MutationMiner: integration of text mining and database management to support protein mutation analysis

Jing Tang
Department of Mathematics and Statistics
University of Helsinki
Helsinki, Finland 00014
Email: jing.tang@helsinki.fi

***Abstract*—This paper gives a comprehensive review on a novel biological database management system called MutationMiner. This system was originally introduced in [1] for enhancing the automatic processing of protein mutation analysis, with an effort to directly link the text mining results from document retrieval with the target protein databases. The document retrieval step features Natural language processing (NLP) techniques for identifying the queried protein mutants from biological literature. The extracted information about protein mutation is further used to access various databases for obtaining the corresponding protein sequence and structure data. The review starts with describing the system architecture and then the implementation details of MutationMiner. A case study of using MutationMiner for *Xylanases* protein extraction is reviewed and some weakness of the system is discussed in the final section .**

***Index Terms*—Text Mining, NLP, MutationMiner, Protein mutation.**

I. Introduction

With the advances of molecular technology, biological research has come to an age of high-density data analysis. Large amounts of biological data have been accumulated at a daily base, so that salient biological databases are becoming increasingly important for the purpose of information collection and retrieval. Categories of databases for biological science include sequences databases and protein databases. There sizes could vary from less than 100Kb to more than 10Gb. Currently most of the biological databases have a web-interface and accessible with a bunch of associated bioinformatics tools. For example, the BLAST (Basic local alignment search tool) is an efficient tool kit for comparing gene and protein sequences in NCBI databank. [2]

When data information is simply obtained through databases, the analysis could be straightforward, since the input and output of databases are generally standardized. However, when data information is distributed in different sources, one needs to develop efficient methods to integrate these sources of information. This is often the real case. In fact, most of new research results are first published in scientific papers and then perhaps formulated into publicly accessible databases. Therefore biological databases tend to have a delayed response when processing those latest data. As a result, a large proportion of data information might be undiscovered by the biological community, even though it is actually accessible in literature. Furthermore, the expanding volume of scientific publication is far beyond what individual scientists can cope with. For this reason we need computer-aid techniques to enhance the speed of information processing.

The main contribution of the paper is the use of NLP (Natural Language Processing) as a text-mining technique, in an effort to extract biological information from research papers. The text-mining results are then automatically incorporated as inputs into biological databases, so that human interference is minimized. The authors described a case study called Mutation Miner that integrates protein mutation text mining into a protein structure database, such that the one can identify sequences of targeted proteins and obtain their 3D structures.

In the sequel, I will discuss the main topics of this paper and structure the remaining part of the review as follows. Section II gives general ideas of text mining and then focuses on the NLP technique. Section III describes the base framework of the introduced system, e.g. the architecture for combining text mining results with biological databases. Section IV covers the main application of their framework, e.g. using the MutationMiner system to retrieve protein mutation information. A case study of the system on *Xylanases* protein extraction is described in Section V. Section VI further discusses the performance of MutationMiner and compares it with another system called Mutation GraB. [3]

II. Text-mining systems

Recent work in the compute-aid techniques for biological databases has focused on biological text mining systems. Biological text mining systems are developed for the extraction of biological information from scientific literature. Two types of tasks are dominating in biological text mining: the extraction of gene and protein names and the extraction of protein-protein interaction. Popular examples include the BioRAT system that enables a user to define a searchable template and deliver

the corresponding extracted text segments from literature. For protein function retrieval, one can use ProFAL to retrieves documents and extracts functional properties from the text. [4]

A. Natural language processing

As a new text mining technique, Natural language processing (NLP) has acquired a lot of applications in biological document retrieval. NLP is a new computer-aid technique that processes of natural language by computers. The NLP in biological domain generally consists of three steps. [5]

1) Text handling. Text handling is a pre-processing step that prepare texts for further computer analysis. During this step, trivial words that contain little biological information are removed. Next, variations of the same word will be restored into its basic form. More sophisticated methods can even identify synonyms so that the vocabulary size is further reduced. After a basic vocabulary is formulated, each word will be tagged according to its syntactic and semantic classes. Syntactic classes define words as nouns, verbs, etc. and semantic classes assign words that are closely related in their biological meanings. Typical semantic classes include gene names, protein names, or drug names. Once a document has been appropriately pre-processed, the text will be represented in a structure that facilitates the subsequential computations.
2) Corpus analysis. This is the essential part of natural language processing which aims at document classification with machine learning approaches. Corpus analysis can be either supervised or unsupervised. Supervised methods require a training set of documents in which their classification labels are known. Popular supervised machine learning methods include *naive bayes* or maximum entropy methods. In contrast, unsupervised methods does not use a training set a prior and the classification of document data is determined by the computation algorithm alone. Popular unsupervised methods include K-means clustering and Principle component analysis (PCA). The output of corpus analysis is generally a summary for a set of text documents that categorizes them into sub-groups.
3) Information extraction. After a set of text documents are grouped by corpus analysis, information extraction methods are used to identify relations between objects of interest, such as gene-gene interactions and protein-protein interactions. Information extraction assumes that genes or proteins that frequently co-occur within documents might correlate in biological reactions. Accordingly by looking for sentences that describe both of the genes or proteins, one can discover the relationship between them. Note that the identification of plausible interaction information is generally formatted in a structured form. The result of information extraction therefore is readily acceptable by biological databases.

A typical NLP for biological information processing starts with a literature search that involves the text handling and corpus analysis, so that the set of candidate documents that might contain the information we want is determined. The information extraction is then carried out on the individual documents so that at least one known gene or protein of interest can be identified. Next, we need to explore the corresponding databases using the NLP result and find the target information.

III. Integration of biological databases and text mining

The provided architecture for connecting text mining results and biological databases is shown in Fig. 1. The primary goal is to facilitate information retrieval with text mining tools from NLP. It is based on the standard multi-tier information system design, in which four tiers are constructed.

The first tier is the clients terminal, which provides a user interface for humans. The client tier usually takes a form of web interface so that input and output of the system can be delivered through a web browser. Client tier could be accessed by humans, or any other automated clients.

The presentation and interaction tier which is next to the client tier receives requests from the client, analyzes and forwards them to the corresponding information retrieval modules. Results are then sent back to the corresponding client. This tier acts as a middle tier between the client tier and the data analysis tier.

The retrieval and analysis tier fulfills all the document retrieval functions as described in the previous section. The natural language analysis function is based on the GATE (General Architecture for Text Engineering) framework. GATE is a general database framework for developing software components for natural language processing. The framework and development environment are written in Java and available as open-source under the GNU licence. For more details, see [8]. In addition, various bioinformatics tools that are needed to analyze the NLP-derived results, are also provided in this tier. For example, to find regions with local similarity between gene or protein sequences, one can include BLAST (Basic local alignment and search tool) or ClustalW.

The resources tier is the place where documents are stored. The documents can be collected either directly from the Web, or from a document database, e.g., PubMed.

The system architecture with the four tiers enables the incorporation of content creation, retrieval and analysis. The standard document retrieval and presentation systems are supplemented with a natural language processing component that contains a number of customized bioinformatics tools. Such integration makes the system more intelligent than traditional separate components.

IV. The MutationMiner system

From now on, the review emphasizes a case study introduced by the authors. It is a biological document processing system developed with the architecture described above. First,

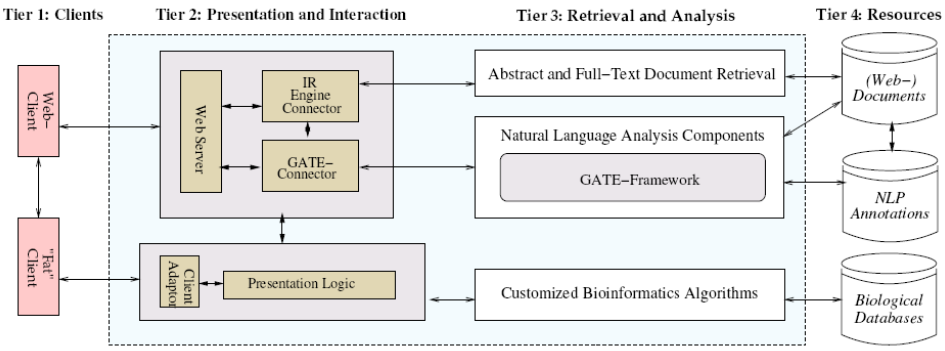

Fig. 1. System architecture for the integration of NLP with biological databases [1]

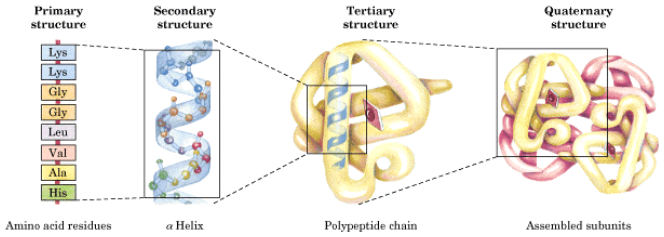

Fig. 2. Structure of proteins. Primary structure is the chain of amino acid sequence. Secondary structure includes alpha helix and beta sheet, which are local repeating structures stabilized by hydrogen bonds (i.e. chemical forces that connect molecules). Tertiary structure is the overall shape of a single protein molecule. The arrangement of two or more tertiary structures in a protein make up its quaternary structure.

the biological motivation for building such a system is explained in the following.

A. Protein mutation

Proteins are essential parts for organisms to perform their functions in biological processes. Many proteins, for example, are enzymes that act as catalysts in specific biochemical reactions and are vital to metabolism. Proteins are made of amino acids arranged in a linear sequence. Such a linear chain is usually folded into a 3-dimensional structure. The amino acid sequence for a protein is often called the *primary structure* and the high-level structures are often referred to as *secondary structure, tertiary structure* and *quaternary structure* (Fig. 2).

Evolution of proteins relies on structural changes, produced by mutations in the amino acid sequence and genetic rearrangements. A protein point mutation refers to as the substitution of a wild-type amino acid (i.e. normal amino acid) with a mutated one. Finding the biological impacts of protein mutation can be important to our understanding of protein functions. Researchers working on the mechanism of an enzyme often introduce mutations such that the importance of a particular residue to the enzyme function can be evaluated. Results of mutant proteins are often represented in scientific publications. Therefore it is important to extract protein mutation information from text documents such that the corresponding protein structures can be predicted.

B. Information extraction for protein mutation

The task of point mutation extraction involves first the identification of protein and mutation names discussed within an article. After these entities are identified, one needs to find a mapping between the point mutation and its correct protein family. Currently available protein mutation databases, such as PMD, can provide this capability. However, these databases are often manually managed, which limits the speed of processing newly published papers. In order to cope with the abundant information that emerges in scientific literature, we need automated processing techniques. Next, I will review on the text mining method applied in MutationMiner.

C. Text mining using NLP

As discussed in section II, NLP is an automated text mining technique to identify relations between objects from documents. Applying NLP to identify and relate proteins, organisms and point mutations is thus straightforward. MutationMiner utilizes a NLP text mining subsystem at a sentence level as in the following procedures.

1) Preprocessing and Gazetteering. Preprocessing identifies meaningful words as individual tokens and classifies them based on a number of precompiled concept categories. For example, we may want to identify words that are chemical names, drug name, gene or protein names and so on. Once words are identified, these words can then be matched by using a gazetteer.
   A gazetteer is a list of key words that describe a particular category. The gazetteers used in MutationMiner are based on three sources, namely MeSH(Medical Subject Hierarchy, Swiss-Prot and hand-made lists. The gazetteers annotate words with two-level categories: major category and minor (sub) category. [4]
2) Sentence Splitting and POS Tagging. During this step the input text is further split into individual sentences and for each sentence a part-of-speech (POS) tag was assigned using the Hepple tagger. POS is used to identify parts of speech in context. A simple form of POS is

the identification of words as nouns, verbs, adjectives, adverbs, etc. For more details on the Hepple tagger and POS, see [8].

3) Named Entity Recognition. This is the stage where individual tokens are combined into complex named entities (NEs). For example, a first name, last name and possibly initials form a name entity called *persons*. During this step *protein expressions* and *mutation expressions* can be identified.
4) Noun Phrase Chunking. Noun phrases are grammatical structures that are built based on named entities and POS tags. In this stage, all those noun phrases that contain a biological named entity will be identified.
5) Relation Detection. All the identified named entities will be compared in their noun phrase patterns. The sentences for the protein expressions and mutation expressions are scanned for identifying entities with matched noun phrase patterns. To ease the searching complexity, it is assumed that all point mutations are described in the sentences where the associated proteins are also mentioned. This is of course a simplified assumption. For complex cases especially when multiple protein mutations are described in a paper, such an assumption does not hold in general. More discussions concerning the assumption justification in various situations will be given in a later section.

In summary, when searching for relations between point mutations and proteins in the literature, MutationMiner relies on a NLP method that provides a set of annotated sentences categorized according to the noun phrase structures. The protein and mutation terms can be identified in the form of Name entities (NEs). After these entities are identified, a correct association can be made between the mutation and its corresponding protein. Furthermore, the text mining in MutationMiner is able to find the taxonomic origin of the associated protein so that we can correctly retrieve amino acid sequences from protein sequence databases.

D. Accessing protein sequence databases

The information obtained in the previous text mining step is further processed by accessing various biological databases. The proteins that have been identified as mutatations are queried into biological databases for the retrieval of their sequence and structure data. The subsequent analysis with biological databases is thus decomposed into two subtasks: protein sequence analysis and protein structure analysis.

1) Protein sequence analysis: MutationMiner accesses the *Entrez* system for the retrieval of protein sequences. The *Entrez* system is an integrated search engine that allows users (MutatinoMiner, in this case) to search the protein sequence databases at the National Center for Biotechnology Information (NCBI) website. The input provided by MutationMiner is a protein/organism pair and the output of the search is a list of candidate sequences in FASTA format.

The candidate sequences are compared in similarity using multiple sequence alignment (MSA). A MSA is often carried out by a bioinformatics tool called CLUSTER W. Only those sequences that satisfy certain similarity criteria are eligible for subsequent structure analysis. A consensus sequence will be generated according to the MSA result.

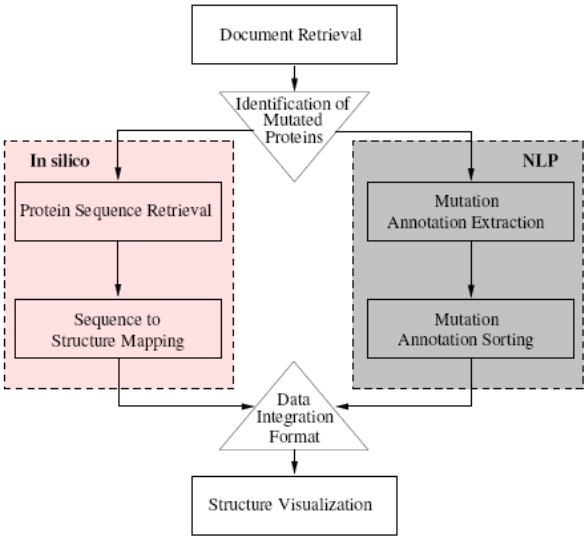

Fig. 3. The MutationMiner framework [7]

2) Protein structure analysis: Selection of a protein structure for its 3-d visualization is determined in a dynamical way. A dynamically selected structure is obtained by finding the top hit when comparing the candidate sequences with the sequences of structures obtained from the Protein Data Bank. The coordinates of the mutated positions on the structure sequence are determined using pairwise alignment with BLAST.

After the optimal structure with mutation information is determined, the 3-d visualization is directly obtained through the Protein Data Bank.

E. System integration framework

As shown in Figure 3, MutationMiner employees a framework by mixing NLP and sequence/structure analysis approaches. The protein mutation information that is described in selected documents is extracted using a series of NLP techniques. Their impact on protein function can be evaluated by accessing the corresponding protein sequence and structure databases. The MutationMiner architecture facilitates an integrated inference of the locations of mutation residues on eligible sequences, by mapping extracted mutation information to a protein homolog that is accessible from databases. The result is readable by protein visualization tools such as ProSAT.

V. Case study and results

To demonstrate the validity of MutationMiner, mutations to *xylanases* enzymes were chosen as the information of interest to us. *Xylanases* enzymes are a protein family that can depolymerize the plant cell wall component *xylan* to simple

Table 1. NLP subsystem partial evaluation results.

	Abstract only		Full paper	
	Protein/Organism	Mutations	Protein/Organism	Mutations
Precision	0.88	1.00	0.91	0.84
Recall	0.71	0.85	0.46	0.97
F-Measure	0.79	0.92	0.61	0.90

Fig. 4. NLP accuracy on the *Xylanases* text mining. [1] Precision, Recall and F-measure are three commonly used indicators to evaluate NLP algorithms in text mining. Precision is the fraction of the proteins or mutations retrieved that are relevant. Recall is the fraction of the proteins or mutations that are relevant to the query that are successfully retrieved. F-measure is a combination of precision and recall measures. F-Measure is defined as $F = ((1 + \beta^2) * Precision * Recall)/((\beta^2 * Precision) + Recall)$, with usually $\beta^2 = 1$. [10]

sugars. *Xylanases* are commonly used to remove xylan in the process of paper bleaching. *Xylanases* have been applied in many industrial processes and it has been known that under certain industrial conditions, e.g. high temperature, alkaline, *Xylanases* experience functional changes that are driven by point mutations. Every year there are numerous publications concerning mutations made to *xylanases* in order to improve their properties.

In the current case study 20 papers describing mutations to *xylanase* proteins were selected. The goal is to retrieve the protein sequences corresponding to these papers. Figure 4 shows the the preliminary results by applying the NLP text mining on either abstracts only or the full papers. It can be seen that the NLP text mining was able to identify most of correct protein names and taxonomic origins, especially when the full paper is included in the analysis.

The identified protein-mutation pairs are then used to search the protein sequence database *Entrez* for inferring the correct protein structures. (This part of results, however, was included only in [7] which seems an earlier implementation of MutationMiner than [1].) The candidate sequences were determined as those having greater than 70% similarity to each other. The mapping of candidate sequences to the structural homolog was achieved by pairwise alignment using BLAST.

With the 20 papers evaluated in the case study, 54 mutated amino acid residues were identified on two *xylanase* families, 14 on family 10 and 40 on family 11. Details of the protein structures can be found in [7].

VI. Discussions

Most of new research results in molecular biology are represented in the form of scientific journals. Currently available databases, however, lack efficient automated systems to extract the new information from text documents. The majority of this information is unfortunately handled by manual labor that prevents a large scale of database storage and access. Accordingly, text mining with the aid of computers is highly recommended for the identification and extraction of literature information.

This review gives a comprehensive description on the text mining method provided in [1], and the corresponding database integration framework. The system architecture, which is implemented in MutationMiner, is capable of automatically extracting mutation information from protein research literature and providing visualization tools for annotation of protein 3-d structures using the extracted information. The integration of text mining components and protein database retrieval provided a unified architecture which facilitate the automated data processing.

However, as pointed out by the authors, the MutationMiner system is still in its early stages of development. Some hidden weakness concerning both the text mining method and the database design might provide poor results and even invalidate their system in real applications. In the following I will discuss three major weakness in their system.

A. The NLP assumption

As mentioned in section IV-C, MutationMiner works with an assumption made on document structure such that all the mutation terms are supposed to occur in the same sentence as the associated proteins and organisms are described. The assumption is favorable by the authors since it decreases the search space down to a sentence level. However, it is often questionable for many of biological documents. First, most papers describe more than one protein mutation. In these cases, further processing is required to correct protein-mutation identification. One idea is to define a distance measure between two entities such that the significance of a plausible protein-mutation association is determined by the distance. Secondly, the spatial location of a mutation term is in some cases not a precise indicator. [3] argued that the frequency of a mutation term, instead of its actual position in the text, is more informative in disambiguation of protein-mutation associations. The ambiguity incurred by the simplified assumption in MutaionMiner would provide many false protein-mutation relations that are actually do not exist according to the literature.

B. The measures of NLP performance

In the case study where 20 papers on the xylanase protein mutation were analyzed, the performance of their NLP method was evaluated by computing precision, recall and F-measure for the identification of proteins and the identification of protein mutations separately (Fig 4). However, it would be more sensible if the same measures are computed for the identification of protein-mutation relations. In fact, the results presented in Fig 4 give little information about how well the NLP performs on the target problem, e.g. the correct identification of protein-mutation associations.

C. MutationMiner's availability

Although the authors published a series of papers based on the MutationMiner system, the system itself seems not available for public access. Therefore it is difficult to compare MutationMiner with other similar systems, such as Mutation GraB [3]. Note that the method underlying MutationMiner was published in 2005. It seems that for some reason the authors are not continuing this project any longer.

ACKNOWLEDGMENT

The author would like to thank Jan Lindström for his great supervision, and also thank two anonymous reviewers for their critical suggestions.

REFERENCES

[1] Ren Witte, Christopher J. O. Baker: Combining Biological Databases and Text Mining to Support New Bioinformatics Applications. NLDB 2005: 310-321.

[2] Altschul SF, Gish W, Miller W; Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 215: 403-10 (1990)

[3] Lawrence C Lee , Florence Horn , Fred E Cohen. Automatic Extraction of Protein Point Mutations Using a Graph Bigram Association. PLoS Comput Biol. 2007 Feb 2;3 (2)

[4] Corney, D. P. A., Buxton, B. F., Langdon W.B. and Jones, D. T. (2004) "BioRAT: Extracting Biological Information from Full-length Papers", Bioinformatics; vol. 20(17); pp.3206-13

[5] http://www.cse.ust.hk/dekai/600G/notes/KM_Slides_Text_KM_Mining.pdf

[6] Baclawski K, Cigna J, Kokar MM, Mager P, Indurkhya B. (2000) Knowledge representation and indexing using the unified medical language system. Pac Symp Biocomput. 493-504.

[7] René Witte. (2004) An Integration Architecture for User-Centric Document Creation, Retrieval, and Analysis.roceedings of the VLDB Workshop on Information Integration on the Web (IIWeb'04). 141-144.

[8] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[9] Christopher J. O. Baker and Ren Witte, Enriching Protein Structure Visualizations with Mutation Annotations Obtained by Text Mining Protein Engineering Literature. The Third Canadian Working Conference on Computational Biology (CCCB'04), October 4th, 2004, Markham, Ontario, Canada.

[10] Didier Nakache, Elisabeth Mtais, Jean Franois Timsit: Evaluation and NLP. DEXA 2005: 626-632