Combining Biological Databases and Text Mining to support New Bioinformatics Applications

René Witte[1] and Christopher J. O. Baker[2]

[1] Institute for Program Structures and Data Organization (IPD)
Universität Karlsruhe (TH), Germany
Email witte@ipd.uka.de
[2] Department of Computer Science and Software Engineering
Concordia University, Montréal (Québec), Canada
Email baker@encs.concordia.ca

Abstract. A large amount of biological knowledge today is only available from full-text research papers. Since neither manual database curators nor users can keep up with the rapidly expanding volume of scientific literature, natural language processing approaches are becoming increasingly important for bioinformatic projects.
In this paper, we go beyond simply extracting information from full-text articles by describing an architecture that supports targeted access to information from biological databases using the results derived from text mining of research papers, thereby integrating information from both sources within a biological application.
The described architecture is currently being used to extract information about protein mutations from full-text research papers. Text mining results drive the retrieval of sequence information from protein databases and the employment of algorithmic sequence analysis tools, which facilitate further data access from protein structure databases. Complex mapping of NLP derived text annotations to protein structures allows the rendering, with 3D structure visualization, of information not available in databases of mutation annotations.

1 Introduction

Biological researchers today have access to vast amounts of research data. Unlike in many other disciplines, these results are not only published in research papers, but additionally in a structured form within several publicly accessible databases. This data describes a unique array of information on biological entities such as DNA, proteins, and small molecules. A large proportion of salient information is however still hidden within individual research papers. Moreover, the rate at which new findings are being published is much higher than individual scientists or engineers can cope with, which is hindering further research and the development of industrial applications. For this reason, NLP techniques are progressively being applied in the area of biology.

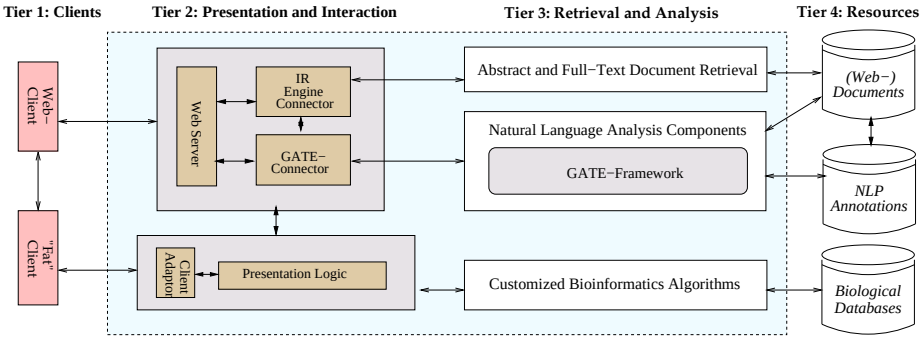

Fig. 1. System architecture for the integration of NLP with biological databases

Existing work in the area of biological text mining systems so far has focused on delivering extraction-based systems for biological research and database curation projects. Examples for such systems are: (1) The *BioRAT* system [3], which combines an information retrieval engine with an information extraction component based on user-definable templates (regular-expression based grammars). Users can then view extracted text segments instead of reading the full-text papers. (2) *ProFAL* (PROtein Functional Annotation through Literature) [4] is a system that annotates entries in biological databases with information found in the scientific literature, supporting manual database curation by proposing information from texts as supplementary data for biological entities. (3) *Textpresso,* an "ontology-based information retrieval and extraction system for Biological Literature" [9], which also aims at supporting biological database curation tasks.

While we also aim at supporting biologists through text mining, our work differs in that we want to provide a foundation for new biological applications by directly linking information obtained through NLP with the available biological databases. In other applications, like BioRAT or Textpresso, the textual results are meant for human consumption, so the often low precision of text mining systems is less critical. In our case, however, there is no human step between the NLP system and further application-specific processing. To ensure a reliable cross-linking is possible, NLP-derived results must be more rigorously structured, filtered, and analyzed before being used for bioinformatics applications. We demonstrate the feasibility of this idea with a biologically relevant application for the visualization of protein 3D structures.

2 Connecting Biological Databases and Text Mining

In this section we present an architecture for combining text mining results with biological databases and in-silico algorithms. It follows a standard multi-tier information system design, similarly to the application discussed in [13]. Figure 2 shows the main components, which we now discuss in detail.

Tier 1: Clients. The first tier provides access to the system, typically used by humans, but potentially also for other automated clients. Most services and data

will be delivered through a web browser, while some programs could require additional "fat clients" or *Java* applets, like a 3D-visualization component.

Tier 2: Presentation and Interaction. Tier 2 is responsible for information presentation and user interaction. In our architecture, it has to deal with both service access and content visualization. A connector for an information retrieval engine allows the dispatch of user queries to an IR system to obtain documents. Retrieved documents can then be queued for processing by an NLP system. Finally, it allows for the control of specialized in-silico applications, the interaction between the user agent, the processed NLP results, and the bioinformatics algorithms.

Tier 3: Retrieval and Analysis. Tier 3 provides all the document analysis and retrieval functions discussed above. In order to access biological documents, the architecture can be equipped with a stand-alone information retrieval engine like *Lucene* or a web-spidering component. The natural language analysis part is based on the GATE *(General Architecture for Text Engineering)* framework [5], one of the most widely used NLP tools. Since it has been designed as a component-based architecture, individual analysis components can be easily added, modified, or removed from the system. Finally, application-specific algorithms are needed to process the NLP-derived results, filtering and supplementing them with data from biological databases. These algorithms, in turn, can reference standard bioinformatics tools like BLAST [1] or CLUSTAL W [12].

Tier 4: Resources. Input documents (research papers) either come directly from the Web (or some other networked source, like emails), or a full-text database. Results from the NLP component are stored as annotations to the original documents in their own database. They can be queried for specific keywords (for example, finding all references to a particular protein), or exported to XML for exchange with other applications. Finally, in order to verify, process, and supplement the text mining results the architecture needs access to various biological databases, for example the PDB, Brenda, or *Entrez*.

3 Case Study: The MutationMiner System

In this section we present the *MutationMiner* system we have developed within the architecture described above. It combines text mining results from protein engineering literature with biological databases to support enhanced 3D structure visualizations of proteins [2]. We give preliminary results and outline areas for further improvement, which are discussed in more detail in section 4.

3.1 Biological Background

The motivation for this work is the ever-increasing amount of scientific literature detailing the effects of mutations to proteins. A bio-engineer working on the improvement of an enzyme, for example for its use within an industrial process, needs an understanding of the impact of all mutations carried out on the particular protein family. This requires a complex mapping of sequence mutants to a

common protein structure. Currently the protein mutation database (PMD) [7] and associated visualization tools can provide this capability. The content of this database is limited however by the speed at which newly published papers can be processed. In 1999 the PMD authors reported a three-year backlog of unprocessed publications. Since the arrival of high-throughput sequence modification techniques, such as directed evolution, a greater number of mutant sequences are produced along with information about their improved performance under precisely defined conditions.

Our goal, therefore, is to develop text mining tools that automatically scan literature and extract information about protein mutations. The extracted information can then be used to access protein sequence information from biological databases for use by a sequence alignment algorithm, which in turn queries protein structure information needed for 3D visualization. A protein engineer can then view structural representations of proteins (obtained from protein databases) combined with annotations describing mutations and their impacts (extracted through text mining from publications).

Protein Mutations in the Literature. Enzymes are proteins that catalyze specific biochemical reactions. Each enzyme family carries out conversions of distinct chemical substrates to chemical products. Within an enzyme family each individual enzyme has different physiochemical operating parameters, like temperature optimum, pH optimum, or thermal stability. Mutation of protein sequences has in many cases resulted in the production of enzymes with altered properties and is a common approach to enzyme improvement. Such mutations are typically the change of amino acids of the protein sequence achieved using molecular biology techniques such as site directed mutagenesis or directed evolution. The properties of the amino acids at specific positions on the protein sequence are the determining factor, however which amino acids and which positions are responsible for particular enzyme properties is not always known. For this reason, protein engineers routinely mutate residues and document their impacts on enzyme characteristics of special interest in scientific publications.

Structure Visualization. The complex structure of a protein is intrinsically related to its function and the elucidation and manipulation of protein structures to enhance protein function has valuable practical benefits. Protein structure visualization tools allow the protein engineer to view and rotate three-dimensional images in various representations. This in turn allows for the interpretation of experimental or computational results in a spatial context and facilitates the generation of hypotheses concerning the mechanistic interactions of the protein with substrate ligands. For these reasons, it is important to be able to link text mining results in an automated fashion to such 3D visualizations.

3.2 System Architecture and Implementation

The system, as outlined above, needs to integrate document retrieval, NLP-based text analysis, protein sequence database access, protein sequence analysis, and output format generation within a single architecture. Figure 2 shows the enhanced architecture based on the design presented in section 2.

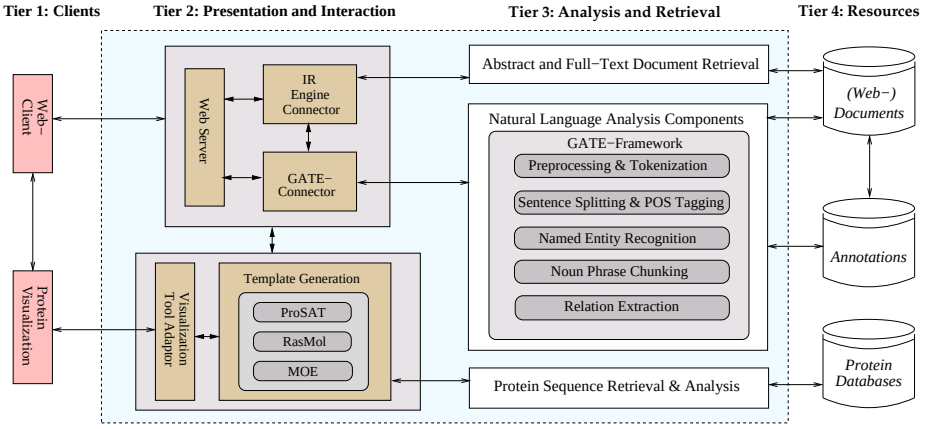

Fig. 2. MutationMiner System Architecture

Users interact with the system using a standard web client (tier 1). A web server (tier 2) receives a query (e.g., for a protein family) and dispatches it to an IR subsystem (tier 3), which retrieves relevant texts from the Web (e.g., NCBI's PubMed) or a local database (tier 4). Retrieved abstracts or full-length papers (if available) are then run through the NLP subsystem (tier 3) to identify mutations and extract relevant information. This information is then used by another tier 3 component to search *Entrez* in order to identify protein accessions and retrieve protein sequences from a biological sequence database. Mutated residues located on eligible sequences are then combined with the information extracted from the documents and converted into tool-specific output formats (tier 2). The user can then access the combined information through a protein visualization tool like ProSAT. In the remainder of this section, we discuss the individual components in more detail and show preliminary results.

Text Mining Subsystem. The NLP step needs to identify the proteins being mutated so that the corresponding amino acid sequence can be retrieved from a database. To do this the retrieved documents are run through an NLP subsystem that extracts proteins, host organisms, mutations, their interrelations, as well as provided accession numbers. A full text or abstract, once retrieved and converted into a suitable input format, is run through a so-called processing pipeline of GATE components, which we describe in more detail below.

Preprocessing and Gazetteering. After dividing the input stream into individual tokens in the *tokenization* step, a lookup phase identifies words and expressions based on a number of precompiled lists, like person names, companies, measurements, and biomedical-related lists, like chemicals, drugs, genetic structures, or protein names. Based on these lists, a *Gazetteer* component annotates words with a major and minor type, which forms a two-level hierarchy, similar to a (very simple) ontology. For non-biomedical information, we rely on lists contained in the ANNIE information extraction system that comes with GATE. Biomedical

lists use the same resources as the BioRAT system described in [3]: lists of entries extracted from the MeSH hierarchy and SwissProt, together holding more than five million words in roughly 650,000 entries.

Sentence Splitting and POS Tagging. The next two components split the input text into individual sentences and then, for each sentence, annotate each word with its *part-of-speech (POS)* tag using the Hepple tagger.

Named Entity Recognition. In the next stage, several finite-state transducers combine individual tokens into more complex named entities (NEs), based on regular-expression grammars and specialized tokenizers, which are run over the annotations generated by the previous steps. Examples for entities we detect are *persons* (containing a first name, last name, and possibly initials), *protein expressions*, or *database accession identifiers*. At this stage we also identify *mutation expressions*, which can occur in many different formats.

Noun Phrase Chunking. Another JAPE (finite-state transducer) grammar analyses the text and builds up more complex grammatical structures, so-called *noun phrases*, which include determiners, modifiers, and head nouns. For example, the words *"The specific enzyme activity"* will be identified as a single noun phrase (NP) with its words marked up as *"The/DET specific/MOD enzyme/MOD activity/HEAD."* An important feature of our NP chunker is its ability to incorporate the named entities detected above in addition to using POS tags. This allows us to alleviate some of the problems that result from using standard POS taggers, which are statistically trained on more general domains like newspaper articles, for biomedical documents. Finally, we mark all those noun phrase structures that contain a biological named entity.

Relation Detection. The last step is the correct identification and interpretation of relations between entities. For our task, we need to be able to identify two kinds of relations: between *proteins* and *mutations,* that is, which protein has been mutated within the described experiment; and between *proteins* and *taxonomic origin,* which we need to correctly retrieve amino acid sequences from protein sequence databases. For the protein-mutation identification, we currently extract all sentences that contain mutation expressions as identified by the corresponding NE grammar. We then scan these sentences for the protein expression, making the simple assumption that the protein mentioned together with the mutations must be the one that has been mutated. For example, in the sentence: *"Wild-type and mutated xylanase II proteins (termed E210D and E210S) were expressed in S. cerevisiae grown in liquid culture."* we identify two mutations, E210D and E210S, and one protein expression, *"xylanase II proteins,"* which we then assume is the protein being mutated. As this approach is quite simplistic, it might fail in a number of cases, especially when more than one protein mutation is described within a single paper. However, since we only extract those mutations where we can identify a corresponding host organism, this approach has been shown to work reliably within our case study on selected xylanase papers. For extracting the second (protein-host) relation we use a template-based approach that matches certain NP-NP patterns where one noun phrase contains

1: P36217. Reports Endo-1,4-beta-xyl...[gi:549461] BLink, Domains, Links

```
>gi|549461|sp|P36217|XYN2_TRIRE Endo-1,4-beta-xylanase 2 precursor
MVSFTSLLAASPPSRASCRPAAEVESVAVEKRQTIQPGTGYNNGYFYSYWNDGHGGVTYTNGPGG
QFSVNWSNSGNFVGGKGWQPGTKNKVINFSGSYNPNGNSYLSVYGWSRNPLIEYYIVENFGTYNP
```

Fig. 3. Protein sequence data in FASTA format for *xylanase 2* retrieved from *Entrez* using protein names and organisms obtained by NLP analysis

CLUSTAL W (1.82) multiple sequence alignment

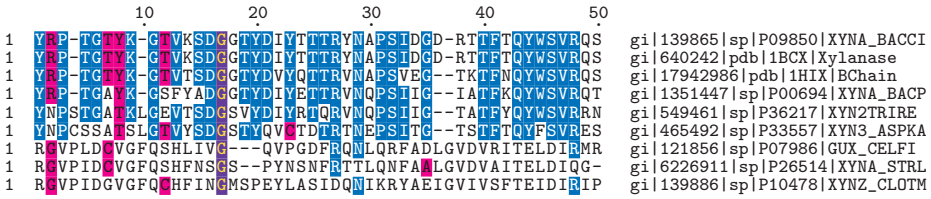

Fig. 4. Alignment of *xylanase* sequences obtained from the *Entrez* database

the protein expression identified as the one being mutated (e.g., *xylanase II*), with NPs containing an expression marked as an organism (e.g., algae or fungi).

Biological Database Integration and Protein Sequence Analysis. As outlined above, information retrieved from documents is used to access various biological databases for the retrieval of protein sequence and structure data, which in turn is used for further processing steps. The end product of our application is a combined data set for protein 3D-structure visualization containing information from both scientific publications and databases. In the following paragraphs, we discuss how data obtained in the text mining step can be processed by in-silico bioinformatics tools and linked to databases.

Protein Sequence Database Access. The second step in the process is the retrieval of protein sequences from a sequence database for each protein/organism combination detected in the text mining subsystem. For this, we access the *Entrez* databases in order to identify protein accessions and retrieve protein sequences in FASTA format [10]. *Entrez* is the integrated, text-based search and retrieval system used at *National Centre for Biotechnology Information (NCBI)* for the major databases, including PubMed Scientific Literature, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, and Taxonomy.[3] The key needed for a successful retrieval is a correct protein/organism pair. Figure 3 shows an example of a retrieved *xylanase* sequence in FASTA format (for programmatic purposes the sequence is obtained in XML format).

Sequence Analysis. The sequence analysis component takes the sequences obtained in the previous step and processes them for similarity. Outlying and du-

[3] The complete list of Entrez databases can be viewed at http://www.ncbi.nlm.nih.gov/Database/index.html

Title	Crystallographic Analyses Of Family 11 Endo–1,4-Xylanase Xyl1
Classification	Hydrolase
Compound	Mol_Id: 1; Molecule: Endo-1,4–Xylanase; Chain: A, B; Ec: 3.2.1.8;
Exp. Method	X-ray Diffraction

```
JRNL TITL 2 ENDO-[BETA]-1,4-XYLANASE XYL1 FROM STREPTOMYCES SP. S38
JRNL REF  ACTA CRYSTALLOGR.,SECT.D V. 57 1813 2001
JRNL REFN ASTM ABCRE6 DK ISSN 0907-4449
...
DBREF  1HIX A    1   190  TREMBL Q59962   Q59962
DBREF  1HIX B    1   190  TREMBL Q59962   Q59962
...
ATOM 1 N   ILE A 4 48.459 19.245 17.075 1.00 24.52 N
ATOM 2 CA  ILE A 4 47.132 19.306 17.680 1.00 50.98 C
ATOM 3 C   ILE A 4 47.116 18.686 19.079 1.00 49.94 C
ATOM 4 O   ILE A 4 48.009 17.936 19.465 1.00 70.83 O
ATOM 5 CB  ILE A 4 46.042 18.612 16.837 1.00 50.51 C
ATOM 6 CG1 ILE A 4 46.419 17.217 16.338 1.00 51.09 C
ATOM 7 CG2 ILE A 4 45.613 19.514 15.687 1.00 54.39 C
ATOM 8 CD1 ILE A 4 46.397 17.045 14.836 1.00 46.72 C
ATOM 9 N   THR A 5 46.077 19.024 19.828 1.00 40.65 N
...
MASTER   321  0  0  2  28  0  0  9 3077  2  0 30
END
```

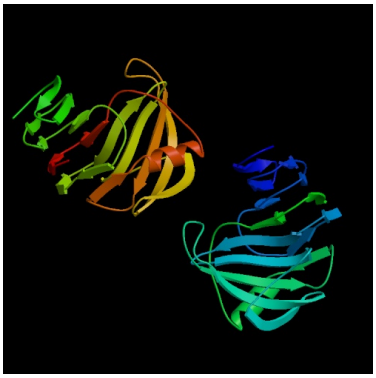

Fig. 5. Protein Data Bank (PDB) record for 1HIX and its 3D-visualization

plicated sequences are identified using multiple sequence alignment (MSA) and statistical scoring with user-specified threshold criteria. Figure 4 shows an excerpt of a MSA with CLUSTAL W [12]. A list of candidate sequences for which protein mutation annotations from the papers may be written to a structure visualization tool input format is generated. Before annotations are written to an input format the sequences are further evaluated for a number of features. Domain complexity is evaluated using CDD *(Conserved Domain Database)* search tools [8] and non-target domains are trimmed. Mutated residues are located on the retrieved sequences and only sequences bearing the declared wild type residues at the specified coordinates with the correct offset between multiple mutations are eligible for subsequent sequence-structure alignment.

Structure Selection. The choice of a protein structure for mapping and visualization of mutations can be generated dynamically or is user-defined. A dynamically selected structure is the top hit obtained when the consensus sequence of all eligible sequences is pairwise aligned using BLAST against the database of sequences of structures contained in the Protein Data Bank. The structure of the selected sequence (top hit) is used as the template to render the mutations and associated annotations from a variety of sequence mutations described in publications. The mapped coordinates of the mutated residues on the structure sequence are identified by pairwise BLAST alignment. More details on the sequence analysis algorithm can be found in [2].

3D-Structure Database Access. We now retrieve the corresponding structure from the Protein Data Bank (PDB). This database is the single worldwide repository for the processing and distribution of 3D biological macromolecular structure

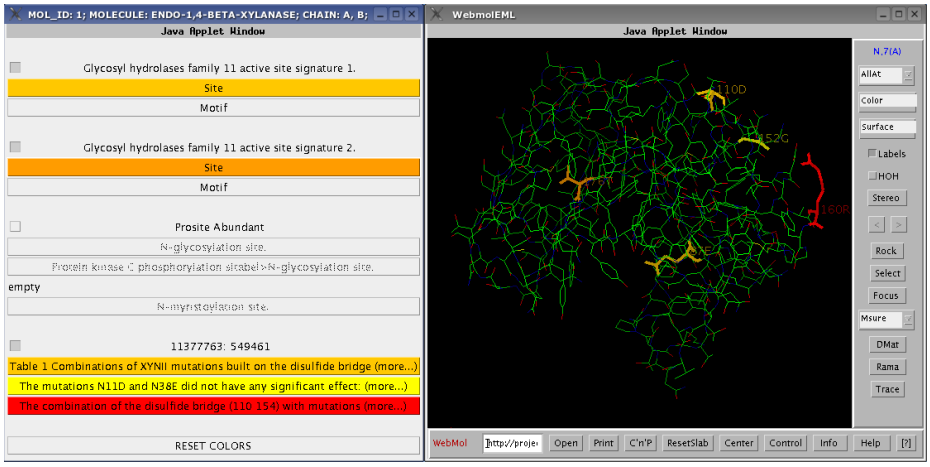

Fig. 6. ProSAT showing a 3D (Webmol) visualization of the endo-1,4-β-protein with mutations extracted through text mining, selected with the interface on the left (sections of the extracted information is displayed on the buttons)

data. The name of the structure, e.g. 1HIX, identified in the selection step is the key term entered in the PDB query engine and facilitates the direct download of the structure file. Figure 5 shows an example of a `pdb`[4] file containing atom coordinates and the corresponding amino acids used by a variety of structure visualization tools for rendering images of protein structures in 3D. Amino acid residue identity and coordinates, columns 4 and 6, facilitate mapping of mutation annotations extracted through text mining.

Application Integration. After sequence analysis has legitimized the transfer of annotations from a particular text to a residue on the structural homolog, ranking and formatting of sentences is necessary. Formatted annotations are produced depending on the input format for a particular visualization tool.

Currently, only the ProSAT template [6] with additional provision for non-database annotations is employed, while other tools could be enhanced for this purpose as well. The annotations, along with the Genbank protein Identifier (GI) and PubMed ID (PMID) for the originating publication, are uploaded to the ProSAT server and rendered alongside the structural homolog through a Webmol interface. Coloured mutated residues are highlighted in structure and described in a corresponding annotation panel, as shown in Figure 6.

3.3 Case Study and Results

Improvement of enzyme features is particularly relevant when the enzyme of interest catalyses an industrially relevant reaction. In our case study we have

[4] Further information on the standard PDB data format can be found at `http://www.rcsb.org/pdb/`.

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

chosen to mine texts describing mutations to *xylanase* enzymes. Such enzymes depolymerise the plant cell wall component *xylan* that is partly responsible for dark colour of unbleached paper. Chlorine based oxidizing chemicals are typically used to bleach paper and result in considerable effluent problems for the pulp and paper industry. Xylanases are now used to remove xylan, which results in less chlorine being required for bleaching. Xylanases have been specifically improved to perform well under industrial conditions (high temperature, alkaline) required by the pulp bleaching process.

For our first system evaluation we selected twenty papers on xylanase mutations. Table 1 shows the results of a preliminary (manual) evaluation of the NLP subsystem. We evaluated (a) whether the system found the correct protein-organism pairs (i.e., it must have identified the protein, the organism, and correctly assigned the protein to its host organism) and (b) how many mutations it found. We are currently preparing more extensive, automated evaluations of the NLP subsystem, the sequence analysis component, and the overall system. However, with respect to the NLP part, the most problematic entities are currently author-invented abbreviations.

3.4 Summary

The case study has addressed a complex biological data integration problem and highlighted the feasibility of integrating literature-derived annotations with in-silico biology. The extent to which the text mining systems combined with sequence analysis tools and existing biological data can provide additional insight to structural biology and protein engineering will be determined from the future employment of the prototype software by expert protein engineers knowledgeable of specific protein families. We consider the use of text mining to drive protein structure visualization as an innovative approach that provides the protein engineer with enhanced access to the knowledge reported by other investigators without the need for time consuming manual literature searches.

4 Future Work

From an application perspective, the inclusion of further information describing enzyme characteristics of wild type proteins for contrasting with the improvements to particular features of these enzymes described in the literature is of

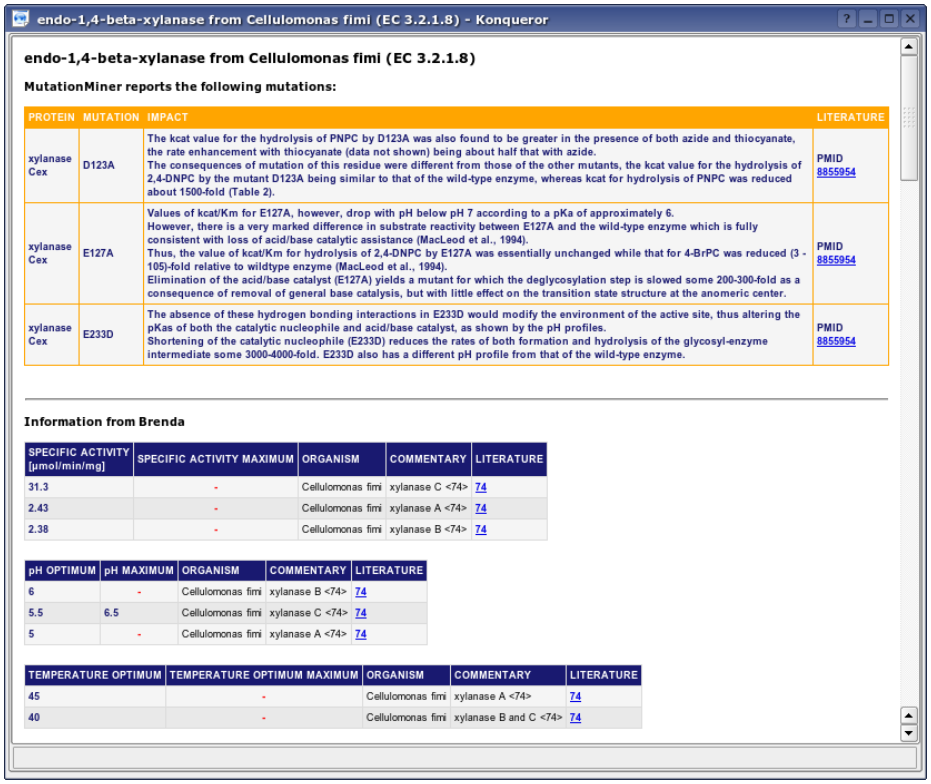

Fig. 7. A simulated screenshot of connecting text mining results of a mutated protein with its corresponding wild-type information from the *Brenda* database

additional value to the protein engineer. Such descriptions in the literature often refer to fold increases without necessarily providing units of measurement. Bringing wild type data together with mutation induced improvements is clearly valuable and information of this kind is of great value in decision making for future investigations. For example, before embarking on major mutational studies to improve an enzyme for a particular property it is important to know if there is a precedent of such an achievement within any protein family. Text mining of mutation literature accompanied with wild type information provided by database searching complements this need and can be achieved by the architecture we describe. Figure 7 shows an example by simulating a connection between our system and the *Brenda* database [11], which we plan to automate in the future.

5 Conclusions

In this paper we present an architecture enabling new biological applications by linking biological databases with text mining results from research papers.

The protein mutation example shows that text mining results of scientific literature can provide enough information to access and link numerous biological databases to build or enhance in-silico bioinformatics applications.

An important insight of our work is that the often imprecise and incomplete results from natural language processing techniques can be automatically filtered through bioinformatics algorithms and supplemented with information from existing databases.

Acknowledgements. Vladislav Ryzhikov implemented and evaluated significant parts of the NLP subsystem. The authors would like to thank Razif R. Gabdoulline and Rebecca Wade for their help and collaboration in adapting their ProSAT system to accept textual annotations.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–310, 1990.
2. Christopher J. O. Baker and René Witte. Enriching Protein Structure Visualizations with Mutation Annotations Obtained by Text Mining Protein Engineering Literature. In *The 3rd Canadian Working Conference on Computational Biology (CCCB'04)*, Markham, Ontario, October 4 2004. Co-located with IBM CASCON.
3. D.P.A. Corney, B.F. Buxton, W.B. Langdon, and D.T. Jones. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, November 22 2004.
4. Francisco M. Couto, Mario J. Silva, and Pedro Coutinho. ProFAL: PROtein Functional Annotation through Literature. In *JISBD*, pages 747–756, 2003.
5. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the ACL*, 2002.
6. R. R. Gabdoulline, R. Hoffmann, F. Leitnern, and R. C. Wade. ProSAT: functional annotation of protein 3D structures. *Bioinformatics*, 19(13):1723–1725, 2003.
7. Takeshi Kawabata, Motonori Ota, and Ken Nishikawa. The protein mutant database. *Nucleaic Acid Research*, 27(1), 1999.
8. A. Marchler-Bauer, A. R. Panchenko, B. A. Shoemaker, P. A. Thiessen, L. Y. Geer, and S. H. Bryant. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, 30(1):281–283, 2002.
9. Hans-Michael Müller, Eimear E. Kenny, and Paul W. Sternberg. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biology*, 2(11):1984–1998, November 2004. www.plosbiology.org.
10. W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. of the National Academy of Sciences of the USA*, 85(8), 1988.
11. I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32, 2004.
12. J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.
13. René Witte. An Integration Architecture for User-Centric Document Creation, Retrieval, and Analysis. In *Proceedings of the VLDB Workshop on Information Integration on the Web (IIWeb)*, pages 141–144, Toronto, Canada, August 30 2004.