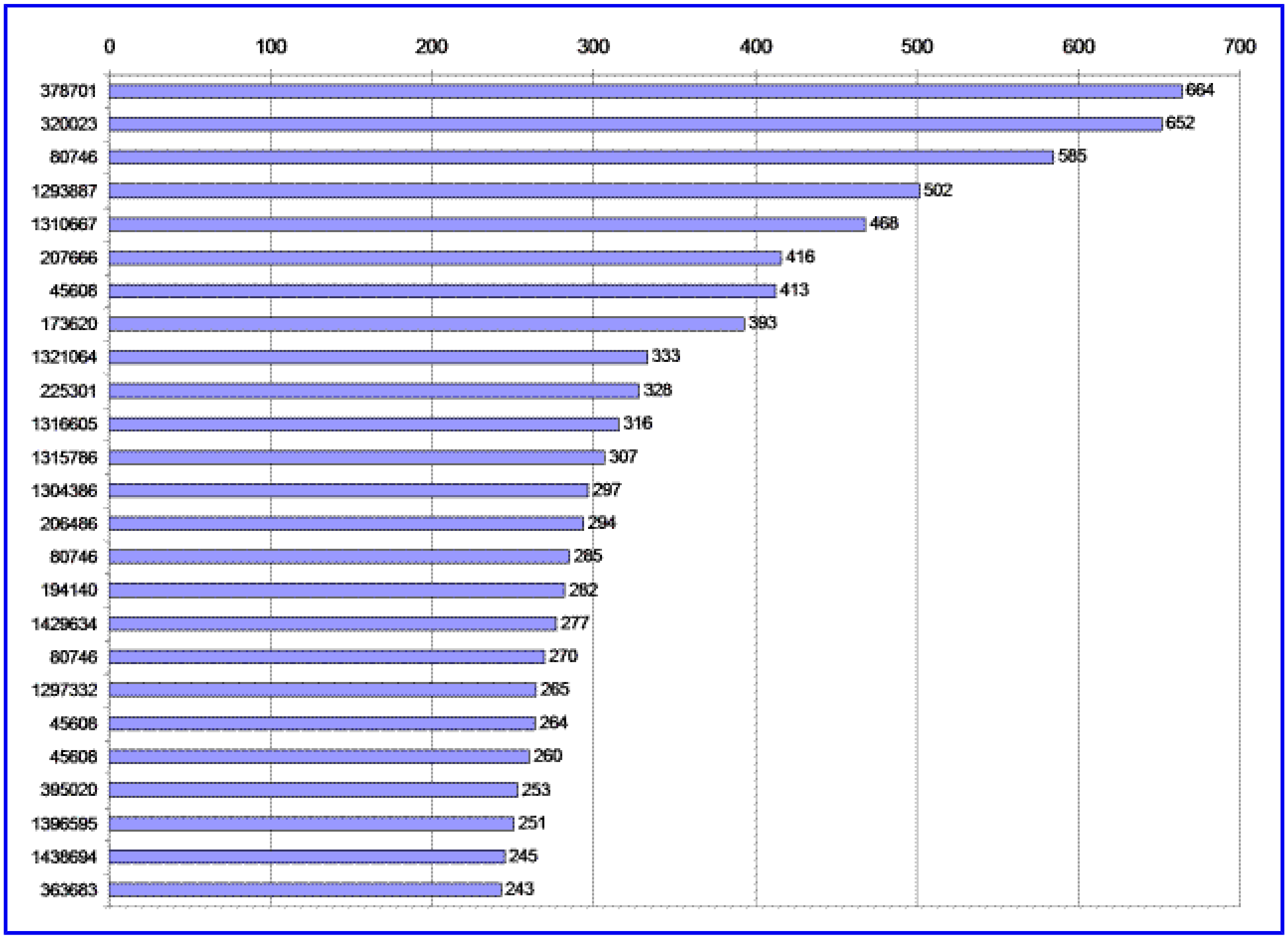


Figure 5. Frequency distribution of top 25 unique SGDIDs in literature file

Table 4 illustrates that the distribution of genes also follows Bradford's Law of Scattering. The top 268 genes account for one third of all annotations.

Table 4

	Unique SGDIDs			
Zone	**Qty.**	**%**	**Cum.**	**%**
1	268	33.33	268	33.33
2	901	33.32	1,169	66.65
3	5,966	33.35	7,135	100.00

Descriptive statistics of GO annotations

The second SGD data file analyzed was the orf_geneontology.tab file, which contains annotations that link individual yeast genes to specific SGD Literature Topics, and to Gene Ontology (GO) terms for molecular function, biological process, and cellular components. Table 5 shows the frequency counts for each type of GO annotation.

Table 5

	Qty.	%
Total annotations in orf_geneontology.tab	5,806	-
Total unique GO annotations	1,907	100.00
- Unique GO Molecular Function annotations	986	51.70
- Unique GO Biological Process annotations	672	35.24
- Unique GO Cellular Component annotations	249	13.06

Figure 6 shows the frequency distributions of the top 10 GO annotations by type. An annotation to GO is mandatory for each SGDID, so a special value of ‘unknown’ is available for cases where one or more of a gene’s function, process, or location in a cell is currently unknown. Clearly, the work of characterizing genes is still in the early stages, as the largest number of annotations for Molecular Function and Biological Process are to ‘unknown’, at 37% and 26%, respectively. The largest Cellular Component annotations are to Cytoplasm (22%) and ‘unknown’ (15%).

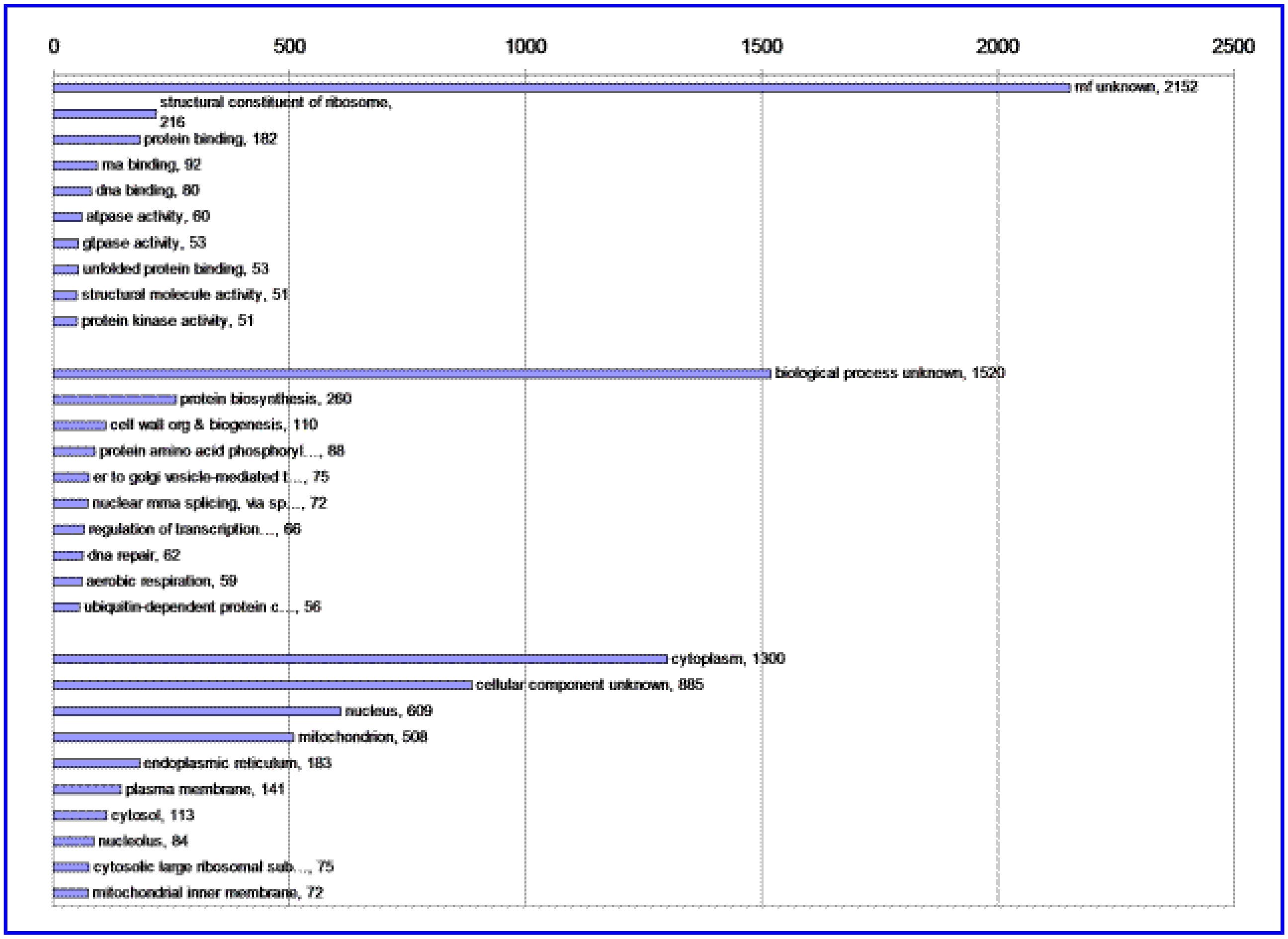


Figure 6. Frequency distribution of top 10 GO annotations by type

Table 6 shows that the distribution of the unique GO annotations follows Bradford's Law of Scattering.

Table 6

	Molecular Function				Biological Process				Cellular Component			
Zone	Qty.	%	Cum.	%	Qty.	%	Cum.	%	Qty.	%	Cum.	%
1	1	37.07	1	37.07	3	32.55	3	32.55	2	37.63	2	37.63
2	50	29.49	51	66.55	59	34.05	62	66.60	6	28.21	8	65.85
3	935	33.45	986	100	610	33.40	672	100	241	34.15	249	99.99

Index term analysis

For this portion of the study, four types of index terms associated with the random sample of 10 citations and genes from the top 10 journals (as described in Methods) were analyzed: MeSH terms, MeSH Journal Descriptors, GO terms, and SGD Literature Topics. The nature and use of these terms is described briefly below.

MeSH terms

The Medical Subject Headings (MeSH) controlled vocabulary from the National Library of Medicine (NLM) is used in pre-coordinate indexing of articles in MEDLINE. It is structured as a polyhierarchical tree and currently contains 22,997 descriptors and 83 topical qualifiers. In addition, some descriptors are designated as 'MeSH Major Topics' when that descriptor describes what a particular article is substantially about. MeSH terms for this data set were acquired through the process described in the Methods section.

Journal Descriptors

Bernhardt, et al. (2005) describe the idea of concept inheritance by individual articles of the set of 127 MeSH terms used to index their parent journals in MEDLINE. These terms, called MeSH Journal Descriptors (Humphrey, 1999), are used as broad classifications of the subject matter typically published by a journal, such as 'biochemistry' or 'cytology'. MeSH Journal Descriptors are annotated to the 'MeSH Subjects' and 'Other Subject(s)' fields of each journal title record in the NLM's LocatorPlus database. For the term analysis below, Journal Descriptors were collected manually using LocatorPlus.

GO terms

The Gene Ontology (GO) is a vocabulary system composed of three separately-rooted ontologies structured as directed acyclic graphs. The system is used by multiple model organism databases such as yeast, fruitfly, and mouse to facilitate cross-organism annotation of genes having similar functions, participating in similar biological processes, or located in the same parts of the cell. GO currently has in excess of 18,000 terms across the three ontologies, and more than 158,000 manually-curated gene products from multiple organisms (Gene Ontology Consortium, 2006). The GO terms for this analysis were collected from the SGD GO file (orf_geneontology.tab).

SGD Literature Topics

SGD has an internally-developed controlled vocabulary called Literature Topics (SGD, 2006) that is used to provide additional access points when searching the curated literature within SGD. The vocabulary contains 45 Topics in 9 categories. The terms for this analysis were collected from the SGD GO file.

Table 7 shows the number of terms annotated to each citation for Journal Descriptors, MeSH terms, and SGD Literature Topics. GO terms are not counted here because only the most frequently-used term from each ontology is represented in the gene_literature.tab file, so a count of 3 for each citation would be inaccurate. The number of Journal Descriptors per citation is very low, in line with the goal of providing a very high level categorization of 'aboutness'.

Table 7

PMID	Journal Descriptors	MeSH Terms	Literature Topics
1. 11959868	1	17	4
2. 7823940	2	14	4
3. 7017711	1	12	4
4. 7957107	2	26	1
5. 7502579	3	12	3
6. 7896088	1	25	3
7. 9461451	2	15	1
8. 9506516	1	20	3
9. 11266464	1	18	6
10. 10394911	2	20	6
Mode	1	20	3,4
Range	2	14	5

A matrix of term co-occurrence was compiled using the four index term types and the full text of the title and abstract of each article, as retrieved from PubMed. Table 8 shows the counts of exact term matches, not conceptual relationships, even if the terms had similar roots (e.g., 'Cytology' as a Journal Descriptor was not considered a match with the GO term 'cytoplasm', because the former is a practice specialty and the latter is a location in a cell). The only synonyms allowed were 'yeast' for 'Saccharomyces'.

The highest incidence of term co-occurrence was between MeSH terms and the title/abstract text, while the lowest was Journal Descriptors, irrespective of the other term types. A single Journal Descriptor was present in each of 5 cases when paired with MeSH terms, and only once did a Descriptor appear in an abstract.

Table 8

	Journal Descriptors					MeSH terms					GO terms					SGD Literature Topics					Title/Abstract				
Citation	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
JDs	-	-	-	-	-																				
MeSH	0	1	0	0	1	-	-	-	-	-															
GO	0	0	0	0	0	1	0	0	1	1	-	-	-	-	-										
SGD-LT	0	0	0	0	0	1	3	0	0	1	0	0	0	0	0	-	-	-	-	-					
Abstract	0	0	0	0	1	5	6	7	5	3	1	0	0	1	0	0	1	0	0	1	-	-	-	-	-

Discussion

Descriptive citation analysis

The concept of multi-faceted term profiles of articles using all four index term types to aid in searching and browsing results sets seems reasonable if the strengths of each type are leveraged. Each may facilitate different types of results set clusters - by discipline, by methods or work tasks, or by gene- or organism-specific features, for example. Other existing facets could assist in this process. One approach to clustering might be to group journals by whether they are ‘clinical’ or ‘basic science’ in their coverage. This decision involves subjectivity, as there is no classification for this in MEDLINE records. Bernhardt, et al. (2005) attempt to automate this to a certain degree using MeSH Journal Descriptors inherited by articles from the journals in which they appear, but their work is focused on differentiating clinical specialties, rather than a clinical/basic science division. MEDLINE records do have a ‘Publication Type’ field (e.g., review, editorial, letter) that is another possible clustering facet. These facets types are not mutually exclusive in terms of use in interface design; multiple facets could be employed to enable clustering by single facets of interest or in combination.

The journal frequency rankings and Bradford analysis may be useful to biomedical librarians who need to identify core journals in the area of yeast genetics for collection management purposes. This approach has been employed by Schloman (1997) and others to analyze the literatures of allied health.

Limitations

This is preliminary work that has certain limitations that inhibit generalization to SGD as a whole and to other model organism databases. Only a small number of citations' index terms were analyzed, and the distribution of citations to genes as shown in Figure 5 means that a random sample may not be a representative one. The very small sample used for index term analysis is probably also not representative.

Future Work

To be more informative, the index term analysis requires a larger-scale automated investigation of cross-type index term co-occurrence that includes more sophisticated linguistic analysis, including synonyms and truncations to normalize inexact matches across the term types, instead of exact match only. While the sample selected here was random, a purposive sample may be required to adjust for over-representation of certain genes. While this work focused on intra-citation term analysis, future work cluster-oriented work might explore the differential effectiveness of clustering based on each index type for a defined document corpus, in terms of utility for different browsing perspectives.

In comparison with the other data above, it would also be interesting to consider GO evidence codes, which provide each annotation with information about how the underlying evidence was derived (such as experiment type), as a possible clustering facet. This information is in the SGD database, but was not available in the GO data file used in this study. Visualizing the distribution of GO annotations in relation to the overall GO vocabularies may indicate clusters of annotation, or gaps in knowledge.

To make the data manipulation process more amenable to automation for future analysis, it would be desirable to acquire citation data from SGD in a non-concatenated format to avoid having to transform PMIDs into full citations from PubMed and then parse them from MEDLINE format to comma-delimited files. However, if analysis of associated MeSH terms or other fields is desired, these would still need to be acquired from PubMed. An open-source script or parser to transform PubMed MEDLINE- or XML-formatted files into relational database-friendly structures is needed from NLM or a third party.

Supplementary Data and Code Access

The data files, MySQL table definitions and queries, eFetch queries, and other utilities used in this study are available for use by request. A supplementary table listing the index term data from which Tables 7 and 8 were derived is available online:

http://macmullen.com/conferences/asist/asist06-s1.pdf .

Acknowledgements

The author was supported by a research fellowship from the Annotation of Structured Data project at the School of Information and Library Science at the University of North Carolina, Chapel Hill, which was funded by a gift from Microsoft Research. Gary Marchionini provided thoughtful comments on an early draft. This study would not have been possible without the prior work of the biological curators at SGD.

References

Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dolinski, K., Dwight, et al. (2006) *Saccharomyces Genome Database* Data files dated January 29, 2006. Available ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/(Accessed: 2006-02-13.)

Bernhardt, P.J., Humphrey, S.M., & Rindflesch, T.C. (2005) Determining Prominent Subdomains in Medicine *Proceedings of the 2005 American Medical Informatics Association (AMIA) Annual Symposium* 46-50

Bradford, S.C. (1948) *Documentation* London: Crosby Lockwood

DuBois, Paul (2000) *MySQL* Indianapolis, IN: New Riders. p. 188

Dwight, S.S., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dolinski, K., et al. (2004) Saccharomyces Genome Database: Underlying principles and organisation *Briefings in Bioinformatics* 5(1):9-22

Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006 *Nucleic Acids Research* 34: D322-D326. PMID: 16381878

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., et al. (2003) Global analysis of protein localization in budding yeast *Nature* 425(6959):686-691. PMID: 14562095

Humphrey S.M. (1999) Automatic indexing of documents from journal descriptors: A preliminary investigation *Journal of the American Society for Information Science* 50(8):661-674

MacMullen, W.J. (2005) Inter-database annotation linkages in model organism databases *Proceedings of the 68th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*

NCBI (2006) NCBI eFetch utility Available http://eutils.ncbi.nlm.nih.gov/entrez/eutils/(Accessed: 2006-02-13.)

RefWorks (2006) RefWorks Web-Based Bibliographic Management Software, January 2006 release Licensed to the University of North Carolina, Chapel Hill. Available http://refworks.com/(Accessed: 2006-02-13.)

Schloman B.F. (1997) Mapping the literature of allied health: project overview *Bulletin of the Medical Library Assoc* 85(3):271-277. PMID: 9285127

SGD (2006) *Saccaromyces Genome Database (SGD) Literature Topics* http://www.yeastgenome.org/help/Literature_Topics.html(Accessed 2006-02-13.)