Enriching Protein Structure Visualizations with Mutation Annotations Obtained by Text Mining Protein Engineering Literature

Christopher J. O. Baker and René Witte
Department of Computer Science and Software Engineering
Concordia University, Montréal, Canada

Abstract

Protein structure visualization tools render images that allow the user to explore structural features of a protein. Context specific information relating to a particular protein or protein family is not easily integrated and must be uploaded from databases or provided through manual curation of input files. We describe a mixed natural language processing and sequence analysis based approach for the retrieval of mutation specific annotations from full text articles for rendering with protein structures.

Keywords: Text Mining, Protein Structure Annotation, Protein Function, ProSAT, Xylanase

1 INTRODUCTION

Natural language processing (NLP) techniques are progressively being applied to support bioinformatic database curation projects as funding for manual expert curation cannot continue indefinitely [4]. Challenges exist however both in the definition of specific bioinformatic requirements and the capabilities of information retrieval techniques.

As a case study for integrating information retrieval and knowledge extraction with bioinformatic applications we selected the annotation of protein structures with segments of literature detailing the consequences of specific mutations. For protein engineers, understanding the impact of all mutations carried out on a protein family requires a complex mapping of sequence mutants to a common structure. Currently the protein mutation database (PMD) [11] and associated visualization tools 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. Coupled with a larger number of protein structures, more sophisticated alignment algorithms, like Fugue [15] or Muscle [9], and structure annotation tools [10], further improvements could be made to the collation, mapping, and rendering of mutant sequence information. Some structure visualization tools allow the mapping of existing sequences to structures primarily to enable overlay of sequence features stored in databases to structures [10, 14]. Our aim is to employ language technology to improve access to annotations concerning the impacts of mutations and apply these to 3D structures of proteins. To do this we have developed a mixed NLP and sequence analysis approach that combines retrieval and analysis of protein sequences described in selected texts with the extraction of specific sentences from the same texts that describe mutations made to the protein sequences and their impact on protein function. Our architecture facilitates a mapping of mutations and legitimate annotations to a structural homolog in a format readable by structure visualization tools (see Figure 1).

The remainder of this paper is structured as follows: In the next section we discuss the system architecture with its individual components. Section 3 describes a case study using the xylanase protein family. The last section summarizes our findings and outlines future work.

2 SYSTEM ARCHITECTURE

A system capable of extracting experimentally introduced mutations from full-text papers and linking them to protein structure visualizations must be able to integrate document retrieval, NLP-based text analysis, protein sequence database access, protein sequence analysis, and output format generation within a single architecture. For this, we designed a multi-tier information system based on the architecture discussed in [18]. Figure 2 shows the main components, organized by tier.

Users interact with the system using a standard web client (tier 1). A web server (tier 2) receives a query (e.g.,

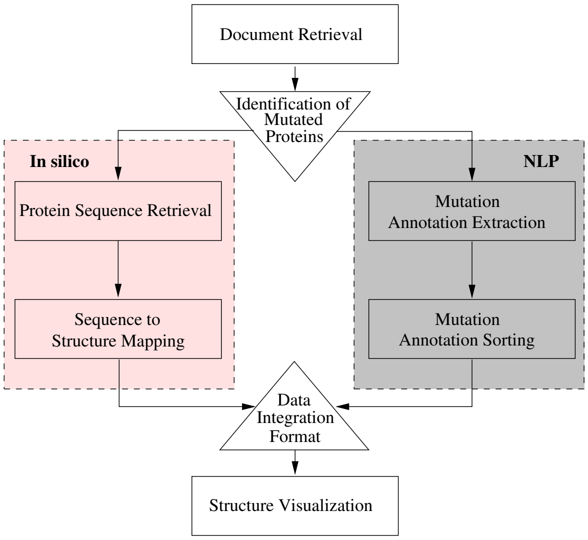

Document Retrieval
Identification of Mutated Proteins
In silico
Protein Sequence Retrieval
Sequence to Structure Mapping
NLP
Mutation Annotation Extraction
Mutation Annotation Sorting
Data Integration Format
Structure Visualization

Figure 1: Data Flow

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 (where 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*[1] in order to identify protein accessions and retrieve protein sequences in FASTA format [13]. 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.

Within this paper, we do not discuss the information retrieval (IR) part of the design. Many of the challenges in document retrieval and conversion, as well as possible solutions, are discussed within the context of the BioRAT system [5].

2.1 NLP 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.

Our NLP component is based on the GATE *(General Architecture for Text Engineering)* framework [6,7], one of the most widely used NLP tools. As it has been designed as a component-based architecture, individual analysis components (called *processing resources*) can be easily added, modified, or removed from the system. GATE is also being used by other biomedical systems, most notably BioRAT [5].

A full text or abstract, once retrieved and converted into a suitable input format, is run through a so-called processing pipeline of NLP 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. This includes lists like person names, dates, locations, companies, measurements, and, most importantly for our task, 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 the non-biomedical information, we rely on lists developed by the CLaC group for the newspaper article domain [2,3], which are based on the ANNIE information extraction system that comes with GATE. Biomedical lists use the same resources as the BioRAT system described in [5]: lists of entries extracted from the MeSH hierarchy and SwissProt, together holding more than five million words in roughly 650,000 entries.

Named Entity Recognition. In the next phase, several finite-state transducers combine individual tokens into more complex named entities (NE), based on regular-expression grammars, which are run over the annotations generated by the previous step. 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.

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 tag*, for example, verb, adjective, or noun. For this, we use the CLaC sentence splitter (an enhanced version of the ANNIE sentence splitter) and the Hepple tagger that comes with the GATE system.

NP 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"*.

[1] *Entrez* is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. See `http://www.ncbi.nlm.nih.gov/Database/index.html`

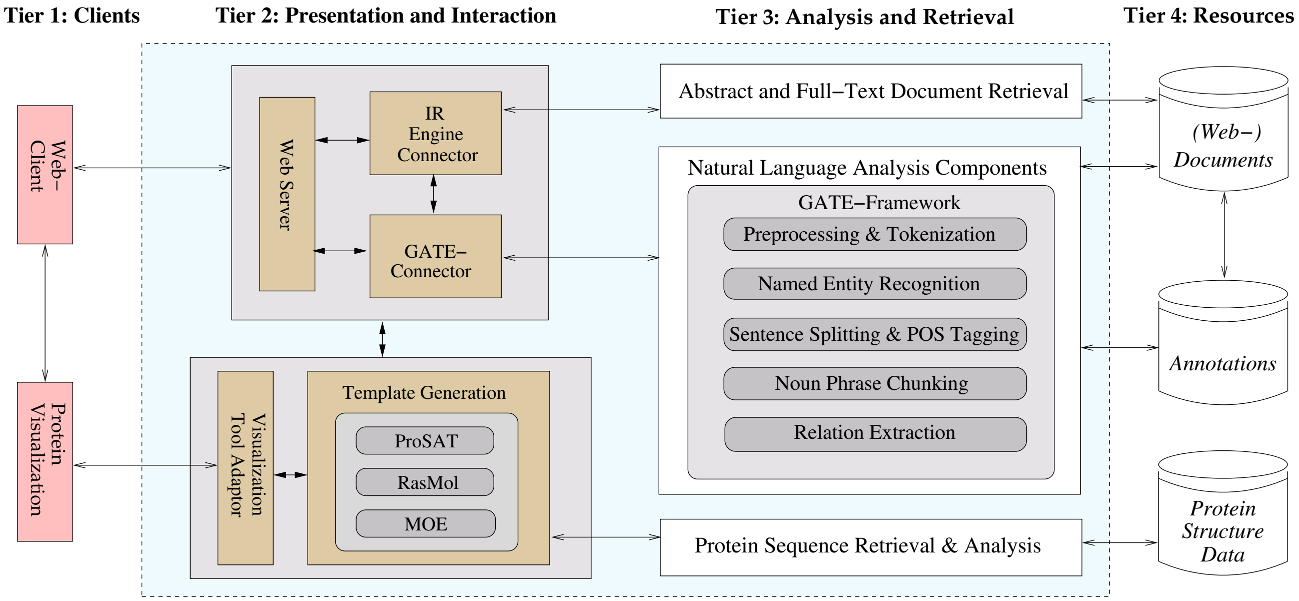

Tier 1: Clients
Tier 2: Presentation and Interaction
Tier 3: Analysis and Retrieval
Tier 4: Resources
Web–Client
Protein Visualization
Web Server
IR Engine Connector
GATE–Connector
Visualization Tool Adaptor
Template Generation
ProSAT
RasMol
MOE
Abstract and Full–Text Document Retrieval
Natural Language Analysis Components
GATE–Framework
Preprocessing & Tokenization
Named Entity Recognition
Sentence Splitting & POS Tagging
Noun Phrase Chunking
Relation Extraction
Protein Sequence Retrieval & Analysis
(Web–) Documents
Annotations
Protein Structure Data

Figure 2: System Architecture

Another grammar stage then joins basic NPs that appear within certain grammatical structures, like prepositions or conjunctions. This *NP chunking* allows us to locate important entities more precisely, for example, in the sentence *"The specific enzyme activity of mutant E210D was 0.8%. . ."* we can identify the whole phrase up to E210D as a single (complex) noun phrase and thus determine that it is really the activity that is 0.8%, not the "E210D" (as a naive approach might infer based on location).

Another 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. This typically results in a number of mis-tagged words, which in turn degrades NP precision.

Relation Detection. The last (and currently most problematic) 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 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).

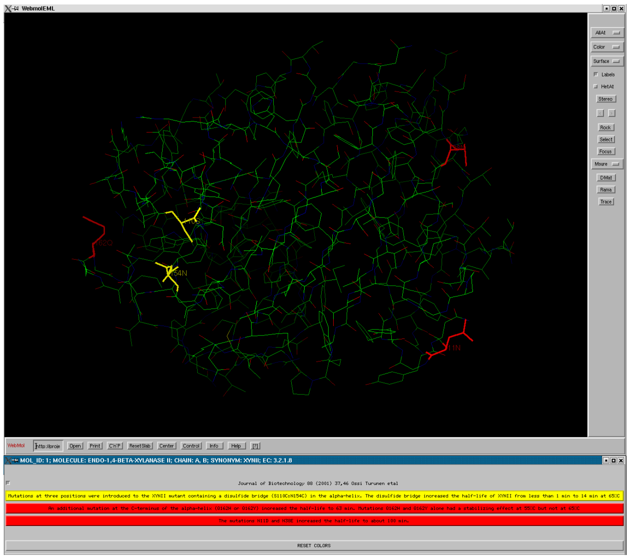
We plan to enhance this step in the future with a more detailed linguistic analysis that first performs a complete syntactical analysis (a full parse) of the sentences and then extracts predicate-argument structures from the parse trees, however, this is still under development.

2.2 Protein Sequence Retrieval and Analysis

The protein sequence retrieval and analysis component attempts to identify protein sequence accessions based on the protein and host organism names obtained in the NLP system. It then retrieves formatted protein sequences and analyzes them for similarity. Outlying sequences are removed, producing a list of sequences for which protein mutation annotations will be retrieved.

```
<menue>
  <status=on>
  <label>
     Journal of Biotechnology 88 (2001) 37,46 Ossi Turunen etal
   </label>
  <item>
    <range>110:A 110:A, 154:A 154:A</range>
    <color=yellow> <status=on>
    <label>Mutations at three positions were introduced to the
       XYNII mutant containing a disulfide bridge (S110C:N154C)
       in the alpha−helix. The disulfide bridge increased the
       half−life of XYNII from less than 1 min to 14 min at 65 C
     </label>
   </item>
  <item>
    <range>162:A 162:A</range>
    <color=red> <status=on>
    <label>An additional mutation at the C−terminus of the
       alpha−helix (Q162H or Q162Y) increased the half−life
       to 63 min. Mutations Q162H and Q162Y alone had a
       stabilizing effect at 55 C but not at 65 C
     </label>
   </item>
  <item>
    <range>11:A 11:A, 38:A 38:A </range>
    <color=red> <status=on>
    <label>The mutations N11D and N38E increased the
       half−life to about 100 min.
     </label>
   </item>
 </menue>
```

Figure 3: Template with extracted information used for ProSAT visualization


WebMol
http://proje
Open
Print
C'n'P
ResetSlab
Center
Control
Info
Help
[?]
MOL_ID: 1; MOLECULE: ENDO-1,4-BETA-XYLANASE II; CHAIN: A, B; SYNONYM: XYNII; EC: 3.2.1.8
RESET COLORS

Journal of Biotechnology 88 (2001) 37,46 Ossi Turunen etal

Mutations at three positions were introduced to the XYNII mutant containing a disulfide bridge (S110C:N154C) in the alpha-helix. The disulfide bridge increased the half-life of XYNII from less than 1 min to 14 min at 65C

An additional mutation at the C-terminus of the alpha-helix (Q162H or Q162Y) increased the half-life to 63 min. Mutations Q162H and Q162Y alone had a stabilizing effect at 55C but not at 65C

The mutations N11D and N38E increased the half-life to about 100 min.

Figure 4: ProSAT showing annotations extracted through text mining (enlarge electronic version for details)

To achieve this, a protein name and originating organism obtained by NLP analysis is used as input to *Entrez* for retrieval of protein sequence accession and the sequence. The FASTA formatted sequence of the top hit is obtained and the identity of the amino acid at the position described as mutated in the publication is checked. Further evaluation of domain complexity on the sequence using CDD *(Conserved Domain Database)* search tools [12] is carried out. Where the retrieved sequences contain multiple domains the non-target protein sequence is removed while maintaining the original residue numbering. The degree of sequence identity between all retrieved sequences is determined by producing multiple sequence alignments (MSA) with CLUSTAL W [16], which are then statistically scored using *alistat* [8] to determine the overall similarity of the sequences. The most distant sequence in the alignment is calculated by finding the maximum pairwise identity (best relative) for all sequences, then finding the minimum of these numbers and hence, the most outlying sequence. Iteratively, the most outlying sequence is removed and the alignment remade and rescored with *alistat* until the most outlying sequence is within a specific threshold. A consensus sequence is generated and a BLAST *(Basic Local Alignment Search Tool)* [1] search is used to identify the closest structural homolog. Each of the sequences in the MSA is then aligned, pairwise, with the sequence of the closest structural homolog using BLAST. Residue alignment is recorded for identification of the equivalent residue in the structural homolog to receive annotations described in a text. A local sequence homology is calculated for the region covering the mutated residue and five amino acids up and downstream to evaluate the legitimacy of the annotation transfer. A threshold of conservation is applied to infer legitimacy.

2.3 Output Template Generation

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

Here, only the ProSAT template [10] with additional provision for non-database annotations is employed (personal communication R. Gabdoulline), while other tools could be enhanced for this purpose as well.

Annotations are uploaded to the ProSAT server and rendered on the structural homolog through a Webmol interface. Coloured mutated residues are highlighted in structure and described in a corresponding annotation panel.

3 CASE STUDY

To demonstrate the feasibility of our approach for annotation of a protein structure with useful mutation annotations we selected xylanases as a protein family of interest to us. Xylanase (EC 3.2.1.8) is a family of enzymes that can depolymerise the hemicellulose and plant cell wall component xylan to simple sugars. Many industrial applications exist for this fibre modifying enzyme and numerous publications describe mutations made to xylanases in order to improve their properties.

PMID	*Entrez* Protein Accession	Found	Accession	Protein Name	Organism	Fam.	#M	Abst.	Trim
8855954	gi|121856|sp|P07986|GUX_CELFI	Yes	None	CEX xylanase	*Cellulomonas fimi*	10	3	Yes	Yes
1359880	gi|1351447|sp|P00694|XYNA_BACPU	Yes	None	Xylanase	*Bacillus pumilus*	11	3	Yes	No
8019418	gi|139865|sp|P09850|XYNA_BACCI	Yes	None	Xylanase	*Bacillus circulans*	11	2	Yes	No
10220321	gi|139865|sp|P09850|XYNA_BACCI	Yes	1bvv, 2bv	Xylanase	*Bacillus circulans*	11	1	Yes	No
10860737	gi|139865|sp|P09850|XYNA_BACCI	Yes	1C5H, 1C5I	Xylanase	*Bacillus circulans*	11	1	Yes	No
11601976	gi|139886|sp|P10478|XYNZ_CLOTM	Yes	None	Xylanase Z	*Clostridium thermocellum*	10	1	Yes	Yes
10752608	gi|17942986|pdb|1HIX|B	No	None	Xyl1	*Streptomyces Sp. S38*	11	5	No	No
9930661	gi|465492|sp|P33557|XYN3_ASPKA	Yes	None	Xylanase C	*Aspergillus kawachii*	11	1	Yes	No
8376336	gi|533366|gb|M97882.1|TEOENDXYLA	No	M97882	Xylanase	*T. saccharolyticum*	11	3	Yes	No
11377763	gi|549461|sp|P36217|XYN2_TRIRE	Yes	None	Xylanase II	*Trichoderma reesei*	11	3	Yes	No
11917150	gi|549461|sp|P36217|XYN2_TRIRE	Yes	None	Xylanase II	*Trichoderma reesei*	11	11	Yes	No
15129722	gi|549461|sp|P36217|XYN2_TRIRE	Yes	None	Xylanase II	*Trichoderma reesei*	11	2	Yes	No
15260499	gi|549461|sp|P36217|XYN2_TRIRE	Yes	P36217, P362	Xylanase II	*Trichoderma reesei*	11	3	Yes	No
15278768	gi|549461|sp|P36217|XYN2_TRIRE	Yes	None	Xylanase II	*Trichoderma reesei*	11	3	Yes	No
7764794	gi|6226911|sp|P26514|XYNA_STRLI	Yes	None	Xylanase A	*Streptomyces lividans*	10	3	Yes	Yes
9201919	gi|6226911|sp|P26514|XYNA_STRLI	Yes	None	Xylanase A	*Streptomyces lividans*	10	2	Yes	Yes
9681873	gi|6226911|sp|P26514|XYNA_STRLI	Yes	None	Xylanase A	*Streptomyces lividans*	10	1	Yes	Yes
10235626	gi|6226911|sp|P26514|XYNA_STRLI	Yes	None	Xylanase A	*Streptomyces lividans*	10	4	Yes	Yes
9731776	gi|640242|pdb|1BCX|	No	None	β-1,4-glycosidase	*Cex*	11	2	Yes	No

Table 1: *Entrez* protein accessions for xylanases using protein names and taxonomic origins extracted from full text articles

In the current study we retrieved 20 texts describing mutations to xylanase proteins using keyword searches. We wished to retrieve the protein sequences corresponding to these papers. In the majority of papers the database accession identifiers for the xylanase proteins were absent. The NLP subsystem was able to identify protein names and taxonomic origins, which were then used to search the protein sequence database *Entrez* for the protein accession identifiers. Table 1 summarizes our case study's main results, including the PubMed IDs (PMID) for the abstracts of each article investigated and the *Entrez* protein accessions retrieved. Additionally, column "#M" in the table shows the number of mutations described in each paper.

Due to missing entries in the Gazetteer lists our system failed to mark up *Thermoanaerobacterium* or *pimulus* as genera and species, respectively, preventing the automated retrieval of the protein identifier and sequences. Multiple papers referred to the same proteins reducing the overall number of sequences retrieved. Three protein sequences also included non-xylanase domains, which we trimmed out by using CDD to find the coordinates of the xylanase domains. These sequences are highlighted in column "Trim." To review the overall similarity of the sequences a multiple sequence alignment of the retrieved sequences was carried out (see Figure 5 in the appendix). Here we can see the degree of sequence divergence between xylanases of different subfamilies, both family 11 and family 10 xylanases were represented (compare with column "Fam."). The final MSA contained only family 11 xylanases and the *alistat* statistical scoring of these sequences identified them as having greater than 70% similarity to each other. This was the minimum similarity threshold for NLP extraction and mapping of mutation specific annotations to proceed. Mapping of sequences to the structural homolog was achieved by pairwise alignment with structure-sequence 1REF, representing the xylanase II from *Trichoderma reesei*. Texts describing mutations on these sequences were analyzed and the NLP annotations extracted and sorted with a relevance score.

The structure-sequence residues equivalent to those in mutated sequences were written along with the highest scoring text annotations into the ProSAT structural visualization template as shown in Figure 3. Together the 20 papers evaluated in this case study describe 54 amino acid residues that had been mutated, 14 on family 10 and 40 on family 11 xylanases. Figure 4 shows a screenshot of ProSAT rendering 1REF (family 11 structure-sequence) with five mutated residues highlighted and additional annotations derived from [17], which describe the impact of mutations on xylanase thermostability through the introduction of new disulphide bridges.

4 CONCLUSIONS AND FUTURE WORK

In this paper we present a system architecture capable of automatically extracting mutation information from protein engineering literature for enriching the information provided by visualization tools. Our system relies to a large extent on components and resources that are available already, but have never before been integrated within a single architecture for the purpose of protein structure visualization.

While our system is still in its early stages of development and more rigorous evaluations are needed, we nevertheless believe it to be important to show how the vast amount of information available online today can be exploited in an automatic fashion for the bioengineer.

One of our main contributions, therefore, is to highlight the challenges involved in integrating literature-derived annotations with in-silico biology and to consider the extent to which the integration of text mining systems with tools and databases already available can provide additional insight to structural biology and protein engineering. While NLP-based approaches cannot retrieve 100% of all relevant protein accessions and their annotations, even a recall rate of 25%–50% would be a vast improvement over the currently available rate of ca. 5% accessible through manually curated databases. Protein engineers get immediate access to current and historical research results, without a need for time-consuming, manual literature search.

Our next step will be the development of a larger corpus of test documents in order to obtain precision and recall measures for our system and to aid in detecting shortcomings for further developments. We also plan on collaborating with visualization tool developers to allow for displaying more complex annotations that can be tied directly with our text analysis component, thus allowing for a more structured and flexible view than it is possible through simple sentence extraction.

In the future, a system as described here could also be integrated with full-text databases like PubMed Central (PMC), enabling automatic extraction of relevant information from newly submitted documents and their delivery, in the form of a web service, to various clients, including structural visualization tools.

Acknowledgments. This work was financed in part through the Génome Québec funded project *Ontologies, the semantic web and intelligent systems for genomics* (V. Haarslev and G. Butler). The authors would like to thank Razif R. Gabdoulline for his timely help in understanding the ProSAT template format and providing the screenshot that appears in this paper. Natural language processing capabilities are delivered in part by the CLaC group (Computational Linguistics at Concordia), which is headed by S. Bergler and L. Kosseim.

References

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Meyers, and David J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–310, 1990.

[2] Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz. Using knowledge-poor coreference resolution for text summarization. In *Proceedings of the HLT-NAACL Workshop on Text Summarization DUC 2003*, pages 43–50, Edmonton, Canada, 2003. NIST.

[3] Sabine Bergler, René Witte, Zhuoyan Li, Michelle Khalifé, Yunyu Chen, Monia Doandes, and Alina Andreevskaia. Multi-ERSS and ERSS 2004. In *Workshop on Text Summarization*, Document Understanding Conference (DUC), Boston Park Plaza Hotel and Towers, Boston, USA, May 6–7 2004. NIST. `http://duc.nist.gov/`.

[4] EMBOSS seeks funding as UK Human Genome Mapping Project Resource Centre closes down. `http://bioinformatics.org/forums/forum.php?forum_id=2663`, 2004.

[5] David P. A. Corney, Bernard F. Buxton, William B. Langdon, and David T. Jones. BioRAT: extracting biological information from full-length papers. *Bioinformatics Advance Access*, July 1st 2004. PMID 15231534.

[6] H. Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254, 2002. `http://gate.ac.uk`.

[7] 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 Association for Computational Linguistics*, 2002.

[8] S. Eddy and E. Birney. *HMMER User's Guide: Biological sequence analysis using profile hidden Markov models*. Washington University, version 2.1.1 edition, 1998. `http://hmmer.wustl.edu/`.

[9] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, March 19 2004.

[10] Razif R. Gabdoulline, René Hoffmann, Florian Leitnern, and Rebecca C. Wade. ProSAT: functional annotation of protein 3D structures. *Bioinformatics*, 19(13):1723–1725, 2003.

[11] Takeshi Kawabata, Motonori Ota, and Ken Nishikawa. The protein mutant database. *Nucleaic Acid Research*, 27(1), 1999.

[12] 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.

[13] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448, April 1988.

[14] Andrea Schafferhans, Joachim E. W. Meyer, and Seán I. O'Donoghue. The PSSH database of alignments between protein sequences and tertiary structures. *Nucleic Acids Research*, 31(1):494–498, January 1st 2003.

[15] J. Shi, T. L. Blundell, and K. Mizuguchi. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*, 310(1):243–257, 2001.

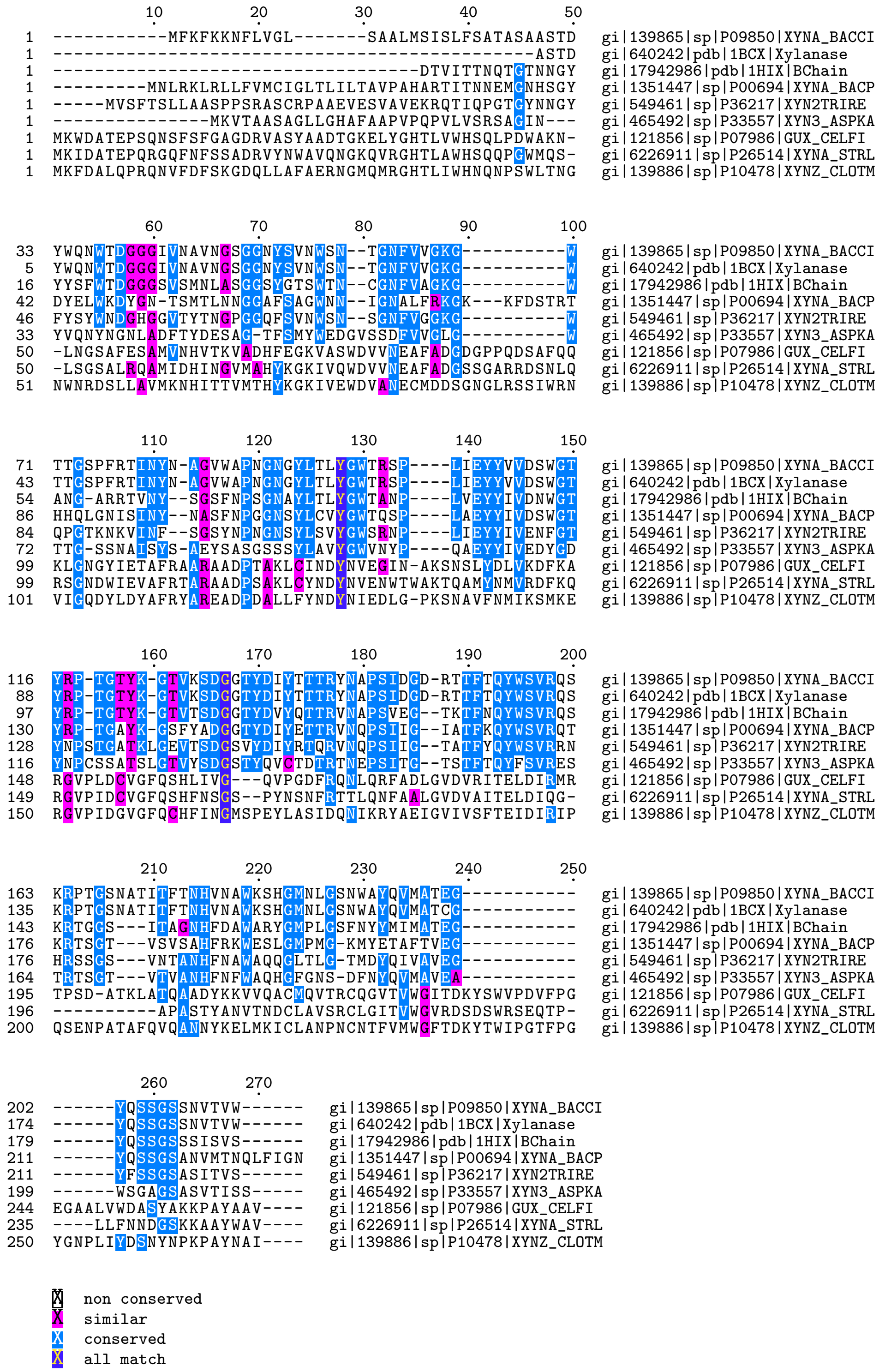
[16] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

[17] O. Turunen, K. Etuaho, F. Fenel, J. Vehmaanpera, X. Wu, J. Rouvinen, and M. Leisola. A combination of weakly stabilizing mutations with a disulfide bridge in the alpha-helix region of trichoderma reesei endo-1,4-beta-xylanase ii increases the thermal stability through synergism. *Journal of Biotechnology*, 88(1):37–46, June 1st 2001.

[18] René Witte. An Integration Architecture for User-Centric Document Creation, Retrieval, and Analysis. In Hasan Davulcu and Nick Kushmerick, editors, *Proceedings of the VLDB Workshop on Information Integration on the Web (IIWeb-2004)*, pages 141–144, Toronto, Canada, August 30 2004.

Appendix

CLUSTAL W (1.82) multiple sequence alignment



```
                  10        20        30        40        50
                   .         .         .         .         .
  1  -----------MFKFKKNFLVGL-------SAALMSISLFSATASAASTD   gi|139865|sp|P09850|XYNA_BACCI
  1  ----------------------------------------------ASTD   gi|640242|pdb|1BCX|Xylanase
  1  -----------------------------------DTVITTNQTGTNNGY   gi|17942986|pdb|1HIX|BChain
  1  ---------MNLRKLRLLFVMCIGLTLILTAVPAHARTITNNEMGNHSGY   gi|1351447|sp|P00694|XYNA_BACP
  1  -----MVSFTSLLAASPPSRASCRPAAEVESVAVEKRQTIQPGTGYNNGY   gi|549461|sp|P36217|XYN2TRIRE
  1  ---------------MKVTAASAGLLGHAFAAPVPQPVLVSRSAGIN---   gi|465492|sp|P33557|XYN3_ASPKA
  1  MKWDATEPSQNSFSFGAGDRVASYAADTGKELYGHTLVWHSQLPDWAKN-   gi|121856|sp|P07986|GUX_CELFI
  1  MKIDATEPQRGQFNFSSADRVYNWAVQNGKQVRGHTLAWHSQQPGWMQS-   gi|6226911|sp|P26514|XYNA_STRL
  1  MKFDALQPRQNVFDFSKGDQLLAFAERNGMQMRGHTLIWHNQNPSWLTNG   gi|139886|sp|P10478|XYNZ_CLOTM


                  60        70        80        90       100
                   .         .         .         .         .
 33  YWQNWTDGGGIVNAVNGSGGNYSVNWSN--TGNFVVGKG----------W   gi|139865|sp|P09850|XYNA_BACCI
  5  YWQNWTDGGGIVNAVNGSGGNYSVNWSN--TGNFVVGKG----------W   gi|640242|pdb|1BCX|Xylanase
 16  YYSFWTDGGGSVSMNLASGGSYGTSWTN--CGNFVAGKG----------W   gi|17942986|pdb|1HIX|BChain
 42  DYELWKDYGN-TSMTLNNGGAFSAGWNN--IGNALFRKGK---KFDSTRT   gi|1351447|sp|P00694|XYNA_BACP
 46  FYSYWNDGHGGVTYTNGPGGQFSVNWSN--SGNFVGGKG----------W   gi|549461|sp|P36217|XYN2TRIRE
 33  YVQNYNGNLADFTYDESAG-TFSMYWEDGVSSDFVVGLG----------W   gi|465492|sp|P33557|XYN3_ASPKA
 50  -LNGSAFESAMVNHVTKVADHFEGKVASWDVVNEAFADGDGPPQDSAFQQ   gi|121856|sp|P07986|GUX_CELFI
 50  -LSGSALRQAMIDHINGVMAHYKGKIVQWDVVNEAFADGSSGARRDSNLQ   gi|6226911|sp|P26514|XYNA_STRL
 51  NWNRDSLLAVMKNHITTVMTHYKGKIVEWDVANECMDDSGNGLRSSIWRN   gi|139886|sp|P10478|XYNZ_CLOTM


                 110       120       130       140       150
                   .         .         .         .         .
 71  TTGSPFRTINYN-AGVWAPNGNGYLTLYGWTRSP----LIEYYVVDSWGT   gi|139865|sp|P09850|XYNA_BACCI
 43  TTGSPFRTINYN-AGVWAPNGNGYLTLYGWTRSP----LIEYYVVDSWGT   gi|640242|pdb|1BCX|Xylanase
 54  ANG-ARRTVNY--SGSFNPSGNAYLTLYGWTANP----LVEYYIVDNWGT   gi|17942986|pdb|1HIX|BChain
 86  HHQLGNISINY--NASFNPGGNSYLCVYGWTQSP----LAEYYIVDSWGT   gi|1351447|sp|P00694|XYNA_BACP
 84  QPGTKNKVINF--SGSYNPNGNSYLSVYGWSRNP----LIEYYIVENFGT   gi|549461|sp|P36217|XYN2TRIRE
 72  TTG-SSNAISYS-AEYSASGSSSYLAVYGWVNYP----QAEYYIVEDYGD   gi|465492|sp|P33557|XYN3_ASPKA
 99  KLGNGYIETAFRAARAADPTAKLCINDYNVEGIN-AKSNSLYDLVKDFKA   gi|121856|sp|P07986|GUX_CELFI
 99  RSGNDWIEVAFRTARAADPSAKLCYNDYNVENWTWAKTQAMYNMVRDFKQ   gi|6226911|sp|P26514|XYNA_STRL
101  VIGQDYLDYAFRYAREADPDALLFYNDYNIEDLG-PKSNAVFNMIKSMKE   gi|139886|sp|P10478|XYNZ_CLOTM


                 160       170       180       190       200
                   .         .         .         .         .
116  YRP-TGTYK-GTVKSDGGTYDIYTTTRYNAPSIDGD-RTTFTQYWSVRQS   gi|139865|sp|P09850|XYNA_BACCI
 88  YRP-TGTYK-GTVKSDGGTYDIYTTTRYNAPSIDGD-RTTFTQYWSVRQS   gi|640242|pdb|1BCX|Xylanase
 97  YRP-TGTYK-GTVTSDGGTYDVYQTTRVNAPSVEG--TKTFNQYWSVRQS   gi|17942986|pdb|1HIX|BChain
130  YRP-TGAYK-GSFYADGGTYDIYETTRVNQPSIIG--IATFKQYWSVRQT   gi|1351447|sp|P00694|XYNA_BACP
128  YNPSTGATKLGEVTSDGSVYDIYRTQRVNQPSIIG--TATFYQYWSVRRN   gi|549461|sp|P36217|XYN2TRIRE
116  YNPCSSATSLGTVYSDGSTYQVCTDTRTNEPSITG--TSTFTQYFSVRES   gi|465492|sp|P33557|XYN3_ASPKA
148  RGVPLDCVGFQSHLIVG---QVPGDFRQNLQRFADLGVDVRITELDIRMR   gi|121856|sp|P07986|GUX_CELFI
149  RGVPIDCVGFQSHFNSGS--PYNSNFRTTLQNFAALGVDVAITELDIQG-   gi|6226911|sp|P26514|XYNA_STRL
150  RGVPIDGVGFQCHFINGMSPEYLASIDQNIKRYAEIGVIVSFTEIDIRIP   gi|139886|sp|P10478|XYNZ_CLOTM


                 210       220       230       240       250
                   .         .         .         .         .
163  KRPTGSNATITFTNHVNAWKSHGMNLGSNWAYQVMATEG-----------   gi|139865|sp|P09850|XYNA_BACCI
135  KRPTGSNATITFTNHVNAWKSHGMNLGSNWAYQVMATCG-----------   gi|640242|pdb|1BCX|Xylanase
143  KRTGGS---ITAGNHFDAWARYGMPLGSFNYYMIMATEG-----------   gi|17942986|pdb|1HIX|BChain
176  KRTSGT---VSVSAHFRKWESLGMPMG-KMYETAFTVEG-----------   gi|1351447|sp|P00694|XYNA_BACP
176  HRSSGS---VNTANHFNAWAQQGLTLG-TMDYQIVAVEG-----------   gi|549461|sp|P36217|XYN2TRIRE
164  TRTSGT---VTVANHFNFWAQHGFGNS-DFNYQVMAVEA-----------   gi|465492|sp|P33557|XYN3_ASPKA
195  TPSD-ATKLATQAADYKKVVQACMQVTRCQGVTVWGITDKYSWVPDVFPG   gi|121856|sp|P07986|GUX_CELFI
196  ----------APASTYANVTNDCLAVSRCLGITVWGVRDSDSWRSEQTP-   gi|6226911|sp|P26514|XYNA_STRL
200  QSENPATAFQVQANNYKELMKICLANPNCNTFVMWGFTDKYTWIPGTFPG   gi|139886|sp|P10478|XYNZ_CLOTM


                 260       270
                   .         .
202  ------YQSSGSSNVTVW------   gi|139865|sp|P09850|XYNA_BACCI
174  ------YQSSGSSNVTVW------   gi|640242|pdb|1BCX|Xylanase
179  ------YQSSGSSSISVS------   gi|17942986|pdb|1HIX|BChain
211  ------YQSSGSANVMTNQLFIGN   gi|1351447|sp|P00694|XYNA_BACP
211  ------YFSSGSASITVS------   gi|549461|sp|P36217|XYN2TRIRE
199  ------WSGAGSASVTISS-----   gi|465492|sp|P33557|XYN3_ASPKA
244  EGAALVWDASYAKKPAYAAV----   gi|121856|sp|P07986|GUX_CELFI
235  ----LLFNNDGSKKAAYWAV----   gi|6226911|sp|P26514|XYNA_STRL
250  YGNPLIYDSNYNPKPAYNAI----   gi|139886|sp|P10478|XYNZ_CLOTM


  X  non conserved
  X  similar
  X  conserved
  X  all match
```

Figure 5: Alignment of xylanase sequences retrieved from *Entrez* using protein names and organisms obtained by NLP analysis