Evaluation of Document Processing Techniques Towards *Escherichia coli* Bibliographic Database

Hisayuki Horai[1,2], Tomoya Baba[1], Kouichi Doi[1], Tomohiro Mitsumori[1],
hisayu-h@is.naist.jp, tmbaba@gtc.naist.jp, doy@is.naist.jp, mitsumor@is.naist.jp

Natsuko Yamamoto[1], Hirotada Mori[1], and Hirofumi Doi[1,2]
nyamamot@gtc.naist.jp, hmori@gtc.naist.jp, and doi@cl-sciences.co.jp

[1] Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan.
[2] Celestar Lexico-Sciences, Inc., MTG D17, 1-3 Nakase, Mihama-ku, Chiba 261-8501, Japan.

Abstract

We are developing *Escherichia coli* bibliographic database, a comprehensive accumulation of knowledge from articles. Because of massiveness of articles and requirements of high exhaustiveness and accuracy, automatic computer supports based on document processing techniques are contemplated to be essential to success. This paper reports the feasibility study of document processing techniques for *Escherichia coli* bibliographic database development. The techniques and methods introduced and evaluated in this paper is so general that it can be applied to develop not only a *Escherichia coli* bibliographic database but also many bibliographic databases specialised for other species. Furthermore, these techniques and methods are applicable for developing an genomic ontology.

大腸菌文献データベース構築に向けた文書処理技術の評価

蓬莱尚幸[1,2], 馬場知哉[1], 土井晃一[1], 三森智裕[1],
hisayu-h@is.naist.jp, tmbaba@gtc.naist.jp, doy@is.naist.jp, mitsumor@is.naist.jp

山本奈津子[1], 森浩禎[1], 土居洋文[1,2]
nyamamot@gtc.naist.jp, hmori@gtc.naist.jp, and doi@cl-sciences.co.jp

[1] 奈良先端科学技術大学院大学
〒630-0192 生駒市高山町 8916-5
[2] セレスター・レキシコ・サイエンシズ株式会社
〒261-8501 千葉市美浜区中瀬 1-3 MTG D17, MTG D17

概　要

我々は文献からの知識を幅広く集めた大腸菌文献データベースを構築中である。大量の文献を対象に高精度かつ漏れなく知識を集めるためには、文書処理技術に基づく計算機支援の重要性が高い。本報告では、大腸菌文献データベース開発のための文書処理技術の評価について述べる。これらの評価対象の技術や方法論は一般的なものであり、大腸菌以外の生物種に関する文献データベースの開発にも利用可能である。さらに、これらの技術や方法論は、ゲノムオントロジーの開発にも適用可能である。

1 Introduction

In bio-medical domain, National Library of Medicine (NLM) provides a fundamental bibliographic database MEDLINE [7] and many researchers utilise them via the internet. Despite their generality and conciseness, researchers are awaited anxiously by more accurate, exhaustive and well-structured bibliographic database in their specific research domain.

We are developing a bibliographic database specialised for *Escherichia coli* research. The currently developing version of the database has at least three advantages over MEDLINE:

- It includes only *E. coli* 'research'. When a researcher retrieves articles from MEDLINE, the results include not only articles on *E. coli* itself but also ones on other subjects where *E. coli* is just an experimental material. The latter articles should be eliminated for *E. coli* research use.

- It is classified. *E. coli* research covers a wide range of biological functions, such as metabolism, regulation, replication, translation and so on. Articles in the database are classified in such functions. Because an article often covers a plural number of functions, multiple classifications are adopted.

- Genes and proteins are annotated. Genes and proteins are extracted from each article and they are annotated to it. It is useful to access to the related entries of gene/protein databases, and to retrieve articles by gene/protein names.

These advantages might be appreciated by *E. coli* researchers, and does not violate the usefulness of MEDLINE for general purpose.

Generally speaking, editing activity determines the feasibility of a bibliographic database. In order to reduce the amount and difficulty of the editing activity, computer supports based on document processing techniques [2] play a vital role. Because the effective techniques for achieving each bibliographic database depend upon the database and its domain, the performance of each technique must be evaluated during the initial stage of a project and in the track of it. This paper reports the initial evaluation of some document processing techniques applicable to the *E. coli* bibliographic database. The current performance of each technique is evaluated using a criterial dataset with correct answer made by researchers. Furthermore, it is discussed on the practical supports by these techniques at the current level and some future works for the improvement of these techniques.

2 Methods

The editing activity of the *E. coli* bibliographic database, called 'the database' hereafter, includes the following tasks: selection of *E. coli* articles, classification of articles, and annotation of genes and proteins. From the viewpoint of document processing, each task is related to a specific sub-domain of document processing. In other words, some different techniques in a specific sub-domain of document processing are applicable to each task. Selecting of *E. coli* articles, classification of article and annotation of genes and proteins are related to document retrieval, multi-value classification and information extraction, respectively. Subsections from 2.1 to 2.3 explain each technique. Furthermore, subsection 2.4 explains the evaluation method.

The database is aimed to comprehensive accumulation of knowledge obtained from research articles of *E. coli*. For collecting *E. coli* articles exhaustively, all techniques to be applied must be based on commonly available information for every article. All techniques and methods in this paper utilise MEDLINE records as the input. Each MEDLINE record includes the title, the abstract, some bibliographic information, and some special annotation by NLM.

2.1 Selection of *E. coli* Articles

This subsection explains three techniques applicable for the supports to the task to select *E. coli* articles from a large number of MEDLINE articles.

2.1.1 MeSH Term

MeSH [8] is a controlled vocabulary for retrieving articles for MEDLINE, and NLM annotates MeSH terms to each article in MEDLINE. MeSH is regarded as a promised material for retrieval of MEDLINE. Furthermore, for further refinement of the retrieval, MeSH terms annotated to an article are classified into major terms and minor terms. A major term of an article is more principal subject than a minor term of the article. Two methods are introduced for selection of *E. coli* articles:

[Method 1.1] Retrieve articles which annotated MeSH term 'Escherichia coli'.

[Method 1.2] Narrow the search to the articles which annotated major MeSH term 'Escherichia coli'.

In this paper, the effectiveness of MeSH terms for selecting *E. coli* articles is evaluated. Furthermore, through the comparison between Method 1.1 and 1.2, the effectiveness of the difference of major and minor MeSH terms for selecting *E. coli* articles is clarified.

2.1.2 Exact Match

String exact match [2] is efficient for document retrieval when the retrieval objective can be easily represented in a restricted set of expression. From our experiences, 'Escherichia coli' and 'E. coli' are used as a text representation for *E. coli* in articles generally and frequently. Based on the empirical finding, the following method is introduced.

[Method 1.3] Retrieve articles whose title or abstract includes string 'Escherichia coli' or 'E. coli'.

Our experiences lead to another empirical finding. In the case that the strings using Method 1.3 exists in the title of an article, the probability that the article is an answer is quite high. Based on the finding, the following method seems to be accurate.

[Method 1.4] Retrieve articles whose title includes string 'Escherichia coli' or 'E. coli'.

In this paper, the effectiveness of each method is evaluated. Furthermore, through the comparison between Method 1.3 and 1.4, the accuracy of the empirical findings concerning the difference between a title and an abstract is clarified.

2.1.3 Support Vector Machine

Machine learning, especially support vector machine (SVM) [3], is regarded as a general and powerful tool for document processing. In this paper, the following naive methods are introduced and evaluated for selection of *E. coli* articles.

[Method 1.5] The feature of SVM is the Bag-of-Word in titles and abstracts. If a word exists in the title or the abstract of an article, then the value concerning the article and the word is 1. Otherwise, the value is 0.

[Method 1.6] The feature of SVM is the Bag-of-Word in titles. If a word exists in the title of an article, then the value concerning the article and the word is 1. Otherwise, the value is 0.

[Method 1.7] The feature of SVM is the Cartesian product of Bag-of-Word and its location, i.e. 'title' or 'abstract'. If word W exists in the title of an article, then the value concerning the article and the value of feature (W, article) is 1. If word W exists in the abstract of an article, then the value concerning the article and the value of (W, abstract) is 1. Otherwise, the value is 0.

Method 1.5 is the most naive method to apply SVM to this task. Method 1.6 and 1.7 are introduced based on the empirical finding concerning titles mentioned in 2.1.2. Because the words in a title are fewer than the ones in an abstract, the recall of Method 1.6 is worse than Method 1.5, but the precision of Method 1.6 is better than Method 1.5 if the empirical finding is true. Method 1.7 is intended to achieve a happy medium.

The evaluation system is developed by means of TinySVM [4]. The kernel of SVM utilised is polynomial kernel of degree 2.

When machine learning is applied to a specific problem, training data are needed. Training data in this paper is selected from the criterial dataset for the comparative evaluation of methods. The details of the criterial dataset and its usage for training data are explained in subsection 2.4.

2.2 Classification of Articles

This subsection explains two techniques applicable for the supports to the task to classify *E. coli* articles. These techniques can be applicable to predict the classification of an article using its title and abstract provided in MEDLINE.

2.2.1 Tf·idf

Tf·idf [2] is an automatically computable index of correspondence of a word to a document, and generally used in document processing. Term frequency of word w_i in document d_j (tf(i, j)) is the number of existence of w_i in d_j. Document frequency of word w_i (df(i)) is the number of documents which includes w_i. Inverted document frequency (idf(i)) of word w_i is calculated by the following formula:

$$\mathrm{idf}(i) = \log\left(\frac{N}{\mathrm{df}(i)}\right)$$

where N is the total number of documents.

Basically, tf(i, j) is positively correlated to the correspondence of w_i to d_j because if a word appears in a document frequently then the document discusses the concept denoted by the word time and again. In contrast, df(i) is negatively correlated to the correspondence of w_i to d_j because words which appears universally in many documents, c.f. pronoun, conjunction and article, do not have strong correspondence to any documents. Both correlations are merged to index tf·idf, i.e. the arithmetic product of tf(i, j) and idf(j). Hereafter, tf·idf is called 'weight' and denoted by w(j, i).

Document d_j can be characterised by a vector V(j) = (w(1, j). w(2, j), ..., w(M, j)) where M is the number of words. Similarity between V(a) and V(b) in M-dimension vector space is regarded as the similarity of documents w_a and w_b (Sim(a, b)). There are some choices for calculation of the similarity in vector space. The methods in this paper utilise cosine similarity, i.e. the cosine of the angle between V(a) and V(b) calculated by the following formula:

$$\mathrm{Sim}(a,b) = \frac{\sum_{i=1}^{M}[w(i,a)\cdot w(i,b)]}{\sqrt{\sum_{i=1}^{M}[w(i,a)^2]}\cdot\sqrt{\sum_{i=1}^{M}[w(i,b)^2]}}$$

Many prediction methods based on this type of similarity have been proposed and successfully accepted in some projects of document processing. In this paper, one of the most naive methods is introduced.

[Method 2.1] The calculation of similarity is based on the weight of words in title and abstract. The classification of an *E. coli* article X is predicted to the classification of another *E. coli* article which is most similar to X.

2.2.2 Support Vector Machine

SVM is applicable to multi-value classification. Multi-value classification is decomposed to a set of independent selection tasks. Each selection task is corresponding to a category of the classification one by one. SVM is applied for each category of the classification independently.

[Method 2.2] The feature of SVM is the Bag-of-Word in titles and abstracts. The value concerning a document and a word is equal to the weight of the word to the document mentioned in 2.2.1 and calculated in Method 2.1.

While a simple existence of a word in an article is used in selecting *E. coli* article (Method 1.5 to 1.7), an accurate correspondence of a word in an article is used in Method 2.2.

The evaluation system is developed by means of TinySVM. The kernel of SVM utilised is polynomial kernel of degree 2 as same as Methods 1.5 to 1.7.

2.3 Extraction of Genes and Proteins

This subsection explains two techniques applicable to extraction of genes and proteins from each *E. coli* article. These techniques are aimed to extract adequate strings from titles and abstracts in MEDLINE.

2.3.1 Exact Match

E. coli is one of species relatively well-investigated at the genome level, and it can be expected that all genes of *E. coli* are enumerated in relatively small amount of efforts. The list of *E. coli* genes provided by EcoCyc [5] is available to use as a source of a dictionary for extracting genes. In the EcoCyc list, more than one gene names are described for every gene. For instance, the version dated May 9th, 2005 includes 4,476 genes and 5,307 gene names. Because a gene name of *E. coli* is usually capitalised to represent the product of the gene, the gene names from EcoCyc can be expanded to protein names. The EcoCyc list includes 16 capitalised and 5,291 uncapitalised gene names. Finally, the dictionary generated from the EcoCyc list automatically, called 'EcoCyc Dictionary' hereafter, consists of the 16 + 5,291 = 5,307 items.

[Method 3.1] Extract every word exactly matched to an item in EcoCyc Dictionary.

Furthermore, names are often modified in articles. In order to cover with such modified names, some specific expansion rules to every item in EcoCyc Dictionary are introduced. These rules can be obtained empirically.

[Method 3.2] Extract every word satisfied at least one of the following conditions:

- Exactly matched to an item EcoCyc Dictionary.
- Prefixed by an item in EcoCyc Dictionary.
- Exactly matched to the concatenation of 'Delta' and an item in Expanded EcoCyc Dictionary.

2.3.2 Pattern Match

From observation to the EcoCyc list and some articles, the following six patterns of gene/protein names are found empirically and inductively.

- P1: 1 alphabet + 2 small letters + 1 capital letter
- P2: 1 alphabet + 2 small letters + 1 capital letter + any string
- P3: 'Delta' + 1 alphabet + 2 small letters + 1 capital letter
- P4: 'sigma(' + any string + ')'
- P5: 'p' or 'P' + integer
- P6: words suffixed by 'ase' or 'ases'

While pattern P1 to P5 are patterns for symbol like gene and protein names, P6 is a pattern for common noun like expression. These patterns are listed in the order of specificity, i.e. P1 is the most specific pattern and P6 is the most general one. In this paper, these patterns are evaluated in three levels of their specificity.

[Method 3.3] Extract every word matched pattern P1.

[Method 3.4] Extract every word matched at least one of patterns P1 to P4.

[Method 3.5] Extract every word matched at least one of patterns P1 to P6.

2.4 Evaluation Method

Each method is evaluated in the following steps:

1. Select a criterial dataset which consists of a number of articles, and the correct answer is made by researchers.
2. For each method, predict the answer for the criterial dataset automatically.
3. For each method, calculate its performance in comparison of the prediction results with the correct answer.

The criterial dataset in this paper consists of 920 articles of Journal of Bacteriology in 2001. Researchers read these articles, select *E. coli* articles, classify every selected article, and extract gene and protein names from each selected article.

In order to predict by methods based on SVM, training data are needed. For predicting each article of the criterial dataset, the other 919 articles of the criterial dataset are used as training data. For predicting each article of the selected *E. coli* articles, all selected articles except the target article are used as training data.

For evaluating the performance of each method, the following three measures are utilised. These measures are commonly used in the domain of document processing [2]. In the following formulae, TP, FP and FN denotes true positive, false positive and false negative, respectively.

- $\text{Precision P} = \dfrac{\text{TP}}{\text{TP} + \text{FP}}$.
- $\text{Recall R} = \dfrac{\text{TP}}{\text{TP} + \text{FN}}$.
- $\text{F-measure} = \dfrac{2 \cdot \text{P} \cdot \text{R}}{\text{P} + \text{R}}$,

  i.e. the harmonic mean of precision and recall.

Table 1: Classification of *E. coli* articles.

Category	Number of classified articles
1. Amino acid metabolism	8 (4.5 %)
2. Biosynthesis of cofactor, prosthetic group, carrier	1 (0.6 %)
3. Cell envelope	12 (6.7 %)
4. Cellular process	14 (7.9 %)
5. Central intermediary metabolism	2 (1.1 %)
6. Energy metabolism	5 (2.8 %)
7. Fatty acid / phospholipid metabolism	4 (2.2 %)
8. Nucleotide metabolism	6 (3.4 %)
9. Regulatory function	76 (42.7 %)
10. Replication	10 (5.6 %)
11. Transport / binding protein	13 (7.3 %)
12. Translation	12 (6.7 %)
13. Transcription	35 (19.7 %)
14. Other categories	40 (22.4 %)
15. Hypothetical	2 (1.1 %)

3 Results

3.1 Criterial Dataset and Correct Answer

The criterial dataset in this paper consists of 920 articles of Journal of Bacteriology in 2001. Researchers read these articles and selected 178 *E. coli* articles (19.3% of criterial dataset). The selected articles are also used as the targets for the classification task and the gene/protein extraction task.

Researchers classify the selected *E. coli* articles into 15 categories, based on GenoBase [1,6]. Table1 shows the results. In 9 categories, only less than 10% of the *E. coli* articles are classified. Especially, only one article is classified into category 'Biosynthesis of cofactor, prosthetic group, carrier', and only two articles are classified into 'Central intermediary metabolism' and 'Hypothetical'.

In extracting gene/protein names, at least one but no more than twenty names are extracted from each *E. coli* article. The results are 898 extractions of 702 different names.

3.2 Selection of *E. coli* Articles

The results of evaluation are shown in Table 2. Fundamentally, all of three different techniques are effective for selection of *E. coli* articles. Especially, the methods based on MeSH are quite effective (f-measure of Method 1.2 is 0.886).

The comparison of precision between Method 1.1 (67.5%) and 1.2 (92.6%) suggests that the major/minor annotation on MeSH terms by NCBI contributes to precision largely.

The comparison of precision between Method 1.3 (48.8%) and 1.4 (92.6%) suggests that titles are more important than abstracts. This suggestion is also derived from the comparison between Method 1.5 (68.7%) and 1.6 (92.6%).

Table 2: Selection of *E. coli* Articles.

Method	Precision
1.1	67.5%
1.2	92.6%
1.3	48.8%
1.4	92.6%
1.5	68.7%
1.6	92.6%
1.7	90.4%

The recall of Method 1.3 is not 100%. It suggests the existence of *E. coli* articles which have neither string 'Escherichia coli' nor 'E. coli' in the title or the abstract.

The comparison between Method 1.4 and 1.6 suggests that other words than 'Escherichia coli' and 'E. coli' contribute little or nothing to the selection.

While the comparison between Method 1.5 (68.7% precision, 50.6% recall and 0.583 f-measure) and 1.7 (90.4% precision, 69.1% recall and 0.783 f-measure) suggests again that the difference of a title and an abstract is quite important, the comparison between 1.6 (92.6%

precision, 77.5% recall and 0.844 f-measure) and 1.7 suggests that words in an abstract have negative effects on the selection.

3.3 Classification of Articles

The results of evaluation are shown in Table 3. Fundamentally, both methods are not effective. Especially, for categories whose correct answers are less than or equal to 10, both methods score 0. For Category 9 and 13, both methods result in relatively high score, where there are relatively many correct answers. These observations suggest that the performance of both methods is largely affected by the number of articles classified in categories. It may concludes that the reason of the low effectiveness of both methods is the paucity of *E. coli* articles in the criterial dataset, and it will be expected to improve the performance as the project progressed.

For categories 1, 3, 12 and 14, Method 2.2 can find correct answers, while Method 2.1 cannot. In contrast, for category 4, Method 2.1 can find correct answers, while Method 2.2 cannot. It is weekly suggested that Method 2.2 is more effective to Method 2.1 in many cases of small number of correct answers, but the opposite case remains to be insoluble.

Table 3: Classification of Articles.

	Method 2.1			Method 2.2		
Category	Precision	Recall	F-measure	Precision	Recall	F-measure
1	0%	0%	0	10.0%	12.5%	0.111
2	0%	0%	0	0%	0%	0
3	0%	0%	0	22.2%	16.7%	0.190
4	15.4%	14.3%	0.148	0%	0%	0
5	0%	0%	0	0%	0%	0
6	0%	0%	0	0%	0%	0
7	0%	0%	0	0%	0%	0
8	0%	0%	0	0%	0%	0
9	45.0%	65.8%	0.535	45.3%	38.2%	0.414
10	0%	0%	0	0%	0%	0
11	8.2%	30.8%	0.129	12.5%	7.7%	0.095
12	0%	0%	0	18.2%	16.7%	0.174
13	22.6%	20.0%	0.212	22.2%	22.9%	0.2252
14	0%	0%	0	12.7%	15.0%	0.138
15	0%	0%	0	0%	0%	0

3.4 Extraction of Genes and Proteins

The results of evaluation are shown in Table 4. The precision of Method 3.1 is not 100%. It suggests that there are some items in EcoCyc Dictionary which cause false positive prediction. The comparison of precision between Method 3.1 (90.9%) and 3.3 (93.0%) suggests that the problematic items in EcoCyc Dictionary may not be matched to the pattern of Method 3.1, i.e. 'aaaA' type.

The comparison of recall between Method 3.1 (45.9%) and 3.3 (48.9%) suggests that EcoCyc Dictionary is not sufficient ,and some 'aaaA' type of gene names of *E. coli* still remains out of the EcoCyc gene list. The results of Method 3.3, 3.4 and 3.5 suggest that these kind of relaxation of matching results in increase of recall and decrease of precision as same as a usual document processing application. It is clear that Method 3.5 is quite less precision (54.4%) and f-measure (0.572) than any other methods

Table 4: Extraction of Genes and Proteins.

Method	Precision	Recall
3.1	90.9%	45.9%
3.2	70.0%	52.9%
3.3	93.0%	48.9%
3.4	70.0%	57.1%
3.5	54.5%	60.2%

4 Discussion

Because the evaluation results in high performance for the task to select *E. coli* articles, the effective support for the task is possible. For instance, the combination of Method 1.2 and 2.2 achieve 91.2% precision and 92.7% recall. This combined method will contribute to increase the entries of *E. coli* bibliographic database quickly.

In contrast, it is difficult to invent more effective methods, because there are no spaces to expand MeSH methods and it may be quite hard to find another effective word for exact match. Methods based on SVM have possibility to be improved when the number of training data increases. Many positive examples are expected to be obtained practically using the above mentioned supports, while it is difficult to obtain negative examples practically. Considering current high performance, the efforts to obtain negative examples is not so worthy to expend.

For the task to classify *E. coli* articles, it is difficult to provide effective supports currently. In order to increase the classified items in *E. coli* bibliographic database, researchers' efforts to classify *E. coli* articles is practically essential. The performance of methods in this paper, especially methods based on SVM, is expected to be improved according to the increase of these items. There are some rooms for improvement to methods based on SVM, but the primary issue is not the method itself but the number of training data.

For the task to extract gene and protein names, the methods in this paper are quite immature. Direct use of EcoCyc gene list as a dictionary causes problems. For instance, the evaluation mentioned in the previous section suggests that some items in EcoCyc Dictionary cause false positive prediction. After the evaluation, we can find gene names 'map' and 'tag' in EcoCyc Dictionary. These common nouns like items in a dictionary usually cause false positive prediction. Furthermore, insufficiency or incompleteness of EcoCyc Dictionary is also suggested. The correct answers include some long protein names suffixed by 'ase', other types of long protein names, and names which consist of a plural number of words. EcoCyc Dictionary does not include these types of names, and pattern P6 causes much false negative prediction. The following solutions are considerable:

- Enrich dictionary using other resources.
- Invent more specific patterns than P6.
- Apply machine learning.

A practically efficient use of current methods to extract names may be restricted to suggest each predicted string as a necessary part of an article to be check.

Considering a bibliographic database for a specific species other than *E. coli*, the three tasks mentioned in this paper, i.e. selection of articles for the species, classification of each selected article into adequate categories and extraction of genes and proteins from each article, are unavoidable to achieve the bibliographic database, and it is essential to keep high accuracy and exhaustiveness of the database that computer supports based on document processing techniques are provided to the creators and editors of the database. The techniques and methods introduced and evaluated in this paper are so general that they can be applied to develop not only a *Escherichia coli* bibliographic database but also many bibliographic databases, each of which is specialised for another specific species.

Furthermore they are applicable to develop an genomic ontology. The editing activity to maintain the relation between genes and research articles is critical to introduce references and detailed research findings into an genomic ontology. The document processing techniques evaluated in this paper provides a basis for supporting to the editing activity.

References

[1] Riley, M., Functions of the gene products of Escherichia coli, *Microbiol. Rev.*, 57(4):862-952, 1993.

[2] Salton, G., Automatic text processing - the transformation, analysis, and retrieval of information by computer, Addison Wesley, 1989.

[3] Vapnik, V. N., The nature of Statistical Learning Theory, Springer, 1995.

[4] http://chasen.org/~taku/software/TinySVM/

[5] http://ecocyc.org/gene-links.dat

[6] http://ecoli.aist-nara.ac.jp/GB5/search.jsp

[7] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[8] http://www.nlm.nih.gov/pubs/factsheets/mesh.html