Experiments in Text-Based Mining and Analysis of Biological Information from MEDLINE on Functionally-Related Genes

Naureen Moon and Rahul Singh
Department of Computer Science, San Francisco State University
San Francisco, CA 94132
numoon@sfsu.edu, rsingh@cs.sfsu.edu

Abstract

Technological advancements such as microarrays have enabled biologists to generate unprecedented quantities of data about biological entities. This has lead to the development of a large number of algorithms for processing and analysis of biological data. Challenges however remain; for instance, genes that function cooperatively need not have similar expression patterns. This suggests the use of non-numerical sources of information to explore the underlying biology. We experimentally study various factors that are inherent in algorithmic methodologies for text analysis. The proposed method accesses MEDLINE dynamically to account for the latest research, with the available literature corresponding to the genes analyzed to develop lists of keywords. Natural language processing (NLP) techniques such as stop-word filtering and stemming are then applied to the lists, and keyword frequencies weighted using the term frequency-inverse document frequency (TFIDF) scheme. The results are input to a hierarchical clustering algorithm to derive groupings of genes by functionality. The process is repeated using z-score weighting and latent semantic analysis (LSA) to determine which yields the most accurate clustering. The study presented examines the importance of these steps and their influence on the overall efficacy of the system. We believe that the analysis conducted as part of this research will be invaluable to development and fine-tuning of text mining methodologies for biological literature.

1. Introduction

Microarray technology and other advanced techniques have enabled biologists to generate vast quantities of experimental data about various biological entities, genes in particular. The challenge now is to obtain meaningful interpretation of the measurements thus obtained. Computing can supply researchers with a macroscopic view of the data, and in particular, can discern patterns across a large spectrum of data, numerical or otherwise.

The proliferation of biological data has indeed driven the development of many algorithms for its quantitative analysis, generally based upon similarities between expression profiles of genes. An important tempering factor for such approaches, however, is the observation that functionally-related genes may not have similar expression patterns. This suggests the use of non-numerical information for the elucidation of gene function and interaction. As such, a corpus of research has recently been developed which utilizes this approach by analyzing biological literature from MEDLINE to identify interactions between genes. Currently at over 12 million references [11], the sheer size of MEDLINE necessitates the development of such techniques for extraction of relevant information and meaningful analysis of it.

Previous work in this area of research includes the use of co-occurrence of gene names in articles to construct a network of genes, where network "neighborhoods" constitute functional clusters [4]. Shatkay et al utilized a representative document for each gene to find similar

documents, comparing top-ranking documents for each gene to determine functional connections [10]. Some approaches applied various schemes for keyword extraction from MEDLINE abstracts to compare term frequencies across genes, with the values thus derived input to a clustering algorithm [2,3,7]. Other systems used MeSH (Medical Subject Headings) keywords from search results for genes to describe genes within already-derived clusters [5,8].

We experimentally examine the various factors involved in computational approaches to text mining. Relevant literature about genes under investigation is accessed from MEDLINE and subsequently analyzed to shed light on interaction between the genes. Since the information underlying MEDLINE is currently growing at a rate of over 10,000 references a week [11], the proposed method accesses MEDLINE dynamically to account for the latest research. Different levels of stop-word filtering are first applied to the data and precision-recall analysis is performed to optimize this parameter. The approach then uses the available literature to develop lists of keywords descriptive of the genes under investigation. NLP techniques such as stop-listing and stemming are then applied to the keyword lists to remove non-descriptive terms and condense related words to their common roots. Keyword frequencies are tabulated and adjusted using the TFIDF weighting scheme, with results input to a hierarchical clustering algorithm whereby the resultant clusters are used to compare performance of stemmed and unstemmed keyword lists. The analysis is then repeated with frequency adjustment using the z-score weighting scheme and LSA.

The aforementioned steps are often used, under various permutations and variants, as the kernel technologies in text-based data analysis and mining. Unfortunately however, few studies have concentrated on elucidating the impact of design decisions within each one of them to the overall quality of a text analysis and mining system, especially when the specificities introduced by an underlying domain are taken into account. This paper addresses these key questions in the context of a concrete problem domain and formulation.

2. Method

2.1. Term extraction and preprocessing of data

Gene-related documents were retrieved by dynamically querying the MEDLINE database using the *esearch* and *efetch* utilities [12]. Title, abstract text, and author last names and affiliations were extracted from each document. The text was subsequently preprocessed to remove capitalization, most punctuation, and strings containing no alphabetic characters (e.g., “0.5%” and “85-90”).

2.2. Stop-word filtering and precision-recall analysis

In the first study, 70 documents containing the word “gene” in the title were obtained (those containing no abstract were not used or counted). Each document was manually curated for biologically-significant keywords. After performing term extraction and preprocessing as described above, each document was subjected to 3 different stop-lists, with no stop-list used as the control. The first stop-list (stop-list 1) used consists of 319 common English words [13], most of which are uninformative outside of the context in which they are employed (e.g., “although” and “your”). We created the second stop-list (stop-list 2) by combining the first with a list of the 1000 most common words in English [14], after having removed redundancies and biologically-significant terms (e.g., “blood”), for a list size of 1044 terms. The last stop-list (dict) is a dictionary containing 41,236 words with all of their inflected forms, for a total list size of 81,520 [15]. The keyword lists thus obtained were compared with their manually-curated counterparts, and the number of words common to both lists, nc, was tabulated. This

data was used to compute the standard information retrieval metrics of precision and recall using:

$$P = nc / na \quad \text{and} \quad R = nc / nm \tag{1,2}$$

where na denotes the total number of words in the computationally-derived list and nm the total number of words in the manually-derived list. This analysis was performed for 70 documents to ensure stability of precision and recall values.

2.3. Term-frequency matrix calculation

The second study used 3 test sets of genes with 46, 26, and 44 genes each, consisting of 3, 4, and 9 well-defined clusters, respectively (Table 1). A query was submitted to MEDLINE for each gene, requesting 10 documents with the gene name in the title. If fewer than 10 documents were available that satisfied the criterion, the difference was made up by documents containing the gene name in the abstract. Terms were then extracted and the data preprocessed as above, with terms from different documents describing the same gene condensed into a single keyword list (after using the optimal stop-list determined in the previous study). The data thus obtained for each gene was combined into a term-frequency table with keywords comprising each row and each column denoting one gene in the test set. The term-frequency values were then weighted using the TFIDF scheme [9], which adjusts each frequency value as follows:

$$tfidf = tf \times \log (N / n) \tag{3}$$

where tf denotes the frequency of the term in the set of documents for a given gene, N the total number of documents in the background set, and n the number of background-set documents in which the term occurs. Since use of the TFIDF weighting scheme necessitates the development of a background set of documents, 1000 gene-related articles were processed for terms and their document frequencies tabulated.

Table 1. Three test sets of genes (set 1 from [5], 2 and 3 from [7])

	GENES (46)	FUNCTION
1	acs2, cdc19, eno2, fba1, gpm1, hxk2, pdc1, pdc5, pdc6, pfk1, pgk1, tdh1, tdh2, tdh3, tkl1, tpi1, tye7	Glycolysis Cluster
2	elo1, ole1, faa4, faa3, sur2, faa1, erg2, psd1, cyb5, pgm1	Fatty Acid Cluster
3	pex1, pex2, pex3, pex4, pex5, pex6, pex7, pex8, pex9, pex10, pex11, pex12, pex13, pex14, pex15, pex16, pex17, pex18, pex19	Peroxin Cluster
	GENES (26)	**FUNCTION**
1	nmda-r2a nmda-r2b nmda-r1 glur6 ka2 glur1 glur2 glur4 glur3 ka1	Glutamate receptor channels
2	dopamine beta-hydroxylase, tyrosine hydroxylase, phenethanolamine N-methyltransferase, catechol-O-methyltransferase, dopa decarboxylase, monoamine oxidase A, monoamine oxidase B	Catecholamine synthetic enzymes
3	alpha-tubulin, beta-tubulin, dynein, actin, alpha-spectrin	Cytoskeletal proteins
4	tyrosine transaminase, chorismate mutase, prephenate dehydratase, prephenate dehydrogenase	Tyrosine/phenylalanine synthesis
	GENES (44)	**FUNCTION**
1	cln1, cln2, gic1, gic2, msb2, rsr1, bud9, mnn1, och1, exg1, kre6, cwp1	Budding
2	clb5, clb6, rnr1, rad27, cdc21, dun1, rad51, cdc45, mcm2	DNA replication and repair
3	htb1, htb2, hta1, hta2, hta3, hho1	Chromatin
4	hhf1, hht1, tel2, apr7	Chromatin
5	tem1	Mitosis control
6	clb2, ace2, swi5, cdc20	Mitosis control
7	cts1, egt2	Cytokinesis
8	mcm3, mcm6, cdc6, cdc46	Prereplication complex formation
9	ste2, far1	Mating

Lastly, the matrix of weighted term frequencies was input to the hierarchical clustering algorithm of the Cluster 3.0 program [16] and the gene clusters obtained by manual inspection of the tree using Java TreeView [16].

2.4. Evaluation of stemming

Stemming is the process by which a word is reduced to its stem by the stripping of suffixes. It can play an important role in information retrieval because it allows inflected forms of the same word to be mapped together. In our case, the use of stemming allows words such as "express," "expressed," and "expression" to be counted as one entry in the list of keywords with frequency three rather than as three distinct keywords, each with frequency one.

We used the Porter stemming algorithm [17], to evaluate the effect of stemming on correctness of clustering. This entailed stemming words before constructing the keyword list and tabulating term frequencies. For TFIDF weighting, it was necessary also to develop a stemmed background set. The resultant matrix of values was used to derive gene clusters as described above, with these results (for all 3 test sets) serving as the control group for all tests conducted.

2.5. Comparison of frequency adjustment schemes

The last step consists of comparing the effectiveness of the TFIDF weighting scheme with z-score weighting and LSA, once again using the accuracy of the clustering result as the metric. Z-score weighting is similar to TFIDF in its use of a background set to adjust frequency values. The z-score is calculated as [1]:

$$z = (f - F) / \sigma \tag{4}$$

where f denotes the frequency of the term in the set of documents for a given gene, F its average frequency in the background set, and σ its standard deviation in the background set.

LSA is a well-known method for inferring word meanings by applying statistical techniques to the contexts in which words appear. One of its applications is determining correlations between documents in a set. It is used here as another method by which to adjust the term-frequency matrix. After a preliminary transformation, LSA applies singular value decomposition to the matrix X, rewriting it as a product of 3 matrices [6]:

$$\{X\} = \{W\}\{S\}\{P\}' \tag{5}$$

The dimensionality of the solution is then reduced by deleting coefficients in the diagonal matrix, S. We delete all but the 2 largest eigenvalues. The adjusted frequency matrix is then reconstructed using only 2 dimensions.

Lastly, the matrices obtained using z-score weighting and LSA are used to cluster the sets of genes for comparison with the TFIDF-derived clusters.

3. Results

3.1. Stop-list comparison

Keyword extraction at 4 levels of stop-word filtering was analyzed using precision and recall metrics for a collection of 70 documents. The no stop-list case suffers from low precision (0.4), while the dictionary gives low values for recall (0.5), due to removal of biologically-significant terms. An inverse relationship between precision and recall measurements is usually observed because removing insignificant terms from keyword lists increases precision, but necessarily also leads to the removal of some relevant terms. As this suggests, however, the increase in precision is generally not matched by an equal decrease in recall, as found in this case. The

small stop-lists give a significant increase in precision (up to 0.6) from the no stop-word case with a negligible loss of recall. As such, we favor the use of the small stop-lists because they seem to offer the best compromise, despite their low precision values (due to relatively few terms being removed for each, 319 for stop-list 1 and 1044 for stop-list 2). The optimal case would be a dictionary from which biologically-relevant terms have been removed, as shown by Liu et al [7]. This result also suggests the use of small stop-lists in information retrieval methods not applied to specialized domains. It should be noted, however, that these metrics were calculated against manually-derived keyword lists, obviously a very subjective standard.

3.2. Effect of stemming

For each of the 3 data sets, term-frequency tables derived with and without stemming were filtered with stop-list 2, weighted using TFIDF, and clustered using the hierarchical algorithm, with boundaries constructed by inspection. The results (Figure 1(A)) show no significant difference due to stemming. The data input to the clustering algorithm was analyzed to explain this aberrant result. It was found that most of the terms that occur in multiple inflected forms are common verbs, and therefore among the lowest-weighted words in the list. For example, the word "associate" occurs with 5 of its inflected forms in the unstemmed keyword list of the 46-gene set, frequently with the minimum weight value. Conversely, the most significant words, such as gene or disease names, have zero or one inflected forms (at most a plural). Furthermore, a word often occurs with one or more of its inflected forms in the set of documents for the same gene. If it does not appear in the keyword list of any other gene, as is often the case, then it is does not affect the clustering result. If it does, however, this serves as a link between genes containing any one of the inflected forms in their keyword lists. For example, the word "plasmalogen" occurs in the term lists of pex7 and pex13, and "plasmalogens" in the lists of pex2 and pex7, so that pex2 and pex13 are linked to pex7, and thereby indirectly to each other.

A

Gene Set	STEMMED		UNSTEMMED	
	no. correct	% correct	no. correct	% correct
46	40	87.0	41	89.1
26	25	96.2	25	96.2
44	27	61.4	27	61.4

B

Gene Set	TFIDF		Z-SCORE		LSA	
	no. correct	% correct	no. correct	% correct	no. correct	% correct
46	40	87.0	40	87.0	35	76.1
26	25	96.2	26	100.0	20	76.9
44	27	61.4	27	61.4	27	61.4

C

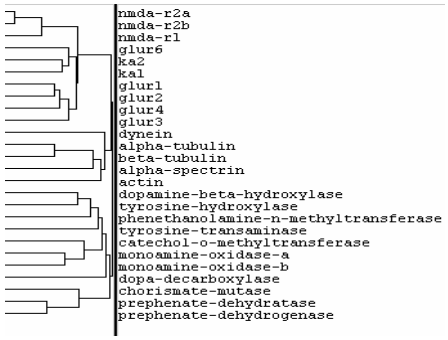

Figure 1. Clustering results for (A) stemmed and unstemmed cases and (B) TFIDF, z-score, and LSA methods. (C) View of hierarchical clustering output of 26-gene data set for TFIDF weighting

While an improvement was not observed in the clustering results due to stemming, this could likely be a peculiarity of its application here in a highly-specialized domain. Moreover, the method is still advantageous in its reduction of data size.

3.3. Effect of frequency adjustment schemes

Each test set in this study was stemmed and filtered with stop-list 2 before the different computational methods were applied to the term-frequency matrix. The hierarchical clustering

output for TFIDF is displayed (Figure 1(C)), with cluster boundaries determined by inspection. Comparison of the clusters derived from the TFIDF and z-score weighting methods show excellent results, with values ranging from 87.0 - 100% correctness (Figure 1(B)). As before, the 44-gene set shows significantly worse results, possibly due to the large number of clusters. It may also be that the 44-set clusters are less functionally coherent than their 46 and 26-set counterparts. The display of cluster functions (Table 1) shows that there are indeed clusters with identical functions, which may have resulted in genes being incorrectly assigned to clusters that perform similar functions as their own.

Results are for data generated using LSA are considerably worse than TFIDF and z-score for the first 2 data sets. It is unclear why this is the case.

4. Conclusions and future work

In this study, we have experimented with different parameters related to keyword extraction from biological literature as a means by which to investigate the factors that affect performance of text-mining systems. We found that the use of a simple stop-word list proves advantageous for information retrieval in this domain, though a large, domain-specific dictionary would likely be better. The same does not seem true for stemming, for reasons which are unclear and require further investigation.

Both TFIDF and z-score weighting techniques were seen to cluster relevant results effectively, and one does not emerge clearly better than the other. Conversely, clustering results obtained using LSA data were unimpressive, the cause of which, however, bears further scrutiny. The dimensionality of the solution can greatly alter the data obtained thereby, so a preliminary test might involve performance comparison of different choices of dimension.

Future work could also apply the techniques and results described above to different knowledge domains, or more general ones.

5. References

[1] M. Andrade and A. Valencia, “Automatic Extraction of Keywords from Scientific Text: Application to the Knowledge Domain of Protein Families,” Bioinformatics, 14:600-607, 1998

[2] D. Chaussabel and A. Sher, “Mining Microarray Expression Data by Literature Profiling,” Genome Biology, 3:1-16, 2002

[3] I. Iliopoulos, A. J. Enright, C. A. Ouzounis, “TextQuest: Document Clustering of MEDLINE Abstracts for Concept Discovery in Molecular Biology,” Proc. Pacific Symposium on Biocomputing, 384-395, 2001

[4] T. K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, “A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression,” Nature Genetics, 28: 21-28, 2001

[5] P. Kankar, S. Adak, A. Sarkar, “MedMeSH Summarizer: Text Mining for Gene Clusters,” Proc. SIAM International Conference in Data Mining, 2002

[6] T. K. Landauer, P. W. Foltz, D. Laham, “An Introduction to Latent Semantic Analysis,” Discourse Processes, 25:259-284, 1998

[7] Y. Liu, B. J. Ciliax, K. Borges, V. Dasigi, A. Ram, S. B. Navathe, R. Dingledine, “Comparison of Two Schemes for Automatic Keyword Extraction from MEDLINE for Functional Gene Clustering,” Proc. CSB 2004

[8] D. Masys, J. B. Welsh, J. L. Fink, M. Gribskov, I. Klakansky, J. Korbeil, “Use of Keyword Hierarchies to Interpret Gene Expression Patterns,” Bioinformatics, 17:319-326, 2001

[9] G. Salton and C. Buckley, “Text-Weighting Approaches in Automatic Text Retrieval,” Information Processing and Management, 24:513-523, 1988

[10] H. Shatkay, S. Edwards, W. J. Wilbur, M. Boguski, “Genes, Themes and Microarrays: Using Information Retrieval for Large-Scale Gene Analysis,” Proc. Intelligent Systems for Molecular Biology, 2000

[11] I. Spasic and S. Ananiadou, “A Flexible Measure of Contextual Similarity for Biomedical Terms,” Proc. Pacific Symposium on Biocomputing, 2005

[12] “Entrez Utilities,” http://eutils.ncbi.nlm.nih.gov/entrez/ query/static/ eutils_help.html

[13] http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

[14] “1000 Most Common Vocabulary Words in English,” http://esl.about.com/library/vocabulary/bl1000_list1.htm

[15] “The 12-dicts Word Lists,” http://wordlist.sourceforge.net/ 12dicts-readme.html#2of12inf

[16] “Open Source Clustering Software,” http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm

[17] “Porter Stemming Algorithm,” http://www.tartarus.org/ ~martin/PorterStemmer