Pacific Symposium on Biocomputing 13:592-603(2008)

INFORMATION NEEDS AND THE ROLE OF TEXT MINING IN DRUG DEVELOPMENT

PHOEBE M. ROBERTS, WILLIAM S. HAYES

Library and Literature Informatics, Biogen Idec, Inc. Cambridge MA USA

Drug development generates information needs from groups throughout a company. Knowing where to look for high-quality information is essential for minimizing costs and remaining competitive. Using 1131 research requests that came to our library between 2001 and 2007, we show that drugs, diseases, and genes/proteins are the most frequently searched subjects, and journal articles, patents, and competitive intelligence literature are the most frequently consulted textual resources.

1. Introduction

Academic research and pharmaceutical research share some common objectives, but there are important differences that influence publishing trends and information needs. Both groups rely heavily on peer-reviewed publications as a source of high-quality information used to formulate hypotheses, design experiments, and interpret results. To remain competitive, both groups must stay abreast of recent developments in order to make informed decisions. Effective search and retrieval is essential for finding high-quality information, which often benefits from integration and visualization due to the sheer volume of information that is available.

Unlike academic biomedical research, where publishing peer-reviewed articles is tied closely to funding, for-profit biomedical research groups are under different constraints. In the competitive marketplace, publishing information can alert competitors to developmental advances. Release of public information, however, is not always avoidable. Drug developers must file applications and data packages with drug approval authorities whose guidelines differ from country to country. Portions of drug application packages are freely available as unstructured text. In addition, drug developers are beholden to patent-granting authorities, filing patents to protect intellectual property and any profits that result from it. This makes legal literature a rich source of early-stage drug discovery information [1]. Publicly traded companies are required by the Securities Exchange Commission to disclose changes in their drug pipeline that have a potential financial impact, all of which are publicly available through the EDGAR database (http://www.sec.gov/edgar.shtml). Conversely, there are times when companies want to make advances known. Publicly traded companies hoping to boost stock prices, or private companies hoping to raise financing, use press releases, industry analyst conferences, and major scientific meetings attended by prescribing physicians to announce advances in their drug pipeline. It is critical to track all of these information resources to stay abreast of competition and spot potential collaborators, and the value of this information is reflected in the success of commercial "competitive intelligence" databases that integrate information in a structured searchable format [2].

Text mining is often raised as an antidote to the exponential expansion of published literature [3, 4]. Instead of relying on one or two keywords to find abstracts and full-text papers, text mining allows more powerful relevance ranking using classification and clustering techniques or class-based searching using entity tagging. Entity extraction adds additional value by structuring unstructured text and generating lists of like items that can be visualized in other ways, allowing the forest to emerge from the trees.

If one were to examine real user information needs, what kinds of questions would benefit from text mining applications? Studies of internet search, and biomedical literature search in particular, indicate that queries tend to be made up of only one or two keywords [5, 6]. Surprisingly, only 1.6% of PubMed queries used the Boolean OR operator [6]. Does this indicate that broadening searches is not important, or does it reflect a lack of familiarity with advanced search capabilities?

One way to understand the potential role of text mining in drug development research is to examine real end-user information needs instead of the terms used to conduct the searches. We describe here classes of queries submitted to the Library and Literature Informatics group at Biogen Idec, a large biotechnology company. The results highlight the entities of greatest value to drug development, and they place in context the utility of peer-reviewed literature versus other information resources.

2. Methods and Results

2.1 Coding Drug Company Research Requests by Subject and Resource

Biogen Idec is the third largest biotechnology company in the world, with strong franchises in multiple sclerosis (MS) and oncology. Historically, Biogen Idec has specialized in developing therapeutic antibodies and biologics, two of which have achieved "blockbuster" status (sales of over a billion dollars a year). The Biogen Idec Library and Literature Informatics group receives requests for research assistance for all aspects of drug development, including research, development, manufacturing, marketing, sales, and post-launch safety. The Library has cataloged 1131 research requests and their results since 2001. This database contains requests for research assistance only. Other Library functions, such as journal article requests or book orders, are not included.

Because of the competitive nature of drug development and the proprietary nature of the research requests, actual user needs will not be explicitly stated here. Instead, we sought a simple classification scheme that would allow us to unambiguously classify queries while maintaining enough information to be valuable to the information retrieval community, even in the absence of user queries. Taxonomies to classify queries have been described for questions asked by clinicians, resulting in an elaborate taxonomy of 64 question types [7]. To simplify our taxonomy, we chose to create controlled vocabularies that captured the main subject(s) of the request (Table 1). Subjects were selected based on their prevalence in the research questions, and questions were coded with as many subjects as applied. Also noted was the resource (e.g. patents, competitive intelligence resources, or journal articles) that was either specified by the requestor or deemed by the information professional to be the best resource for the question (Table 2).

To evaluate the terminologies and their consistent use, both authors (who annotated the full query set) independently coded approximately one-tenth (n=100) of the queries with the controlled vocabulary Subject, Resource, and Text Mining terms shown in Tables 1, 2, and 5 (results are shown in the last column of each table). Interannotator agreement was calculated as the ratio of matches between annotated requests and all requests annotated positively for a specific controlled vocabulary term by either annotator.

Table 1. Information Need Subject Classes

Subject	Description	# requests	Interannotator Agreement (# of Matches)
Drug	Substance administered to humans or animals to reduce or cure disease	355	.82 (46)
Disease	Human disorder or animal model of human disorder. Includes adverse drug reactions.	310	.78 (47)
Gene (includes Protein)	Biological substance that can be mapped to a discrete genetic locus. May be target of a drug.	297	.65 (20)
Company	Institution, public or private, industrial or academic	192	.59 (26)
Methods	Protocols for conducting scientific experiments or administering treatment	120	.47 (9)
Author	Individual who publishes or patents information	89	.70 (7)
Geography	A country or region	64	.62 (5)
Sales/Pricing	Income from or cost of a marketed drug	57	.54 (7)

General	Topics that do not map to subjects above	148	.44 (4)

Table 2. Information Resource Classes

Resource	Description	# requests	Interannotator Agreement (# of Matches)
Journal articles	Scientific literature from biomedical journals	389	.70 (32)
Competitive intelligence resources	Databases and trade publications that draw on company websites, SEC filings, scientific meetings and press releases for information about drugs in development	344	.77 (27)
Patents	Legal literature from worldwide patent agencies	217	.84 (16)
News sources	Newspapers and magazines (not specific to the pharmaceutical industry)	74	.71 (5)
Health statistics resources	Incidence and prevalence of diseases	59	.44 (4)
Other	Sources that do not map to information resources above	123	.29 (4)

Frequently occurring representative queries based on actual user needs are shown in Table 3, illustrating how the controlled vocabularies were applied to categorize query types. Note that the Subject terms were applied to both the input and output of the research request, i.e. the subject of the question, as well as the desired answer. When subject classes were not explicitly stated in the query, they were inferred during query coding based on implicit reference to the subject type. For example, the question, “What’s in Phase II for arthritis?” mentions disease as a subject, and drug is inferred. Company information and the gene or gene product targeted by the drug are also provided in the interest of completeness. In our experience, providing drug information in the absence of manufacturer (Company) and mechanism of action (Gene) prompts follow-up requests for that information. Furthermore, by limiting subjects only to those explicitly stated would understate the frequency at which relationships between entities are of interest (see Section 2.2, Table 4). Including subjects from the question and the answer regardless of whether they are explicitly stated impacted interannotator agreement for the Company and Gene subjects, which were most frequently inferred (data not shown).

Table 3. Representative Queries

Representative Query	Subject Terms	Resource Terms	# results
What drugs are in development to treat multiple sclerosis?	company, disease, drug, gene	Competitive intelligence	138
What companies have drugs in Phase II to treat multiple sclerosis, and what	company, disease,	Competitive intelligence	52

are the drugs?	drug, gene		
What patents have been published about TNF-alpha?	company, gene	Patents	49
In what tissues is TNF-alpha expressed?	gene	Journal articles	6
What protocols have been patented for producing large quantities of therapeutic antibodies? By what companies?	methods, company	Patents	4

2.2 Query Analysis

Requests were classified as “navigational” (directed toward a specific piece of information) or “informational” (collecting data about a topic) [8]. Typical navigational queries included information about a patent family, sales figures for a drug, or a recent news article about the pharmaceutical industry. Navigational queries made up 20.2% (228/1131) of research requests. This is lower than the 25.6% mark noted for PubMed queries [6], and it may reflect differences in query analysis methodology, or in how users employ the services of PubMed versus a corporate library. Interannotator agreements for “navigational” and “informational” queries were .37 (10) and .79 (70) respectively.

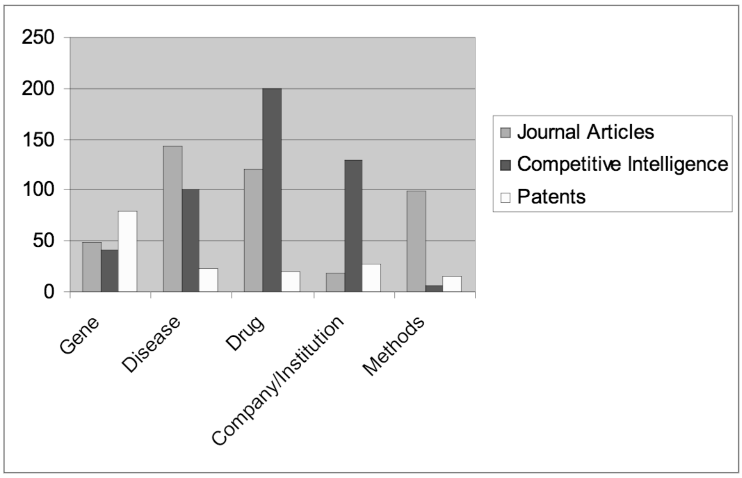
Questions about drugs, diseases and genes made up the largest class of search requests, representing 31.4% (355/1131), 27.4% (310/1131) and 26.2% (297/1131) of all queries, respectively (Table 1). The first two classes are not surprising when the corporate mission is to create drugs to treat diseases. Gene-based queries are also to be expected, considering that genes and proteins are the targets of drugs, and they provide the key to understanding origins of disease and the mechanism of therapeutic action. Consistent with how authors refer to genes and proteins in the literature [9], Biogen Idec employees favored the long names or synonyms of genes rather than using the official gene symbol the vast majority of the time (data not shown).

Journal articles were the most frequently requested resource type, followed by Competitive Intelligence resources, Patent resources and, to a lesser extent, News. Most competitive intelligence questions could be answered by using commercial databases such as Pharmaprojects (http://www.pharmaprojects.com) or the Investigational Drugs Database (IDdb; http://www.iddb.com), which periodically survey corporate websites, press releases, major conferences, and Securities and Exchange Commission reports (complete listing at http://www.iddb.com/cds/faqs_info_sources.htm) (data not shown). Competitive Intelligence databases also include selected information from journal articles and patents, blurring the lines between our Resource definitions

(Table 2), but they do not constitute enough of the database content to impact our results.

To determine if query topics vary by resource, search subjects from journal articles, competitive intelligence resources, and patents were examined individually (Figure 1). Gene and protein names are common search terms across different resource types, and they are the preferred search subjects in the patent literature. Disease and drug searches are directed primarily to the scientific literature and pipeline databases. Company and Institution queries are largely confined to the competitive intelligence literature, and methods searches are limited to journal articles.

Figure 1. Query Subject by Resource



Compound queries, in which more than one subject is represented in the question and/or answer, represented 36.2% (409/1139) of research questions, four examples of which are shown in Table 3. These questions demonstrate the importance of identifying relationships among entity types. Questions requesting information from multiple resources occur in 6.4% (73/1131) of requests. These require answers that involve some degree of data integration, whether it is combining unstructured text from news and journal articles, or merging structured data with unstructured text. This figure is a gross underestimation of data integration requirements, as most journal article, competitive intelligence and patent searches generate results from more than one database [10]. Merging results into a unique set involves extensive post-processing to remove duplicate records, map controlled vocabularies from each database, and apply a uniform format to records from disparate databases.

Table 4. Frequency of Pair-wise Subject Combinations

	Drug	Disease	Gene	Company	Methods	Author	Geography
Drug							
Disease	138						
Gene	22	29					
Company	52	13	13				
Methods	10	10	6	4			
Author	8	12	11	17	0		
Geography	22	23	5	11	1	5	
Sales	41	18	3	7	1	1	14

2.3 Where Does Text Mining Fit In?

Cohen and Hersh define text mining first by distinguishing it from information retrieval, text summarization and natural language processing, then by sub-dividing it into named entity recognition (NER), text classification, synonym and abbreviation extraction, relationship extraction and hypothesis generation [3]. Synonym and abbreviation extraction can be grouped with NER if one assumes that synonyms and abbreviations for each entity are part of the entity extraction process. Similarly, relationship extraction is dependent on NER as a means of identifying which entity classes are related. If the extraction techniques are grouped with NER, that leaves three criteria with which to evaluate the Biogen Idec Library research requests for text mining: extraction, text classification, and hypothesis generation.

A research request was classified as being an Extraction request if the question asked for specific facts ("what are annual sales in Japan?" or "what is the incidence of disease x?"), versus asking for a general search ("please search the patent literature", "I need general information about this disease"). Text Classification was used to describe requests for which large positive training corpora exist. Theoretically, classification can include automated techniques such as unsupervised clustering, which can be applied to all the research requests. Our objective with this category was to quantify the frequency of requests for queries that are executed weekly or monthly over a period of several years, and for which positive training data exist, thereby justifying the effort of building a classifier. A prominent example is product safety literature. The FDA mandates periodic comprehensive literature searches for reports of marketed products in the literature (21 CFR 314.80[1]), which generates a positive training set of documents that can be used to build a classifier. Hypothesis Generation was not used to code the queries, as discussion between the

[1] http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.80

annotators did not result in a viable protocol for annotation into this proposed category.

Out of the 1131 queries, 304 (26.9%) were classified as Extraction (286/304) or Classification (18/304). Search requests not coded as Extraction (73.1%) typically were at the general search level, suggesting that requesters were conducting a broad search, they wanted context around the facts they were looking for, or they were unaware that entity extraction tools are available.

We examined the queries coded as Extraction further to determine if individual Subjects or Resources were over-represented. The majority of extraction research requests called upon Competitive Intelligence (189/286) or Statistics (53/286) resources (data not shown). Interestingly, the answers for these requests were available in proprietary databases such as IDdb, Adis R&D, and others. Extraction questions not answered using databases were spread across subject categories, with journal articles as the primary Resource type (63/286 queries; data not shown).

Table 5. Text Mining Techniques

Technique	Description	# Requests	Interjudge Agreement
Extraction	Named entity recognition, synonym and abbreviation extraction and relationship extraction	286	.60 (31)
Classification	Text Classification – supervised machine learning	18	.50 (2) * *[n=229]

3. Discussion

3.1. *Impact of Assistance on Research Requests*

Information needs have been studied by examining query logs of search engines and inferring the intended need based on query terms and user sessions [5, 6]. Other studies have gathered information needs directly from clinicians [7] or academic and industry researchers [11]. Our study differs in that the information needs represent questions that require professional assistance, i.e. end-users were not able to find results on their own or they could not find results efficiently. This may be influenced by the query subject; gene and protein names are notoriously difficult to use as search terms due to complicated nomenclature and ambiguity [9]. Drugs also undergo name changes as they traverse the developmental pipeline [12]. Diseases are represented in myriad ways as observed in the Medical Subject Headings terminology. In the absence of a sophisticated indexing and query translation system like the one behind PubMed (http://www.pubmed.org), the low frequency of Boolean OR operator use [6] suggests end-users are missing relevant results, prompting them to seek assistance. Variations in search engine algorithms, database design, and content may also place a naïve end-user at a disadvantage. Even though Competitive Intelligence and Patent end-user tools are available at Biogen Idec, the high frequency of requests for assistance suggests that they are too complex for the casual user to efficiently obtain information.

3.2. *Research Request Subjects and Resources: Why Are Questions Asked?*

A frequently cited application of text mining is database curation; e.g. the extraction of gene names, protein-protein interactions, expression data, and subcellular localization. The predominant subjects in the Biogen Idec research requests overlap with entity types frequently studied in text mining research, notably genes and diseases. Our results support the selection of tasks in text mining challenges such as BioCreAtIvE and the TREC Genomics track as representing real information needs, especially named entity recognition of gene and protein names. Genes were the only subject type of interest across resource types (Figure 1), which may reflect the need to understand gene function throughout the drug development process. Selection of a protein as a drug target requires understanding what it does (a journal article search) and who else is working on it (competitive intelligence and patent searches). As named entity recognition of gene names improves, our results suggest that testing algorithms against multiple text sources is a worthwhile endeavor.

Genes were the primary search subject of patent literature, which was unexpected considering that patents are a significant source of drug development information, especially small molecules and their chemical synthesis [1, 13]. The dearth of patent drug searches in our results is due to chemical structure searches being performed by groups outside the Library who do not need our assistance.

Information about drugs is the most common request subject (Table 1). The high cost of drug development makes awareness of research with comparable compounds essential for maximizing efficacy and minimizing unintended adverse effects. Although named entity recognition of chemical compounds has received some attention in the text mining literature [14], to our knowledge, a broader approach to identify any substance with therapeutic benefit has not. In particular, therapeutics for a specific disease (138/378; Table 4) or against a class of targets (represented by drug-gene compound queries, 22/378, Table 4) are of sufficiently high interest to drive Biogen Idec employees to seek assistance.

Searches about companies or institutions were enhanced in the competitive intelligence literature (Figure 1). One reason for this phenomenon may be the ease with which institution searches can be performed against databases that house journal articles and patents. The second reason reflects the fundamental *raison d'etre* of competitive intelligence literature: to find out what other companies are doing.

3.3. *Existing Databases and Entity Extraction*

The Biogen Idec Library does not typically receive requests to interpret results from transcript profiling or proteomics experiments. There are a number of public and proprietary databases that address these needs, providing extracted entities and relationships among them based on the published literature. Numerous public and proprietary databases permit high-throughput analysis of gene lists and extraction of relationships between genes and diseases, expression patterns, or Gene Ontology terms.

Similarly, in the competitive intelligence space, so-called "pipeline databases" allow users to search by and export lists of drugs, indications (i.e. diseases treatable by drugs), companies, and developmental stages [15]. The success of these databases highlights the importance of entity extraction as a means of managing the vast amount of information available. Furthermore, our quantification supports the need for these resources. Literature and competitive intelligence queries are well-served by existing databases. Patent literature, however, is underserved in this regard. The high incidence of patent gene queries illustrates the need for a reliable and comprehensive resource with extracted information about genes or proteins and their patented use. To some extent, GeneSeq and GeneIT perform this task by isolating nucleotide and amino acid sequences, but not all patents about specific targets contain sequences.

3.4. *Requests in the Future*

The Library tends to receive queries that can be answered, consistent with results from analyzing questions asked by clinicians [7]. To add qualitatively new query types to the ones currently serviced requires training and awareness. New queries resulting in new deliverables often require changing customer behavior to take advantage of new capabilities. An example is inferential analysis, which uses indirect relationships to generate or validate hypotheses. Examples of inferential analysis have been described in the literature [16, 17], but demand for this technique has not surfaced in research requests to our library. The Biogen Idec customer base is increasingly aware of inferential analysis as the tools to service those requests are being deployed and the customer base learns what qualitatively new requests will result in answers.

Acknowledgments

The authors thank Suzanne Szak, Pam Gollis and Lulu Chen for critical reading of the manuscript.

References

1. Grandjean, N., et al., *Competitive intelligence and patent analysis in drug discovery: Mining the competitive knowledge bases and patents.* Drug Discovery Today: Technologies, 2005. **2**(3): p. 211-215.
2. Carlucci, S., A. Page, and D. Finegold, *The role of competitive intelligence in biotech startups (Reprinted from Building a Business section of the Bioentrepreneur web portal).* Nat Biotechnol, 2005. **23**(5): p. 525-527.
3. Cohen, A.M. and W.R. Hersh, *A survey of current work in biomedical text mining.* Brief Bioinform, 2005. **6**(1): p. 57-71.
4. Scherf, M., A. Epple, and T. Werner, *The next generation of literature analysis: integration of genomic analysis into text mining.* Brief Bioinform, 2005. **6**(3): p. 287-97.
5. Chau, M., X. Fang, and O.R.L. Sheng, *Analysis of the query logs of a web site search engine.* J Am Soc Inf Sci Technol, 2005. **56**(13): p. 1363-1376.
6. Herskovic, J.R., et al., *A day in the life of PubMed: Analysis of a typical day's query log.* J Am Med Inf Assoc, 2007. **14**(2): p. 212-220.
7. Ely, J.W., et al., *A taxonomy of generic clinical questions: classification study.* British Medical Journal, 2000. **321**(7258): p. 429-32.
8. Broder, A., *A taxonomy of web search.* SIGIR Forum, 2002. **36**: p. 3-10.
9. Chen, L., H. Liu, and C. Friedman, *Gene name ambiguity of eukaryotic nomenclatures.* Bioinformatics, 2005. **21**(2): p. 248-56.
10. Biarez, O., et al., *Comparison and evaluation of nine bibliographic databases concerning adverse drug reactions.* Dicp, 1991. **25**(10): p. 1062-5.

11. Stevens, R., et al., *A classification of tasks in bioinformatics.* Bioinformatics, 2001. **17**(2): p. 180-8.

12. Snow, B., *Drug Nomenclature and Its Relationship to Scientific Communication*, in *Drug Information: A Guide to Current Resources*, B. Snow, Editor. 1999, Medical Library Association and The Scarecrow Press, Inc.: Lanham, Maryland and London, England. p. 7-19.

13. Simmons, E.S., *Prior art searching in the preparation of pharmaceutical patent applications.* Drug Discov Today, 1998. **3**(2): p. 52-60.

14. Mika, S. and B. Rost, *Protein names precisely peeled off free text.* Bioinformatics, 2004. **20 Suppl 1**: p. i241-7.

15. Mullen, A., M. Blunck, and K.E. Moller, *Comparison of some major information resources in pharmaceutical competitor tracking.* Drug Discov Today, 1997. **2**(5): p. 179-186.

16. Wren, J.D., et al., *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-98.

17. Swanson, D.R., *Medical literature as a potential source of new knowledge.* Bull Med Libr Assoc, 1990. **78**(1): p. 29-37.