BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 22 2006, pages 2813–2818
doi:10.1093/bioinformatics/btl480

Databases and ontologies

ADAM: another database of abbreviations in MEDLINE

Wei Zhou, Vetle I. Torvik and Neil R. Smalheiser*

Department of Psychiatry and Psychiatric Institute, MC912, University of Illinois at Chicago, Chicago, IL 60612, USA

Received on June 14, 2006; revised on September 7, 2006; accepted on September 8, 2006
Advance Access publication September 18, 2006
Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Abbreviations are an important type of terminology in the biomedical domain. Although several groups have already created databases of biomedical abbreviations, these are either not public, or are not comprehensive, or focus exclusively on acronym-type abbreviations. We have created another abbreviation database, ADAM, which covers commonly used abbreviations and their definitions (or long-forms) within MEDLINE titles and abstracts, including both acronym and non-acronym abbreviations.

Results: A model of recognizing abbreviations and their long-forms from titles and abstracts of MEDLINE (2006 baseline) was employed. After grouping morphological variants, 59 405 abbreviation/long-form pairs were identified. ADAM shows high precision (97.4%) and includes most of the frequently used abbreviations contained in the Unified Medical Language System (UMLS) Lexicon and the Stanford Abbreviation Database. Conversely, one-third of abbreviations in ADAM are novel insofar as they are not included in either database. About 19% of the novel abbreviations are non-acronym-type and these cover at least seven different types of short-form/long-form pairs.

Availability: A free, public query interface to ADAM is available at http://arrowsmith.psych.uic.edu, and the entire database can be downloaded as a text file.

Contact: neils@uic.edu

1 INTRODUCTION

In recent years, numerous online text mining tools have been created in the biomedical domain to assist scientists in their research (Krallinger and Valencia, 2005; Jensen *et al*., 2006), including lists of abbreviations. An abbreviation is a short-form of a word or phrase used in place of the corresponding long-form. The biomedical literature is growing by over 900 000 articles per year (Stead *et al*., 2005), which makes it hard for thesauruses, such as the Unified Medical Language System (UMLS) (http://umlsks.nlm.nih.gov/) to keep track of all the abbreviations. To help resolve this problem, many techniques have been introduced to identify abbreviations and their long-forms (or definitions) automatically from biomedical articles, and several online abbreviation databases have been created (Wren *et al*., 2005). Identifying long-forms is important for resolving the meaning of abbreviations in biomedical articles, which in turn facilitates information retrieval and information extraction applications (Friedman, 2000; Aronson, 2001).

Abbreviations can be classified as acronyms or non-acronyms. An acronym is a word formed from the initial letter or letters of each of the successive parts or major parts of the long-form: 'NASA' is an acronym for 'National Aeronautics and Space Administration'. A more relaxed definition of acronym would include words formed from initial letter or letters of each of major parts of the long-form, e.g. 'CKB' can arguably be considered an acronym for 'brain creatine kinase', although 'CKB' is out of normal order. In contrast, non-acronym abbreviations do not follow particular lexical patterns with the long-forms. For example, '11p' is a common abbreviation for 'the short arm of chromosome 11'. Note that the letter 'p' does not occur in its long-form at all.

In this paper, we present a systematic method for recognizing frequently used abbreviations and their long-forms within MEDLINE titles and abstracts, based solely on their statistical features and not employing any lexical information, in order to capture both acronyms and non-acronyms. The original motivation for this study was to create a look-up list that would assist us in identifying abbreviations among so-called B-terms, as part of the text processing algorithms supporting the Arrowsmith two node search interface (Swanson and Smalheiser, 1997; Smalheiser, 2005) (http://arrowsmith.psych.uic.edu). (B-terms are title words and phrases that are shared in two different sets of articles in MEDLINE; these B-terms may point to meaningful links across these often disparate literatures.) As well, while programming the Anne O'Tate tool that summarizes features of papers retrieved by a PubMed query (http://arrowsmith.psych.uic.edu), we noticed that abbreviations constitute a significant percentage of words that are 'important' in a given set of topical articles (i.e. occur frequently within that set of articles but relatively infrequently in MEDLINE as a whole). Thus, we have chosen our criteria of inclusion to focus our attention on terms that are particularly likely to appear as B-terms or important words in these tools. However, the database (ADAM) should be useful in a wide number of text processing applications, and can be freely downloaded or queried for non-commercial purposes.

2 METHODS

Our method consists of five sequential steps: step 1, extract candidate abbreviations (short-forms) and the contexts (surrounding text) in which they occur; step 2, identify candidate long-forms by using the statistical information found in the contexts; step 3, filter the short-form/long-form pairs according to a rule of length ratio and an empirically-validated cut-off value; step 4, verify that the short-forms are used in text separately from their

*To whom correspondence should be addressed.

© 2006 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

long-forms; step 5, group together morphologically similar long-forms that correspond to the same short-form or its lexical variants.

To assist in our modeling effort, we characterized certain features of abbreviation/long-form pairs listed as EXCELLENT or GOOD acronyms in the Stanford Abbreviation Database (graciously provided by Jeff Chang and Russ Altman). This will be presented below, but it should be emphasized that we used the Stanford abbreviations in an advisory capacity only and not for example as a training set for machine learning algorithms.

2.1 Step 1: extract candidate abbreviations (short-forms) and the contexts (surrounding text) in which they occur

As will be justified below, we extracted all single-words within parentheses in titles and abstracts of articles in MEDLINE (2006 baseline) as raw candidate short-forms. To obtain the context of the short-form, we extracted 3N (N is the number of alphanumeric characters in the candidate short-form) words to the left of the open parenthesis within the same sentence. For example, given a text: '. . .To assess the proportion of hospitalized patients who tested positive for human immunodeficiency virus (HIV) by a routine inpatient testing service. . .', 'HIV' was collected as the candidate short-form and 9 (3×3) words before the open parentheses 'hospitalized patients who tested positive for human immunodeficiency virus' was collected as the context. In case of nested parentheses, the expression inside of the outermost parenthesis is extracted. For example, in the context '. . .decrease in serum free triiodothyronine (FT(3)) levels. . .', 'FT(3)' will be collected as the candidate short-form.

Why capture only abbreviations inside of parentheses. We are admittedly aware of exceptions: e.g. the abbreviation K252a (a tyrosine kinase inhibitor) does not appear next to its chemical name within any MEDLINE abstract. However, the vast majority of abbreviations are defined as 'long-form (short-form)' on one or more occasions within MEDLINE titles or abstracts.

Why capture only single-words. Multi-word abbreviations do exist. However, the current model is not designed to distinguish multi-word abbreviations from parenthetical expressions, including biomedical terms that are not abbreviations (Liu and Friedman, 2003). To assess whether it was important to capture multi-word abbreviations, we examined the most frequent multi-word abbreviations listed in the Stanford Abbreviation Database. We observed that these are usually compound abbreviations. For example, 'DPP III' is an abbreviation for 'Dipeptidy1 Peptidase III'. We feel that this is in a sense redundant with the single-word abbreviation/long-form pair 'Dipeptidy1 Peptidase (DPP)'. Thus, we decided to include only single-word abbreviations in ADAM.

We also judged that single-letter abbreviations, such as 'A'–'Z', are not important to capture. Of single-letter abbreviations in the Stanford database, only '1-adrenaline (A)', '1-phosphate (P)' and '1-hour (H)' have been frequently used. Some of them, such as 'A', 'B' or 'C', are used commonly for indentation.

Thus, we restricted ADAM abbreviation candidates to single-words with two or more alphanumeric characters. We also excluded Roman numerals 'I' through 'XIV', which are used commonly for numbering in texts.

Why capture the pattern 'long-form (abbreviation)', instead of '(long-form) abbreviation', or '(abbreviation) long-form', or 'abbreviation (long-form)'. We examined a random sample of the abbreviations listed in the Stanford Abbreviation Database, and observed that 99.2% followed the pattern: 'long-form (short-form)' in MEDLINE titles and abstracts rather than the other three patterns. To further validate this assumption, after creating the ADAM database we selected a random sample of 1000 pairs that are in ADAM but not in the Stanford database. For each pair, we counted the following four occurrence frequencies within MEDLINE titles and abstracts: f_1: frq[long-form (abbr.)]; f_2: frq[(long-form) abbr.]; f_3: frq[abbr. (long-form)]; f_4: frq[(abbr.) long-form]. In 98% of cases, they were expressed in text as 'long-form (abbr.)'.

Why capture 3N words in the context. Chang *et al*. (2002) demonstrated that the correct long-form can always be found within 3N words of an acronym-type abbreviation, and our results also suggested that this applies to non-acronyms as well (data not shown). The long-form had to lie within the same sentence as the short-form, using a Perl program for sentence boundary identification (http://l2r.cs.uiuc.edu/~cogcomp/tools.php).

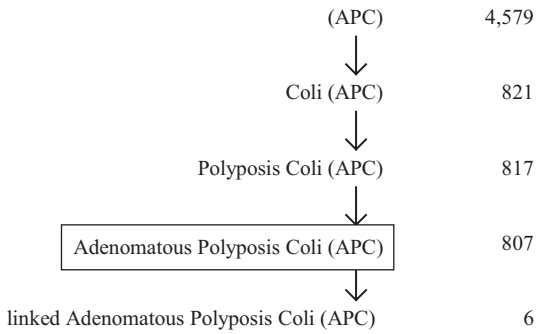

Fig. 1. A simple example of identifying candidate long-forms from contexts.

Morphologically similar candidate short-forms were put together and treated as variants of the same term. For example, 'APC' could be written as 'APC', 'Apc', 'ApC', 'aPC', 'APc', 'apc', 'AP-C' or 'Ap-C' in the literature. These words are the same after removing '-' and changing to the same case. By grouping similar short-forms, more statistical information is gathered, which assists in identifying their long-forms.

2.2 Step 2: identify candidate long-forms

This section describes the task of identifying candidate long-forms within the block of 3N words lying to the left of a short-form (or its lexical variants) given in parentheses. For example, 'APC' (or its variations 'Apc', 'ApC', etc.) has been mentioned in parentheses 4579 times in 4472 articles. The phrase 'Adenomatous Polyposis Coli' has occurred 807 times in 705 of these articles next to '(APC)' on the left. How can we recognize this long-form as the proper expansion of 'APC', rather than shorter or longer long-forms, without making use of any lexical information, such as matching of letters?

We start from '(APC)' and examine the counts for each step in the progression '(APC)' → 'Coli (APC)' → 'Polyposis Coli (APC)' → 'Adenomatous Polyposis Coli (APC)' → 'linked Adenomatous Polyposis Coli (APC)' (Fig. 1). Notice that the count drops significantly from 'Adenomatous Polyposis Coli (APC)' to 'linked Adenomatous Polyposis Coli (APC)'. In this example, 'Adenomatous Polyposis Coli' is determined as a candidate long-form.

The whole process is formalized and divided into several steps, as discussed below: first, tokenize the contexts; second, count the number of times each k-gram ($1 \leq k \leq 3N$) occurs in the contexts; third, determine the candidate long-forms; fourth, get rid of redundant candidate long-forms.

2.2.1 Tokenize the contexts This step is to remove delimiters, such as periods, commas or parentheses, and change the texts into lower case. There are many abbreviations for chemical compounds or substance names that have special nomenclature, e.g. '5-hydroxytryptamine(3)' is the long-form for '5HT(3)'. We want to keep these chemical names in their original forms as much as possible, so we kept all the inner parentheses, brackets and commas.

2.2.2 Count the number of times each k-gram occurs Given a context, 'hospitalized patients who tested positive for HIV' for 'HIV', the following k-grams ($1 \leq k \leq 9$) are extracted: 'virus' (single-word), 'immunodeficiency virus' (bi-gram), . . ., and 'hospitalized patients who tested positive for human immunodeficiency virus' (9-gram). For each

distinct k-gram, we count the number of times it has occurred immediately to the left of the same candidate short-form in all the contexts.

2.2.3 Determine the candidate long-forms For each k-gram, we seek to assign it a score that will be used to determine the candidate long-forms for a given short-form. One possibility is to use the raw proportion of counts (pr) as our score, e.g. if a long-form occurs to the left of the short-form in 50 out of 100 (pr = 50%) total occurrences, then the score would be 0.5. However, the situation 2/4 and 50/100 are not equivalent; although they have the same value 0.5, 50/100 is better than 2/4 because statistically, the SD gets smaller when the sample size is larger. We adjusted the raw proportion pr by the SD assuming that the count $[w_{i+1}w_i \ldots w_2w_1(w) \,|\, w_iw_{i-1} \ldots w_2w_1(w)]$ follows a binomial distribution (Dunning, 1993). Here, w is the candidate short-form enclosed in parentheses and $w_iw_{i-1} \ldots w_2w_1$ is a sequence of words left to the open parentheses. This adjusted proportion is denoted by apr. As a result, 50/100 and 2/4 are different: apr(50/100) = 0.45 > apr(2/4) = 0.25.

To restate this, the adjusted proportion (apr) is defined as follows: Given a short-form w and a k-gram $w_kw_{k-1} \ldots w_2w_1$, the adjusted proportions are defined as:

$$\mathrm{apr}_i = \mathrm{pr}_i - 2*\sqrt{\frac{\mathrm{pr}_i*(1-\mathrm{pr}_i)}{\mathrm{count}[w_{i-1} \ldots w_2w_1(w)]}}, \quad 1 \le i \le k \tag{1}$$

and pr_i is defined as:

$$\mathrm{pr}_i = \frac{\mathrm{count}[w_iw_{i-1} \ldots w_2w_1(w)] - 1}{\mathrm{count}[w_{i-1} \ldots w_2w_1(w)]}, \quad 1 \le i \le k, \tag{2}$$

where $\mathrm{count}[w_iw_{i-1} \ldots w_2w_1(w)]$ is the number of times the i-gram $w_iw_{i-1} \ldots w_2w_1$ occurs in the contexts. As an example, the adjusted proportion for the transfer 'Polyposis Coli (APC)' → 'Adenomatous Polyposis Coli (APC)' is computed as:

$$\mathrm{pr} = \frac{807-1}{817} = 0.9877; \mathrm{apr} = 0.9877 - 2*\sqrt{\frac{0.9877*(1-0.9877)}{817}} = 0.9785.$$

Subtracting 1 from $\mathrm{count}[w_iw_{i-1} \ldots w_2w_1(w)]$ is to distinguish between cases like 4/4 and 40/40. Without subtracting, apr(4/4) = apr(40/40) = 1; after subtracting, apr(40/40) = 0.9503 > apr(4/4) = 0.5335. Intuitively, the adjusted proportion will indicate the likelihood of the i-gram being a phrase or a part of a phrase.

In the progression '(APC)' → 'Coli (APC)' → 'Polyposis Coli (APC)' → 'Adenomatous Polyposis Coli (APC)' → 'linked Adenomatous Polyposis Coli (APC)' (Fig. 1), the apr scores for 'Coli', 'Polyposis Coli', 'Adenomatous Polyposis Coli' and 'linked Adenomatous Polyposis Coli' are 0.1790, 0.9884, 0.9785 and 0.0006, respectively. The apr score drops significantly during the progression 'Adenomatous Polyposis Coli (APC)' → 'linked Adenomatous Polyposis Coli (APC)' and thus 'Adenomatous Polyposis Coli' is determined as a candidate long-form.

Determining the cut-off value for the adjusted proportion. At what point should one stop the progression $[w_iw_{i-1} \ldots w_2w_1(w) \rightarrow w_{i+1}w_i \ldots w_2w_1(w)]$ and deem the resulting phrase to be a candidate long-form? To assist in assigning optimal rules, we examined the way in which long-form scores were distributed for the 691 638 pairs (single-word abbreviations with two or more alphanumeric characters, scored as EXCELLENT or GOOD) in the Stanford Abbreviation Database. We extracted all the k-grams from the long-forms and counted the number of times they occur in the contexts in MEDLINE titles and abstracts, and then computed the adjusted proportions for all the k-grams. We chose 0.05 as the cut-off value for the adjusted proportion, which means that if the apr drops below 0.05 when $w_iw_{i-1} \ldots w_2w_1(w)$ is expanded to $w_{i+1}\ w_i \ldots w_2w_1(w)$, then $w_iw_{i-1} \ldots w_2w_1$ is determined as a candidate long-form.

In some cases, the apr_1 for a valid long-form could be very small. For example, 'aqueous protein concentration' is another valid long-form for the abbreviation 'APC'. In the progression '(APC)' → 'concentration (APC)' → 'protein concentration (APC)' → 'aqueous protein concentration (APC)' → 'between aqueous protein concentration (APC)', their counts are 4795, 10, 9, 9 and 1, respectively. The corresponding apr scores are 0.0019, 0.5470, 0.6793 and 0. In this case, apr_1 (0.0019) is far below 0.05 and we will lose 'aqueous protein concentration' for 'APC' according to our cut-off value. We found out that this problem is caused by the ambiguity of 'APC', which has 27 different long-forms in ADAM. 'Aqueous protein concentration' is not a frequent long-form for 'APC'. To capture relatively minor but valid long-forms, we decided not to apply the cut-off criteria to apr_1 and instead we required that $\mathrm{count}[w_1(w)]$ be ≥10.

In summary, given a short-form w and all its k-grams ($1 \le k \le 3N$, N is the number of alphanumeric characters in w), the criterion for determining its candidate long-forms is given as follows:

A unigram w_1 is a candidate long-form if

$$\begin{cases} \mathrm{count}[w_1(w)] \ge 10 \\ \mathrm{apr}_2 < 0.05 \end{cases}$$

A k-gram, $w_kw_{k-1} \ldots w_2w_1$ ($2 \le k \le 3N$) is a candidate long-form if

$$\begin{cases} \mathrm{count}[w_1(w)] \ge 10 \\ \mathrm{apr}_i \underset{2 \le i \le k}{\ge} 0.05 \\ \mathrm{apr}_{k+1} < 0.05 \end{cases}$$

2.3.4 Get rid of redundant candidate long-forms In the example of 'Adenomatous Polyposis Coli (APC)', another longer phrase, 'mutations of Adenomatous Polyposis Coli (APC)', has been mentioned in 129 citations. Our method identified both 'Adenomatous Polyposis Coli' and 'mutations of Adenomatous Polyposis Coli' as acceptable candidate long-forms for 'APC', since the adjusted proportions for both were over 0.05. However, although 'mutations of Adenomatous Polyposis Coli' is a compound phrase, APC is not used by authors as shorthand to refer to that entire phrase.

To choose among multiple candidate long-form phrases that are both acceptable candidate long-forms based on the criterion described above, we measured how much the apr changes when 'Adenomatous Polyposis Coli' is expanded to 'mutations of Adenomatous Polyposis Coli'. In this example, the apr has decreased significantly [(0.9706 − 0.1339)/0.9706 = 86.20%). The idea is that if the relative difference in apr is greater than a certain amount, the candidate long-form with the lower apr can be eliminated as being a redundant or less-preferred candidate. Again, to determine the optimal cut-off values for the change in apr, we examined the features of single-word abbreviation/long-form pairs in the Stanford Abbreviation Database, where the long-forms listed were rated EXCELLENT or GOOD and hence are thought to be already optimized for acronyms. For each long-form in the Stanford database, we examined its contexts in MEDLINE titles and abstracts and computed its long-form score as above. By expanding the long-form one word further to the left within the MEDLINE contexts, and seeing how the apr changes accordingly, we observed that the apr decreased 18% or more after the expansion for 95% of the long-forms listed in the Stanford database.

The process of eliminating redundant candidate long-forms is described as follows: given two candidate long-forms of the same short-form, $w_m \ldots w_2w_1$ and $w_n \ldots w_m \ldots w_2w_1$, $m < n$ and $w_m \ldots w_2w_1$ is part of $w_n \ldots w_m \ldots w_2w_1$. The change of apr is defined as:

$$\Delta\mathrm{apr} = \frac{\mathrm{apr}_m - \mathrm{apr}_n}{\mathrm{apr}_m}, \tag{3}$$

if $\Delta\mathrm{apr} \ge 0.18$, remove $w_n \ldots w_m \ldots w_2w_1$, otherwise, remove $w_m \ldots w_2w_1$. apr_m and apr_n are the adjusted proportions for the two candidate long-forms.

Scoring the long-forms: the last adjusted proportion [apr_k in Equation (1)] is assigned as the score of the candidate long-form and represents the

proportion of cases in which w_k appears, given $w_{k-1} \ldots w_2 w_1(w)$. Note that the long-form with the highest score may not be the most frequently used. If the first word of the long-form is on an official PubMed stopword list consisting of 132 extremely common words, such as 'the' or 'by' (http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords), the first word is removed from the long-form.

2.3 Step 3: filter the short-form/long-form pairs according to a rule of length ratio and an empirically-validated cut-off value

Long-forms are usually much longer than their abbreviations. We used their length ratio (long-form_length/short-form_length, length is defined as the number of alphanumeric characters) to filter the candidate abbreviation/long-form pairs. 95% of the single-word abbreviation/long-form pairs in the Stanford Abbreviation Database have length ratios $\geq$2.5, and we also chose this value as our minimal cut-off value for the length ratio of ADAM candidate pairs.

2.4 Step 4: verify that the short-forms are used in text separately from their long-forms

This step verifies that the candidate abbreviation has been used as a free-standing term; the idea is that if the candidate has not appeared on its own, it is unlikely to represent a shorthand way of referring to another entity. To do this, we look at the titles and abstracts of articles which mention the 'long-form (abbreviation)' and check whether the abbreviation also occurs outside the parenthesis in the same article. For example, in the titles and abstracts that mention 'North Carolina (USA)', the term 'USA' has never been used on its own, so 'North Carolina (USA)' is removed from the candidate abbreviation/long-form pair list. At this step, 6212 pairs were removed, only 5% of which were judged to be useful pairs.

2.5 Step 5: group together morphologically similar long-forms that correspond to the same short-form (or its lexical variants)

Three distinct types of morphologically similar long-forms were observed in the ADAM database: (1) Plural, e.g. 'antigen presenting cells (APC)' versus 'antigen presenting cell (APC)'; (2) Hyphen or catenation, e.g. 'human papilloma viruses (HPV)' versus 'human papillomaviruses (HPV)'; (3) Extra words, e.g. 'adenomatous polyposis coli (APC)' versus 'adenomatous polyposis coli gene (APC)' versus 'adenomatous polyposis coli gene product (APC)'.

These morphologically similar long-forms represent the same meaning, and grouping them together will show how the same long-form could be written by different authors in the literature. For type 1 and 2 similar long-forms, an approximate string matching algorithm was employed (Gusfield, 1997). This algorithm uses dynamic programming techniques to calculate an edit distance between the source string and the target string, i.e. can one transform the target string to the source string using at most k additions, deletions, and substitutions? For example, 'human papilloma viruses' can be converted into 'human papillomaviruses' by deleting one space. A maximal value of $k = 2$ was chosen to group similar long-forms. The idea is to match strings that are nearly identical, in contrast to the BLAST-like method described in Krauthammer *et al*. (2000) which matches similar names that may diverge appreciably. For type 3 similar long-forms, two long-forms were clustered together if one of them is $w_m \ldots w_2 w_1$ and the other is $w_m \ldots w_2 w_1 y_k \ldots y_1$, i.e. they are overlapped word by word in the beginning.

3 RESULTS

All 15 433 668 citations (i.e. titles and abstracts) in the 2006 baseline of the MEDLINE database (http://mbr.nlm.nih.gov/) were examined, of which about half (7 806 798) contain abstracts. Before grouping morphological variants, 512 314 abbreviation/long-form pairs were identified using the methods described above. After grouping morphological variants, ADAM consists of 59 405 abbreviation/long-form pairs.

3.1 Error analysis

To measure the quality of the abbreviation/long-form pairs in ADAM, we first investigated how many false assignments (i.e. frank errors) were present in the database. In two random samples of 1000 distinct abbreviation/long-form pairs, 23 and 29 errors were found, giving an error rate of about 2.6%. Three types of errors were observed:

(1) Some errors (8/52) were inherent in the assumptions made in our model. For example, in a few cases, the abbreviation did not lie to the right of the long-form, but was embedded in the middle of it. For example, in the case of 'electron (EM) microscopic examination', 'electron' was extracted as the long-form of 'EM' and 'microscopic' was missed.

(2) Sometimes there is no standard long-form for the abbreviation. 24/52 errors were of this type. For example, for the abbreviation 'CelB', the system identified the candidate long-form as '*Pyrococcus furiosus*', whereas the candidate long-form should be 'the beta-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus*.' This occurred because there were many different ways of writing the equivalent long-form (e.g. 'hyperthermostable beta-glycosidase from *Pyrococcus furiosus*'). For these errors, the long-forms usually have more than three words, which indicates that the longer the long-form is, the more likely it could be written in different ways.

(3) In some cases (20/52), the same abbreviation referred to multiple long-forms that have different beginning words but end with the same word or sequence of words. For example, 'CCQ' could be 'Cancer Coping Questionnaire', 'Cocaine Craving Questionnaire' or 'Common Core Questionnaire'. No one of these long-forms dominated or occurred frequently. Our model assigned 'Questionnaire' as the long-form.

These errors are infrequent and are best regarded as incomplete assignments of the long-form. As mentioned below, our web interface links the abbreviation/long-form pairs with the contexts where they are defined within PubMed abstracts, allowing the user to see the correct long-forms immediately.

3.2 Coverage of the ADAM database, compared to Stanford and UMLS databases

Different abbreviation databases are created for different purposes. The Stanford Biomedical Abbreviation Server (http://abbreviation.stanford.edu/) uses lexical heuristic rules to extract abbreviation/long-form pairs that are well matched in letters. Note that 85.3% of the pairs in the Stanford database occur only once in MEDLINE. The UMLS SPECIALIST Lexicon 2005 (http://umlsks.nlm.nih.gov/) is manually curated and covers commonly occurring English words as well as biomedical vocabulary. Our intent was that ADAM should include the common abbreviation/long-form pairs in the biomedical domain, including both acronyms and non-acronyms.

To compare the coverage of ADAM with the Stanford Abbreviation Database and with the UMLS Lexicon, we considered only

Table 1. Types and examples of non-acronym short-forms and their long-forms

Type	Examples
1. Chemical compound	Tritiated Water (^3H_2O) Mercury (Hg)
2. Gene or substance name	Aromatase gene (CYP19) interstitial collagenase (MMP-1)
3. Synonym[a]	Day of birth (P0) leukosialin (CD43)
4. Hyponym[b]	Fibroblast cell line (3T3) rat glioma (C6)
5. Metonymy[c]	Gamma radiation (^{60}Co) phototherapy (UVB)
6. Regular word[d]	2,6,10,14-tetramethylpentadecane (pristane) hydroxymethylglutaryl coenzyme A reductase inhibitors (statins)
7. Brand name or manufacturer	human insulin (NOVO) braun oral-B ultra plaque remover (D9)

[a]The abbreviation and long-form have similar or identical meanings and are interchangeable in the context used by the author.
[b]The abbreviation is more specific than the long-form; it can also be described as 'a-kind-of', 'a type of', or 'an instance of'.
[c]The abbreviation refers to a feature of the long-form.
[d]The short-form is a regular word, not a typical abbreviation, but is used nonetheless as standard shorthand for a long-form.

those pairs that obeyed similar criteria across all databases, namely: (1) the abbreviations are single-words with two or more alpha-numeric characters; (2) the long-form is at least 2.5 times longer than the abbreviation and (3) the abbreviation/long-form pairs occur at least 10 times across MEDLINE in which the abbreviation is in parenthesis and the long-form is on the left. Of this cohort, ADAM contains 93.5% of those listed in the Stanford Abbreviation Database and 92.4% of those in the UMLS SPECIALIST Lexicon.

3.3 Abbreviation/long-form pairs only in ADAM

About one-third (18 293) of the abbreviation/long-form pairs in ADAM are not included in either the Stanford database or UMLS Lexicon at all. To assess these abbreviations, 300 of these novel pairs were selected randomly. These could be classified into three types: Type I (78.3%) consisted of acronym-type abbreviations that appeared more recently in MEDLINE than 2001. (Note that ADAM was built upon the 2006 baseline of MEDLINE, whereas the Stanford database was created as of 2001.) For example, 'shRNA' is an abbreviation for 'short hairpin RNA', which was first defined in PubMed in 2002. Type II (2.6%) consisted of abbreviations that are not strict acronyms, e.g. 'Brain Creatine Kinase (CKB)', or that are acronyms in languages other than English, e.g. 'Spanish Collaborative Study of Congenital Malformations (ECEMC)'. Type III (19.1%) consisted of frank non-acronym abbreviations. The frank non-acronyms comprised at least seven different types of short-forms and the corresponding long-forms (Table 1).

4 DISCUSSION

Recognizing abbreviations and their long-forms from biomedical articles has been an active area of NLP research interest (Yoshida *et al*., 2000; Pustejovsky *et al*., 2001; Wren and Garner, 2002; Chang *et al*., 2002; Yu *et al*., 2002; Liu and Friedman, 2003; Adar, 2004; Ao and Takagi, 2005; Egorov *et al*., 2005; Gaudan *et al*., 2005). Most of the existing techniques were developed for identifying acronyms based on hand-crafted patterns or rules (see Wren *et al*., 2005 for a comprehensive review). Though less numerous than acronyms, we believe that non-acronym abbreviations are also important to capture, e.g. K_m for 'Michaelis constant', or '11p' for 'short arm of chromosome 11'. To date, only one systematic method for capturing non-acronym abbreviations has been reported (Liu and Friedman, 2003). Like Liu and Friedman (2003), we have analyzed statistical information about collocations of the type 'long-form (abbreviation)' in MEDLINE. However, our method differs in the specifics of our inclusion criteria, the modeling and scoring of long-forms, and choice of numerical cut-offs. ADAM also clusters morphologically similar abbreviations and long-forms together and treats them as single terms.

The paired short-forms and long-forms within the ADAM database may be useful for a number of text mining projects. Identifying the long-forms associated with a given abbreviation may help disambiguate the meaning of a given instance of the abbreviation in text, may assist information retrieval or information extraction applications, and may assist in query expansion. For example, ADAM is being used to assist in classifying, ranking and merging B-terms in the Arrowsmith two node search interface (see Introduction; Swanson and Smalheiser, 1997; Smalheiser, 2005) (http://arrowsmith.psych.uic.edu). In addition, we have also used ADAM to find lexical variants of terms used in PubMed queries for research submitted to the 2006 Genomics Text Retrieval Conference (TREC) (http://ir.ohsu.edu/genomics/2006protocol.html). Although ADAM is not designed to be a lexicon for gene or protein related terminology, we found that about three-fourths of gene-related terms used in the 2006 TREC questions were listed either as short-forms or long-forms within ADAM.

ADAM can be accessed freely via a public Web-based query interface, or alternatively, it can be downloaded in its entirety as a text file. Using the Web interface, users can enter an abbreviation and retrieve its long-forms, or enter a long-form and retrieve its abbreviations. Pairs are displayed ranked by their counts, i.e. the number of occurrences in MEDLINE; optionally, users can also rank them according to their long-form scores. For each abbreviation/long-form pair, a link takes the user to the actual PubMed citations (and sentences therein) where the highlighted abbreviation pair is defined.

ACKNOWLEDGEMENTS

The authors thank Jeff Chang and Russ Altman (Stanford University) for graciously providing their abbreviation database. This research is supported by NIH Grants LM 007292 and LM 08364. Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of Interest: none declared.

REFERENCES

Adar,E. (2004) SaRAD: a simple and robust abbreviation dictionary. *Bioinformatics*, **20**, 527–533.

Ao,H. and Takagi,T. (2005) ALICE: an algorithm to extract abbreviations from MEDLINE. *J. Am. Med. Inform. Assoc.*, **12**, 576–586.

Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.*, **2001**, 17–21.

Chang,J.T. *et al.* (2002) Creating an online dictionary of abbreviations from MEDLINE. *J. Am. Med. Inform. Assoc.*, **9**, 612–620.

Dunning,T. (1993) Accurate methods for the statistics of surprise and coincidence. *Comput. Linguistics*, **19**, 61–74.

Egorov,S. *et al.* (2004) A simple and practical dictionary-based approach for identification of proteins in medline abstracts. *J. Am. Med. Inform. Assoc.*, **11**, 174–178.

Friedman,C. (2000) A broad-coverage natural language processing system. *Proc AMIA Symp.*, **2000**, 270–274.

Gaudan,S. *et al.* (2005) Resolving abbreviations to their senses in medline. *Bioinformatics*, **21**, 3658–3664.

Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, NY.

Jensen,L.J. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nature Rev. Genet.*, **7**, 119–129.

Krallinger,M. and Valencia,A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, **6**, 224.

Krauthammer,M. *et al.* (2000) Using BLAST for identifying gene and protein names. *Gene*, **259**, 245–252.

Liu,H. and Friedman,C. (2003) Mining terminological knowledge in large biomedical corpora. *Pac. Symp. Biocomput.*, 415–426.

Pustejovsky,J. *et al.* (2001) Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo.*, **10**, 371–375.

Smalheiser,N.R. (2005) The Arrowsmith project: 2005 status report. *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin, **3735**, pp. 26–43.

Stead,W.W. *et al.* (2005) Achievable steps toward building a National Health Information infrastructure in the United States. *J. Am. Med. Inform. Assoc.*, **12**, 113–120.

Swanson,D.R. and Smalheiser,N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.*, **91**, 183–203.

Wren,J.D. and Garner,H.R. (2002) Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods Inf. Med.*, **41**, 426–434.

Wren,J.D. *et al.* (2005) Biomedical term mapping databases. *Nucleic Acids Res.*, **33**, D289–D293.

Yoshida,M. *et al.* (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, **16**, 169–175.

Yu,H. *et al.* (2002) Mapping abbreviations to full forms in biomedical articles. *J. Am. Med. Inform. Assoc.*, **9**, 262–72.