1、General introduction

1.1 Motivation
In BC Workshop 12, a system for a specific biocuration task based on text mining or natural language technology is demanded. To a system of this kind, we are interested in acetylation. In our previous work and research, we did some studies about this subject but never developed a system for it.

1.2 Background
Heavy jobs have been done to acetylation. A great many academic papers have been published. Most of them are stored in database. However, there is no database, which specifically targets acetylation. In this situation, it is troublesome for people in this territory locate the literature they need.

1.3 Our advantages
First, in our previous work, something relevant to natural language processing in the biological domain has been done. Second, we accumulated technology for general problems in natural language processing.

1.4 Feasibility
In our discussion, we found out there were some troubles in retrieving resources in the biological domain. With feedback from the people researching in the biological domain, we selected the final aim--acetylation. To prevent the barriers beyond us, we simplified the system into a level which is within our ability.
1.5 Innovation

In the general database, various kinds of academic literatures are stored together. There is no database which is made specifically for acetylation. It is troublesome and time-consuming to retrieve the relevant literature. The system we will make is to cope with this trouble.
2. Interface

The interface is as follows:

There is an input box for searching and the links for the high-frequency terms-searching.

The searching results will be ranked according their relevance to acetylation. Each item of the results includes the title, authors, abstract and PubMedID. The full text is not available due to copyrights. The full text can be accessed by searching PubMedID.

*Figure 1 General interface*
Establishment of Methylation-Specific PCR for the Mouse p53 Gene.

Okawaki R, Ootsumma A, Yoshida Y, Nomura T.

Abstract

Methylation-specific PCR (MSP) of the mouse p53 gene has not yet been reported. We searched the CpG islands, sequenced the bisulfite DNA, and designed PCR primers for methylation and unmethylation sites. DNA from a young mouse produced a strong PCR product with the unmethylated primer and a weaker band with the methylated primer. DNA from an old mouse produced bands of similar intensities with both primers. In radiation-induced tumors, DNA from an old mouse yielded similar bands with both types of primers. We suggest that MSP is a valuable technique for the epigenetic study of the mouse p53 gene.

PMID: 22201902
3. **Working description**

There are two parts in the program including the part for documents pre-processing and documents-retrieving.

3.1 **Pre-processing**

First, the abstract of each paper is read and the document object is constructed. After that, the index, segmentation will be done. Following that, the score based on the relevance to acetylation will be given. Set the score as $A$, which will be recorded in the index file.

3.2 **Retrieving**

In this process, there are two steps. The first is search. The second is ranking.

**Search**: After the user inputs a keyword, the search for the key words based on the index will be done. The results will be stored in hit.

**Ranking**: Give the score to each doc according to the general search. The score is based on TF-IDF calculation. Set this score as $B$. The final score of the doc will be given as the combination of $A$ and $B$ (\( S = A \cdot x + B \cdot y \)), where $x$, $y$ is the weight of basic relevance and ranking respectively. After that, the output will be given in the descending order of the final scores.
Figure 3 Overall description

Preprocessing

Original document

Extraction

Document

Construct index

Index

Retrieval

KeyWord

By index

Hits

Re-rank by scores.

Result

By index

Result
Specific Procedure:

Document pre-processing

First, the program fetches information about the documents and extracts PMID, Title, Abstract, Author, RelevancyScore, and construct a field to each of them, which represents the properties of the document. After that, the program will construct index with words.
Read documents

For each document:

Field: PMID
Field: Author
Field: Title
Field: Abstract

Iterate.

Construct Documents containing all fields

Calculate Relevancy Score

Add to index

**TOKENIZED**
Both tokenization and indexing will be done.

**UN_TOKENIZED**
Tokenization will not be done, but index will be done.

**NO_NORMS**
Index will be done, but score will not be given.
Document Indexing

Based on indexing, the key words the user inputs will then be matched. In ranking, the weights of title and abstract are different. The score of the document will be calculated as the score of the document related to the keyword. After that, the weighted mean of RelevanceScore will be calculated to generate the overall scores of the documents. The output will listed in descending order.

*Figure 5 Document indexing*

```
Keyword

Query by Index

Hits

Iteration

Frequency of words in Title and abstract will be given different weights and scores will be calculated accordingly.

Get the stored RelevancyScore of acetylation.

Combine the two scores

Descending order

Output
```
4、Program description

The structure of the project is as follows

Figure 6 Project structure

word.dic is used to store the key words and the relevance score.
The folder WEB-INF store configuration information and reference libs the whole project.

The folder page store s jsp pages.

WEB-INF/src stores java codes.

com.ProAce is the prefix of packages ,including all the codes of this project.

Filter is used to store the retrieving results and score it for the second time.

Index is the code for indexing.

PreTre contains the codes for pre-processing.

Search contains the codes relevant to searching.

app.properties stored some information in configuration, which will be read out by PropertyConfiguration.
Appendix

Scoring algorithm for the documents of acetylation for a single word

For the relevance of key words to index and acetylation, we consider issues in two aspects. First, the relevance of each document and acetylation should be got from the database. Second, the relevance of the key word and the documents, which are relevant to acetylation, should be got.

To the relevance of a document and acetylation, we can get accurate value in pre-processing. The algorithm is as follows.

Scoring for the key word

We looked up to a great many resources relevant to acetylation and searched for the key words which may be relevant to acetylation. After that, select n key words which are close to acetylation. The number n will be set according to the specific experiment performance. According to the relevance between the document and acetylation, we set a rational value for each of the key word manually.

For acetylation, after we made the value table for the key words which are closely related to acetylation, after we studied some resources. This step is crucial. And it is also subjective. The values are due to adjustment based on the experimental results.
### Table 1 Scores assignment for key words

<table>
<thead>
<tr>
<th>No.</th>
<th>Key word</th>
<th>Score (ten as the bigger number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GNAT</td>
<td>9.5</td>
</tr>
<tr>
<td>2</td>
<td>NAT</td>
<td>8.5</td>
</tr>
<tr>
<td>3</td>
<td>nsN-alpha-terminal Acetylation</td>
<td>8.5</td>
</tr>
<tr>
<td>4</td>
<td>HAT and HDAC</td>
<td>9.0</td>
</tr>
<tr>
<td>5</td>
<td>Histone acetylation transferase , HAT</td>
<td>9.0</td>
</tr>
<tr>
<td>6</td>
<td>acetylation</td>
<td>8.5</td>
</tr>
<tr>
<td>7</td>
<td>deacetylation</td>
<td>8.0</td>
</tr>
<tr>
<td>8</td>
<td>Hat1</td>
<td>7.8</td>
</tr>
<tr>
<td>9</td>
<td>PACF</td>
<td>6.8</td>
</tr>
<tr>
<td>10</td>
<td>Sas2</td>
<td>5.0</td>
</tr>
<tr>
<td>11</td>
<td>SIRT1~~SIRT7</td>
<td>6.5</td>
</tr>
<tr>
<td>12</td>
<td>ARD1</td>
<td>5.0</td>
</tr>
<tr>
<td>13</td>
<td>SAGA</td>
<td>7.6</td>
</tr>
<tr>
<td>14</td>
<td>NuA4</td>
<td>7.2</td>
</tr>
<tr>
<td>15</td>
<td>Rtt109</td>
<td>7.8</td>
</tr>
<tr>
<td>16</td>
<td>Rtt109_Vps75</td>
<td>8.2</td>
</tr>
<tr>
<td>17</td>
<td>P53</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Suppose there are n key words $K_i$ ($1 \leq i \leq n$), the score of them is $V_i$ ($1 \leq i \leq n$).

We employed the scoring rule in BM25. We give scores to the document based on the score of key words to represent the relevance of a document and acetylation. The scoring algorithm is as follows.
\[ SA(D_j) = \sum_{i=1}^{N} \left[V_i \cdot \frac{f(K_i, D_j) \cdot (k_i + 1)}{f(K_i, D_j) + k_i \cdot (1 - b_i + b_i \cdot \frac{|D_j|}{\text{avgdl}})} \right] \]

* \(SA(D_j)\) represents the score of document \(D_j\) with respect to acetylation.

* \(f(K_i, D_j)\) represents the frequency of existence of the \(i\)th key word in Document \(D_j\)

* \(|D_j|\) represents the length of document \(D_j\) (i.e. The number of words)

* \(\text{avgdl}\) represents the average length of all the documents in the database

Another two parameters \(k_i\) and \(b_i\) are used to adjust the precision. The default values of them are \(k_i = 2, b_i = 0.75\).

According to this, we will get the relevance of each document and acetylation. For the consistence of measurement, the relevance is normalized.

\[ A(D_j) = \frac{SA(D_j) - \min_{1 \leq k \leq N} \{SA(D_k)\}}{\max_{1 \leq k \leq N} \{SA(D_k)\} - \min_{1 \leq k \leq N} \{SA(D_k)\}} \]

To the relevance of the words in the input of the user and the document, we employed BM25 algorithm and got the following scoring algorithm.

\[ SB(q, D_j) = IDF(q) \cdot \frac{f(q, D_j) \cdot (k_2 + 1)}{f(q, D_j) + k_2 \cdot (1 - b_2 + b_2 \cdot \frac{|D_j|}{\text{avgdl}})} \quad (\*) \]

\[ IDF(q) = \begin{cases} \log \left(\frac{N - n(q) + 0.5}{n(q) + 0.5} \right) & n(q) < 0.5N \\ 0 & otherwise \end{cases} \]
• $SB(q, D_j)$ the score of the relevance of the query word and Document $D_j$.

• $f(q, D_j)$ the frequency of relevance of the query word in Document $D_j$.

• $|D_j|$ represent the length of Document $D_j$ i.e. the number of words in $D_j$.

• $avgdl$ represents the average length of the documents in the entire database.

• $k_2$ and $b_2$ are used to adjust precision.

• $IDF(q)$ is inverse document frequency.

• $N$ is the total number of documents in the database.

• $n(q)$ represents the number of documents, in which, the key word $q$ exists.

We divide the whole document into three parts--Title abstract, main body. We gave scores to each of them and calculated the weighted sum.

For example: suppose the title of document $D_j$ is $D_{j1}$, the abstract of document $D_j$ is $D_{j2}$, the main body of document $D_j$ is $D_{j3}$. The score will be given in the following way.

\[
SB(q, D_{j1}) = \begin{cases} 
(k_3 + 1) \cdot IDF(q) & q \text{ exists in } D_{j1} \\
0 & \text{otherwise} 
\end{cases}
\]

\[
SB(q, D_{j2}) = IDF(q) \cdot \frac{f(q, D_{j2}) \cdot (k_3 + 1)}{f(q, D_{j2}) + k_3 \cdot (1 - b_3 + b_3 \cdot \frac{|D_{j2}|}{avgdl_2})}
\]

\[
SB(q, D_{j3}) = IDF(q) \cdot \frac{f(q, D_{j3}) \cdot (k_3 + 1)}{f(q, D_{j3}) + k_3 \cdot (1 - b_3 + b_3 \cdot \frac{|D_{j3}|}{avgdl_3})}
\]

$k_3$ and $b_3$ are used to adjust precision the default value of them are $k_3 = 2, b_3 = 0.75$
Because the measurements of relevance are different the scoring process in Title, Abstract and main body, we need to normalize them. Set the normalized score of Title, Abstract and main body as 

$$SB'(q, D_{j1}), SB'(q, D_{j2}), SB'(q, D_{j3})$$

The weighted average score of them are calculated as follows.

$$B(q, D_j) = \alpha \cdot SB'(q, D_{j1}) + \beta \cdot SB'(q, D_{j2}) + \gamma \cdot SB'(q, D_{j3})$$

This score is relevance score of the document and the query words.

$$\alpha > \beta > \gamma > 0$$, they are the weights of the three different parts \( \alpha + \beta + \gamma = 1 \). The default value is \( \alpha = \frac{1}{2}, \beta = \frac{1}{3}, \gamma = \frac{1}{6} \).

Combining the two steps above, we get the relevance of each document and acetylation and the relevance of the key word and each document. According to two relevancies, we can get the weighted sum, and get the final score of relevance between the key word and the document with respect to acetylation.

$$Score(q, D_j) = x \cdot A(D_j) + y \cdot B(q, D_j)$$

In it \( x, y > 0 \) and both \( x \) and \( y \) represent the weights.