BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1515–1522

Finding relevant references to genes and proteins in Medline using a Bayesian approach

Julie E. Leonard[1,2], Jeffrey B. Colombe[1] and Joshua L. Levy[1,*]

[1]Incellico Inc., 2327 Englert Dr., Durham, NC 27713, and [2]Program in Bioinformatics, NCSU, Raleigh, NC 27695, USA

Received on March 27, 2002; revised on May 17, 2002; accepted on May 21, 2002

ABSTRACT

Motivation: Mining the biomedical literature for references to genes and proteins always involves a tradeoff between high precision with false negatives, and high recall with false positives. Having a reliable method for assessing the relevance of literature mining results is crucial to finding ways to balance precision and recall, and for subsequently building automated systems to analyze these results. We hypothesize that abstracts and titles that discuss the same gene or protein use similar words. To validate this hypothesis, we built a dictionary- and rule-based system to mine Medline for references to genes and proteins, and used a Bayesian metric for scoring the relevance of each reference assignment.

Results: We analyzed the entire set of Medline records from 1966 to late 2001, and scored each gene and protein reference using a Bayesian estimated probability (EP) based on word frequency in a training set of 137 837 known assignments from 30 594 articles to 36 197 gene and protein symbols. Two test sets of 148 and 150 randomly chosen assignments, respectively, were hand-validated and categorized as either good or bad. The distributions of EP values, when plotted on a log-scale histogram, are shown to markedly differ between good and bad assignments. Using EP values, recall was 100% at 61% precision (EP $= 2 \times 10^{-5}$), 63% at 88% precision (EP $= 0.008$), and 10% at 100% precision (EP $= 0.1$). These results show that Medline entries discussing the same gene or protein have similar word usage, and that our method of assessing this similarity using EP values is valid, and enables an EP cutoff value to be determined that accurately and reproducibly balances precision and recall, allowing automated analysis of literature mining results.

Contact: jlevy@incellico.com; jleonard@incellico.com; jeffcolombe@hotmail.com.

INTRODUCTION

Literature searching is one of the most common information-processing tasks in the biomedical sciences. Medline search engines such as Entrez/Pubmed (Roberts, 2001; Schuler *et al.*, 1996) allow retrieval of document references based on text matching in an entry's fields. A user then hand-filters these results by reading the titles and/or abstracts, keeping relevant entries and discarding irrelevant ones. The final set of relevant articles represents the high-quality references on a given topic.

The above information retrieval (IR) problem of finding relevant documents is simple but time consuming for a human to perform. Retrieval of relevant documents can be the first step in information extraction (IE) processes that build knowledge from documents in a domain, such as the automatic detection of protein–protein interactions in the literature (Blaschke *et al.*, 1999; Humphreys *et al.*, 2000; Marcotte *et al.*, 2001; Ono *et al.*, 2001; Thomas *et al.*, 2000; Yoshida *et al.*, 2000), or the construction of literature networks (Jenssen *et al.*, 2001). The first example depends on being able to correctly identify documents relevant to protein–protein interactions, while both examples rely on being able to correctly identify documents relevant to specific genes and proteins, usually by detecting gene and protein symbols or abbreviations in the article. In this paper, we address the problem of retrieving relevant articles for a gene or protein using an automated process for document filtering, followed by an assessment of the probabilistic significance of each article.

Software for finding relevant documents must be able to first identify terms related to the search in the titles and abstracts and second must distinguish between relevant and irrelevant results. The success of a given program's ability to do this has typically been measured using precision (percent of relevant results) and recall (percent of total known results), as well as false-positive and false-negative rates. Stringency parameters can be varied to balance precision versus recall, and to shift the balance to either side. A curve in which high stringency gives low recall and high precision, while low stringency gives high recall and low precision is usually observed (see Marcotte *et al.*, 2001, Figure 2b).

Automated detection of relevant words or terms can use linguistic, statistical and/or machine-learning approaches.

*To whom correspondence should be addressed.

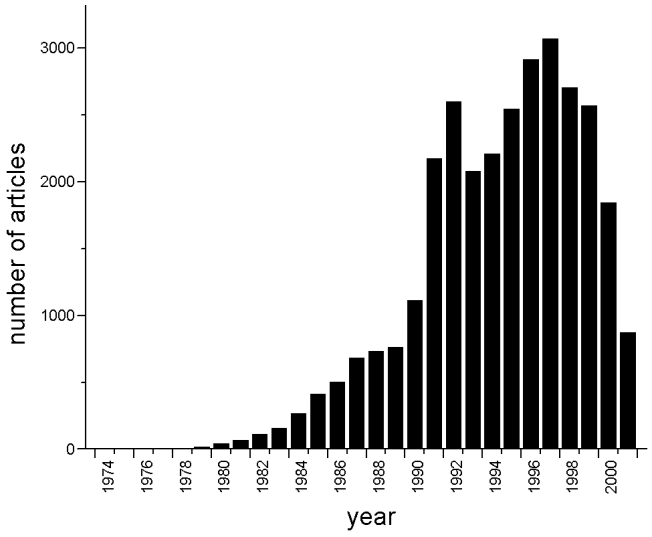

Fig. 1. Histogram of Medline entries per year in our training set. The training set was constructed by taking the entire set of Medline entries that had cross-references to sequence databases such as GenBank. Accession numbers were then translated to gene and protein symbols by traversing each record's cross-references, for example from GenBank to LocusLink, which contains a symbol. Note that 1974, 1975 and 1976 have one, zero, and three counts, respectively.

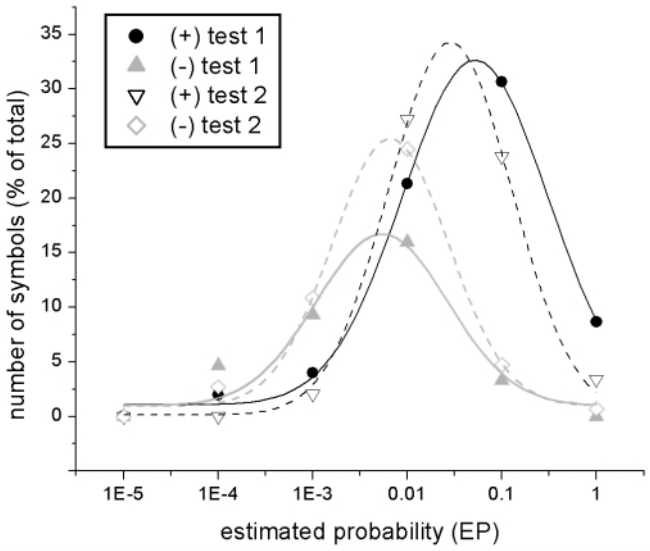

Fig. 2. Distribution of good (+) and bad (−) symbol assignments over the range of estimated probability (EP) values. Dashed lines show data from test set 1, 148 manually checked randomly-sampled assignments, while solid lines show the data from test set 2, 150 manually checked randomly-sampled assignments. The two test sets are distributed similarly, showing that samples were not biased. The distributions of good assignments are shifted to higher EP values relative to the distribution of bad assignments, validating the use of EP values for determining relevance.

Linguistic approaches include dictionary- and rule-based methods that look for specific words and sentence features in order to identify terms. In the biomedical literature, dictionary terms can be gene and protein names or symbols, or subject-specific terms. English words can be filtered using an English dictionary. The rules usually take into account variable features such as punctuation, hyphenation and capitalization in order to distinguish real from false biomedical terms, and also to identify new symbols or terms (Fukuda *et al.*, 1998; Proux *et al.*, 1998). Though dictionary-based approaches are useful, they suffer from high false-positive rates when used without other filters or subsequent statistical analysis (Mutalik *et al.*, 2001; Nadkarni *et al.*, 2001).

Machine-learning approaches use features within the abstract and title text for training adaptive algorithms. For example, an iterative keyword-extraction method has been applied to the problem of assigning proteins to classes using SwissProt descriptions (Tamames *et al.*, 1998), naive Bayesian, decision trees and inductive learning methods have been used to differentiate between biomolecule types (Hatzivassiloglou *et al.*, 2001) and a hidden Markov model has been used for gene name recognition (Collier *et al.*, 2000). The features chosen for training can be similar to features recognized in rule-based approaches, such as capitalization, Greek letters, or neighboring words.

Statistical approaches can use metrics such as word frequency and term weighting to filter irrelevant articles and to identify relevant symbols or terms (Andrade and Valencia, 1998). Metrics can include Bayesian probability (Marcotte *et al.*, 2001), maximum entropy (Raychaudhuri *et al.*, 2002) or other methods, such as C/NC value (Frantzi and Ananiadou, 1999). The Bayesian approach used by Marcotte et al. involved term and word frequencies to calculate a likelihood score of Medline abstracts discussing protein–protein interactions. Using a training set of abstracts known to discuss protein–protein interactions, they showed that the distribution of log likelihood scores was higher for Medline abstracts discussing interactions than for a random set of abstracts. This allowed them to choose an optimal cutoff score that distinguished between the two distributions. Because the two distributions overlapped, they had to choose a cutoff score that balanced precision and recall. Raychaudhuri et al. used a maximum entropy method to assign Gene Ontology (GO) codes to genes associated with abstracts. In their method, a confidence measure was calculated that correlated well with prediction accuracy.

Using similar logic, we propose that abstracts discussing

Table 1. Example assignments of Medline entries to gene and protein symbols, including the calculated Bayesian estimated probability (EP) value

PubMed-ID	Organism	Symbol	EP-value	Assessment
9054946	Human	STFB	0.576192515	Relevant
9185522	Human	MSP	0.160713616	Relevant
9185698	Human	PACE4	0.155983971	Relevant
9197532	Human	MYC	0.150897133	Relevant
9088342	Human	ERCC4	0.145598152	Relevant
8157699	Mouse–human	RARG	0.144040377	Relevant
9018118	Human	BEK	0.13627135	Relevant
9006941	Human	MACH	0.134491264	Relevant
8180497	All	RAB17	0.094318479	Relevant
9046055	Mouse–human	DMD	0.075282465	Relevant
9152386	Human	CYP2D	0.070965125	Relevant
9055826	Rat	PEM	0.064503465	Relevant
9174094	Human	SCN4A	0.062759006	Relevant
9177787	Human	ARHG	0.052642562	Relevant
9176151	Rat	RAB6	0.049656803	Relevant
9128246	Mouse	G	0.047954011	Irrelevant
9120417	Human	MOG	0.047910074	Relevant
9022004	Mouse	RAG1	0.045212177	Relevant
9188795	Rat	MUC2	0.044945587	Relevant
8112598	Human	MAR	0.044362471	Irrelevant
9010225	Mouse–human	EPS8	0.043756746	Relevant
9188856	Human	HGF	0.042814845	Relevant
9093908	Human	ICE	0.042224029	Relevant
9166284	Human	MT1-MMP	0.040641578	Relevant
9021013	Human	PAX3	0.036484183	Relevant
9143297	Human	MT2	6.79×10^{-4}	Irrelevant
9203629	All	AMP	6.02×10^{-4}	Irrelevant
9177684	All	TRP	5.40×10^{-4}	Irrelevant
9166733	All	TRANSFERRIN	5.10×10^{-4}	Irrelevant
9119016	Mouse–human	UPA	4.85×10^{-4}	Relevant
8139576	All	FTZ	4.54×10^{-4}	Irrelevant
9104036	All	HIS3	4.41×10^{-4}	Irrelevant
8152905	Mouse–human	HS	3.75×10^{-4}	Irrelevant
8977179	Human	RAF1	3.65×10^{-4}	Relevant
9075785	Human	S	3.35×10^{-4}	Irrelevant
8980296	Human	S	3.11×10^{-4}	Irrelevant
9143508	Human	P32	3.11×10^{-4}	Irrelevant
9103614	All	FST	3.07×10^{-4}	Irrelevant
9168617	All	FLA	2.46×10^{-4}	Relevant
9027492	Human	RMSA1	1.72×10^{-4}	Relevant
9002272	All	INA	1.18×10^{-4}	Irrelevant
9174597	Human	P50	1.14×10^{-4}	Relevant
9108071	Human	E	1.02×10^{-4}	Irrelevant
9202289	Mouse–human	ED	7.71×10^{-5}	Irrelevant
9093942	Human	E	7.55×10^{-5}	Irrelevant
9108393	Human	E	7.50×10^{-5}	Irrelevant
9058790	Human	P50	7.35×10^{-5}	Relevant
9159467*	All	PRP21	7.00×10^{-5}	Relevant
8978713	All	NG	6.77×10^{-5}	Irrelevant
9048912	Human	E	5.60×10^{-5}	Irrelevant
9151810	Human	S	4.19×10^{-5}	Irrelevant

The examples are taken from test set 2, which includes entries from 1997. Assignments were assessed by reading the title and abstract, then deciding if the article was either relevant or irrelevant to the gene or protein. The assignment marked by the * is an example of a false negative.

not only a particular topic or GO category, but also abstracts discussing a particular gene or protein, use similar language and therefore have similar word frequencies. We show, using a Bayesian probabilistic method that

this hypothesis can be validated and the results used to create a rich and relevant set of literature references for a given set of genes and proteins.

SYSTEM AND METHODS

In order to validate our hypothesis, we developed a system for detecting gene and protein symbols in Medline entries, and for determining the relevance of each Medline entry to the gene or protein it referenced. This required the following components:

(1) the complete set of Medline entries from 1966 to late 2001;
(2) dictionaries of gene and protein symbols, as well as English words;
(3) a high-quality training set of Medline entries which were known to discuss specific genes and proteins in our dictionaries;
(4) rules for use with above dictionaries to find gene/protein symbols in Medline articles while filtering out non-symbols;
(5) Bayesian probabilistic method for calculating an article's relevance to a particular gene/protein, using word occurrences found in the training set.

Training data set

The first requirement for the system was a set of known cross-references between gene and protein symbols and Medline entries. Medline was obtained from the National Library of Medicine through a licensing agreement. We found that many Medline articles contain cross-references to accession numbers from biological databases such as GenBank, so we chose the entire set of Medline entries with accession numbers to create our training set.

In order to translate accession numbers to the gene or protein symbols they represent, we used our CELL^{TM} Annotate/Translate software (http://www.incellico.com/products.html). This software automates the process of traversing cross-referenced records in different biological databases. For example, by traversing cross-references between GenBank accession numbers and LocusLink entries, we were able to retrieve the appropriate gene symbols. Though it was not possible to successfully translate all accession numbers to symbols, the end result was 137 837 cross-references between 30 594 unique Medline entries and 36 197 gene or protein symbols.

We then determined word frequencies for the training set. For each gene or protein, the set of words from all relevant abstracts and titles was determined. Next, we calculated the estimated probability of each of those words appearing in an abstract or title discussing that gene or protein. In this way, a table of gene and protein symbols and their associated word probabilities was created.

Table 2. Results of Student's t-test performed on EP values from both good (+) and bad (−) assignments from the two test sets

Groups tested	*p*-value
Test set 1 good vs test set 2 good	0.0547
Test set 1 good vs test set 1 bad	1.02×10^{-4}*
Test set 2 good vs test set 2 bad	0.00982*
Test set 1 bad vs test set 2 bad	0.865
Test set 1+2 good vs test set 1+2 bad	2.32×10^{-4}*

*$p < 0.05$
The distribution of EP values is significantly different between good (+) and bad (−) assignments for each test sets, as well as for both test sets combined. In addition, good (+) distributions for both test sets are not significantly different, as are bad (−) distributions from both test sets.

Identification of gene/protein symbols in Medline entries

To identify symbols in Medline entries, we applied a dictionary- and rule-based approach to the entire Medline database spanning the years 1966 to late 2001. The dictionaries, consisting of lists of gene and protein symbols, were compiled by extracting gene and protein symbols from the public databases HUGO (Povey *et al.*, 2001), OMIM (Hamosh *et al.*, 2002) and LocusLink (Pruitt and Maglott, 2001). In many instances, the dictionaries contained multiple symbols for a particular gene or protein due to the existence of alternate names or synonyms.

Because gene and protein symbols can sometimes be ambiguous, referring to different products in different organisms, we created species-specific dictionaries. We chose to use model organisms, primarily due to the availability of information in the public databases. Dictionaries were split into the following organisms: mouse, human, rat, zebrafish, and fruit fly. In addition, we constructed an all-inclusive dictionary for Medline entries that either did not contain a specific species or that referred to a different species than the five listed above, and a human+mouse dictionary for entries referring to both human and mouse. At the time of analysis, the dictionaries contained between 1856 and 194 535 gene symbols (Table 3).

Each word in a Medline title and abstract was compared to the appropriate dictionary (either organism-specific or multiple-organism). Gene symbols that were also English words were stored in species-specific tables and were used to help determine whether a word was a true gene symbol or an English word. In this way, we were able to better filter out potential bad assignments.

In addition to the dictionary-look-up step, numerous rules were used to determine if a particular word in an entry's title or abstract was a gene or protein symbol, and that it was not an English word. These rules included removing concatenating tokens from words to see if the

Table 3. Number of gene and protein symbols contained in each species-specific symbol dictionary. Gene and protein symbols were taken from HUGO, LocusLink, and OMIM

Species	Number of symbols
Human	94 867
Mouse	59 140
Rat	7 741
Zebrafish	1 856
Fruit fly	52 634
Human + mouse	141 861
All	194 535

result was a gene symbol (for example, changing IL-4 to IL4) and replacing Greek characters with their Roman equivalents (for example, changing PKC-alpha to PKCA). In addition to these rules, another rule was used to filter out Medline entries that did not discuss genes or proteins. This rule selected genetics-related entries based on the presence of genetics-related MESH headings or the presence of the words 'gene' or 'oncogene', or their variants in an abstract or title, and discarded all other Medline entries. This step substantially reduced processing time, as most entries in Medline do not discuss specific genes or proteins. Approximately 90% of the Medline entries were discarded during this step in the program, leaving a little over 1 million entries to be searched for gene and protein symbols. We also included rules to find new gene and protein symbols on-the-fly, based on proximity to words such as 'gene' or 'oncogene'. For example, if a word in the title or abstract was not found in the gene symbol dictionary, was not an English word, and was immediately followed by the word 'gene', among other specific criteria, then it was considered to be a gene symbol. Finally, if a word was in the appropriate gene symbol dictionary, and was determined by our rules to be a gene or protein symbol, then we created a cross-reference between the symbol and the article, which we call an assignment.

Determining the relevance of an article to a gene/protein

In order to calculate the relevance of a given article to a gene/protein symbol to which it referred, we used a Bayesian method for calculating the estimated probability (EP) based on the words that appear in its text. The method is based on our hypothesis that articles discussing the same gene or protein will use similar language, or in other words, that the articles will be more similar to each other than to the background set of all words in the genetics-related subset of Medline. We determined the probability of a given word occurring in any Medline entry in the subset as:

$$\hat{p}(\text{word}) = f(\text{word}) = \frac{\#\text{ articles with word}}{\text{total }\#\text{ articles}}$$

where $\hat{p}(\text{word})$ is the estimated probability of a word, $f(\text{word})$ is the relative frequency of the word, determined by counting word occurrences for all Medline entries in the genetics-related subset. We also determined the estimated probability of the presence of a given word in the title or abstract of an article discussing a particular gene or protein as:

$$\begin{aligned}\hat{p}(\text{word} \mid \text{gene}) &= f(\text{word} \mid \text{gene}) \\ &= \frac{\#\text{ articles with word AND gene}}{\#\text{ articles with gene}}\end{aligned}$$

We then calculated the probability of any article in the genetics-related subset referring to a given gene as:

$$\begin{aligned}\hat{p}(\text{gene}) &= f(\text{gene}) \\ &= \frac{\#\text{ articles with gene}}{\text{total }\#\text{ articles discussing gene or proteins}}\end{aligned}$$

The probability of a symbol for a given gene/protein co-occurring with a given word was calculated as:

$$\hat{p}(\text{gene} \mid \text{word}) = \hat{p}(\text{word} \mid \text{gene}) \cdot \frac{\hat{p}(\text{gene})}{\hat{p}(\text{word})}$$

Finally, estimated probability (EP) values were calculated using a nave Bayesian, or factorial, prior as follows:

$$\begin{aligned}EP_{j,I} &= \hat{p}(\text{gene}_j \mid \text{word}_{i=1\ldots n}) \\ &= 1 - \prod_{i=1\ldots n} (1 - \hat{p}(\text{gene}_j \mid \text{word}_i))\end{aligned}$$

for all genes j in the set of gene symbols $J = 1 \ldots m$ with all words i in the set of words $I = 1 \ldots n$ comprising each abstract. The estimated probabilities were subtracted from 1 in order to treat the *insignificance* of each word, rather than the significance, as a factorial cause. This way, multiple words that are significant in predicting gene/protein references are made to have a mutually reinforcing effect, rather than a mutually undermining effect in the joint prediction when they are multiplied as factors, while insignificant words are made to have a diminishing effect on the final prediction when treated as factors. In addition, it is necessary to subtract the probability of the gene given the word from 1 to account for words that were not in articles in the training set. Subtracting each probability from 1 prevents these 'new' words from being assigned a probability of 0, which when multiplied by all of the other probabilities would return an overall EP value of 0, even if all of the other words were highly probable and very indicative of that particular gene.

Table 4. The 20 symbols referenced by the most articles in the training set

Symbol	Number of Articles
RAB20	316
RAB19	178
PPY	154
PPY-55A	146
D33	144
PPD33	144
NUP214	139
KIAA0023	139
CAIN	139
D9S46E	139
CAN	139
CANA1	134
RH30A	111
RH4	111
RHPI	111
RHCE	111
VEGFA	109
VEGF	109
ADH-3	106

RESULTS

Our hypothesis is based on the assumption that Medline abstracts and titles that discuss the same protein or gene will tend to use similar words. Using a dictionary- and rule-based method, we processed the entire set of Medline entries from 1966 to late 2001. An estimated probability (EP) value was calculated for each result, based on a Bayesian conditional probability of a Medline entry's title or abstract containing words that have been used by other articles known to discuss that particular gene or protein.

To identify any bias in our training set, we examined the number of articles per symbol and the number of article per year in the training set. The number of articles per symbol follows an exponential decay curve, with about 42% of symbols referenced by one article, 18% referenced by two articles, 10% referenced by three articles, etc. with a mean of 3.7 articles per symbol (data not shown). More than 2500 symbols had 10 or greater articles, the top 20 of which are shown in Table 4. To examine the distribution of cross-references in our training set over time, we plotted a histogram of Medline entry counts per year, shown in Figure 1. The dates range from 1974 to 2001, with more than 87% of entries having publication dates of 1990 or later. This is most likely due to the sharp increase in the publication of gene sequences in the 1990s (see http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

In order to check our results, and to validate that our EP values truly reflected an article's relevance to a gene or protein, a set of assignments from our results were hand validated. Medline is released on a single digital tape and is organized into more than 400 single files of entries in XML, which we will refer to as sections, each ranging from <1 megabyte to >100 megabytes. Entries from a single year usually span many sections. In order to perform a completely unbiased analysis, two test sets were created by randomly choosing processed documents from different sections of Medline. The first test set, from 1994, contained 148 assignments, and the second test set, mainly from 1997, had 150. Each Medline entry was retrieved, and the title and abstract read and evaluated for relevance to the gene or protein in our assignment. Results were binned as being either good (+) or bad (−). In order to be classified as a good assignment, the symbol must refer to a gene or protein, and not to a disease, cell line, or other acronym or English word. Examples of good and bad assignments are given in Table 1.

The distributions of EP values from each bin (good (+) and bad (−)) for both test sets were then plotted on a histogram (Figure 2). The histogram shows that for both test sets, the distribution of EP values for bad assignments is different than that for good assignments, which peak at between 0.001 and 0.01 and between 0.01 and 0.1, respectively. Student's t-tests were performed to verify that test sets were similar to each other, and that good (+) and bad (−) assignments were significantly different (Table 2).

A precision versus recall curve was calculated using both test data sets (total 298 assignments; Figure 3). At varying EP values, we achieved the following recall and precision rates, respectively: 100% and 61% at EP $= 2 \times 10^{-5}$, 63% and 88% at EP $= 0.008$, and 10% and 100% at EP $= 0.1$. We empirically decided that EP $= 0.008$ was a good cutoff for future analyses requiring high precision and good recall, though other EP values offer different balances between precision and recall, depending on the demands of the application.

DISCUSSION

In this paper, we showed that Medline abstracts and titles discussing the same gene or protein use similar language, a similarity which we measured in the form of word frequencies and calculated using a Bayesian estimated probability (EP). It has been previously shown that Medline entries discussing protein–protein interactions can be distinguished from entries not discussing interactions using a set of discriminating words which appear at unexpectedly high or low frequencies in Medline abstracts (Marcotte *et al.*, 2001). Instead of calculating discriminating word sets, our method uses all words in the title and abstract, and our EP values are based on the probability of a word appearing in an abstract that discusses the given gene or protein, *not by the frequency* it appears in any single abstract. A large training set of Medline entries ($>30\,000$) was required for this purpose, and was also necessary for having

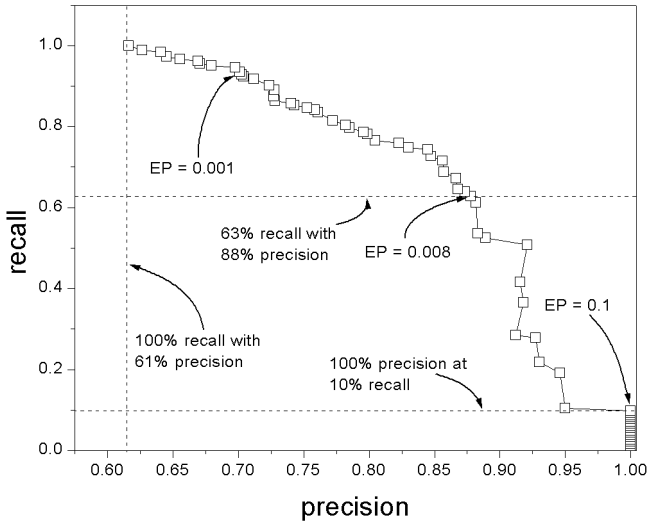

Fig. 3. Precision versus recall at varying Bayesian estimated probability (EP) values. The lowest precision is 61% at 100% recall, indicated by the vertical dotted line. Precision and recall stay relatively proportional until about EP = 0.008 (upper horizontal dotted line), above which precision increases little while recall diminishes significantly. EP = 0.008 offers a good balance between precision and recall, which are 88% and 63% respectively. At EP values of 0.1 or greater, precision is 100%, shown by the lower horizontal dotted line.

a large number of gene and protein symbols (>36 000) available for EP value calculation. The process of creating the training set included the translation of accession numbers to gene and protein symbols, which involved automating the process of traversing references between biological databases. Using our CELL^{TM} Annotate/Translate software, we were able to translate GenBank accession numbers to LocusLink gene symbols, for example, though other methods might have been used as well.

One of the problems of using a dictionary of symbols is that finding the symbol in the text does not mean that the text refers to a gene or protein with that symbol. Since symbols are generally acronyms, abbreviations, or other combinations of letters and numbers, other concepts that share that symbol also appear in abstracts and titles. For example, CA2 refers to the carbonic anhydrase-II protein, but it also refers to the Ca^{2+} ion of calcium. If we were to add a rule eliminating symbols with a '+' after them, another problem would be introduced in that cell lines with a particular protein are also indicated with a '+', and those references would be lost. Another problem is that symbols can refer to different proteins in different organisms. CAR, which is arrestin-C in humans, refers to a cytoplasmic arginine transducer in bacteria, and is also an acronym for Central African Republic in some abstracts. While we developed rules to filter out many bad assignments, it was impossible to make rules for every possible usage of every symbol. Therefore we assumed that a rule-based system for finding symbols could be made more sensitive using a word-occurrence method for assessing relevance. Although our training set included ~36 000 symbols, there is the potential of not having accurate word frequencies for EP-value calculation due to the symbol in question being absent from our training set.

Another problem of symbol identification is false negatives. For example, the entry in Table 1 for PubMed id = 9159467, symbol = PRP21 (marked by a *), which our system correctly identified, was given a low EP value. The abstract in question is clearly relevant to PRP21, but our training set had only one article for PRP21 in which the context was different, leading to a difference in word occurrences, and therefore a low EP value. Because about 42% of symbols had only one abstract in our training set, this may be one of the sources of overlap in the distributions of good (+) and bad (−) EP values, and therefore finding new ways to add articles to the training set could potentially alleviate this problem. One possible strategy would be to take assignments from a first-pass run of our system where EP $\geqslant$ 0.1, precision is 100%, and add those to the training set before re-running. This improvement would lead to increased accuracy over time.

Most of the entries in the training set are from 1990 or later, and therefore there is a possibility that earlier abstracts containing gene and protein symbols might receive low EP values if the context of relevant articles has significantly shifted since they were published. Additionally, if a new context is discovered for a gene or protein, the word frequencies may be different enough that the abstracts discussing the new context may get a low EP value for the gene or protein symbol. In order to prevent this, frequent updates to the word/gene tables are planned.

The difference in EP-value distributions between good and bad assignments reflects the differences in word occurrences between relevant and irrelevant articles that contain a given symbol. The peaks of the two distributions overlap between EP-values of 0.001 and 0.01, but precision is always above 70% for the overlapping region. Recall falls sharply with higher EP-values between ~0.01 and <0.1, but above EP = 0.1, precision is 100%. In general, when comparing good results with high and low EP-values, assignments with higher EP-values seemed to be more relevant to the given gene or protein, as assessed by the subject or general focus of the paper (data not shown).

We intend to use our results to construct a network of protein- and gene-to-literature references, and to use our EP-values to calculate a 'confidence' for each assignment. Future work will focus on refining the EP-value method to minimize the overlap between good and bad assignments, and on text mining for other entities in the literature,

including chemicals, diseases and tissues, each of which will require new rules and dictionaries. The use of synonyms in query expansion has been shown to improve retrieval (Hersh *et al.*, 2000). Thus, we also plan to use the concept of synonyms for gene and protein names, in that we will merge word-probability sets for all known symbols for a given gene or protein, and use a dictionary with multiple synonyms for each biomolecule. In addition, the results of expanded text mining is expected to result in a less fragmented network of literature references, and also allow for faster literature searching with fewer queries.

REFERENCES

Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 60–67.

Collier,N., Nobata,C. and Tsujii,J. (2000) Extracting the names of genes and gene products with a hidden Markov model. *Coling2000*, 201–207.

Frantzi,K.T. and Ananiadou,S. (1999) The C/NC value domain independent method for multi-word term extraction. *Journal of Natural Languge Processing*, **6**, 145–180.

Fukuda,K., Tamura,A., Tsunoda,T. and Takagi,T. (1998) Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.*, 707–718.

Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.

Hatzivassiloglou,V., Duboue,P.A. and Rzhetsky,A. (2001) Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, **17** (Suppl 1), S97–106.

Hersh,W., Price,S. and Donohoe,L. (2000) Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proc. AMIA Symp.*, 344–348.

Humphreys,K., Demetriou,G. and Gaizauskas,R. (2000) Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac. Symp. Biocomput.*, 505–516.

Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, **28**, 21–28.

Marcotte,E.M., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.

Mutalik,P.G., Deshpande,A. and Nadkarni,P.M. (2001) Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J. Am. Med. Inform. Assoc.*, **8**, 598–609.

Nadkarni,P., Chen,R. and Brandt,C. (2001) UMLS concept indexing for production databases: a feasibility study. *J. Am. Med. Inform. Assoc.*, **8**, 80–91.

Ono,T., Hishigaki,H., Tanigami,A. and Takagi,T. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.*, **109**, 678–680.

Proux,D., Rechenmann,F., Julliard,L., Pillet,V.V. and Jacq,B. (1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *Genome Inform. Ser. Workshop Genome Inform.*, **9**, 72–80.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Raychaudhuri,S., Chang,J.T., Sutphin,P.D. and Altman,R.B. (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, **12**, 203–214.

Roberts,R.J. (2001) PubMed central: the GenBank of the published literature. *Proc. Natl Acad. Sci. USA*, **98**, 381–382.

Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.

Tamames,J., Ouzounis,C., Casari,G., Sander,C. and Valencia,A. (1998) EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, **14**, 542–543.

Thomas,J., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.*, 541–552.

Yoshida,M., Fukuda,K. and Takagi,T. (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, **16**, 169–175.