THE DEVELOPMENT OF A TOOL FOR MAPPING PROTEIN MUTATIONS TO SEQUENCE STRUCTURES

ASHWIN BHAT GURPUR

A THESIS
IN
THE DEPARTMENT
OF
COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

AUGUST 2005

Your file Votre référence
ISBN: 0-494-10286-1
Our file Notre référence
ISBN: 0-494-10286-1

Abstract

The Development Of A Tool For Mapping Protein Mutations To Sequence Structures

Ashwin Bhat Gurpur

Related work has been done in the NLP area to extract protein mutation information directly from PubMed papers and storing it in an XML file. This thesis describes a tool that processes this NLP output for the purpose of visualizing the mutations. The tool uses the NLP output file as input and extracts the details of the protein being discussed, along with the mutation information and these details are used to extract the sequence information from the NCBI protein database. Next, for each protein, it extracts the conserved domain information from the NCBI conserved domain database. Each extracted sequence is split into its respective conserved domains and these are placed sequentially. ClustalW and Alistat are used to remove sequences that fall below a particular threshold. For the remaining sequences, a consensus sequence is generated and a structure that best matches it, is selected. Mutations corresponding to the remaining sequences are mapped on to the structure and a reliability score is calculated. All this information is written on to a visualization file. This is the final output of this tool. This file can be uploaded to the PROSAT protein visualization tool and the mutations can be visualized. The results obtained when the tool was tested on three protein families – xylanases, dehalogenases and biphenyl dioxygenase are presented.

Acknowledgments

I thank Dr. Greg Butler, Dr. Volker Haarslev and Dr. Christopher Baker for their invaluable guidance and support.

Contents

List Of Figures

List of Tables

Chapter 1 Introduction

1.1 The Problem

A protein[Onl05a] is composed of one or more chains of amino acids (organic macromolecules having carbon, hydrogen, oxygen, nitrogen and sulphur). They include enzymes, hormones and antibodies, which are necessary for the proper functioning of an organism. Protein engineers perform mutations on proteins and enzymes to alter their properties[Doi05]. Mutation means a change in the amino acid sequence of a protein. A mutation could be an addition of an amino acid, a deletion of an amino acid, or a substitution where one amino acid gets replaced by another one. Mutations can be natural or induced by man.

Some mutations lead to better properties in industrial enzymes like more heat tolerance while other mutations could make the enzyme completely incapable of performing some functions. Thus, knowledge of the effect of mutations on proteins and enzymes is very helpful for making better use of them.

The Protein Mutant Database (PMD)[KNO99] has the natural and artificial mutants of all proteins except the globin and immunoglobulin families. For each paper that discusses mutations, PMD has the complete list of mutations on the proteins and the effects of such mutations on the protein.

The problem is that PMD is not up-to-date with the latest information[BW04]. In 1999, PMD authors reported a 3 year backlog of unprocessed papers. This would severely restrict protein research that depends on the availability of such latest information. The tool we develop in this research is focused on this problem and also to provide a component for visualizing all mutations of a protein family on a common structure.

Another limitation of PMD is that it has no facility for visualizing all mutations of a particular closely related group of protein sequences by automatically selecting the right structure and mapping the mutations on it.

Summary of the problem: We have a set of PubMed papers that discuss mutations on a particular family F of proteins (each paper can discuss the mutations of one or more proteins). Each protein is represented by a sequence of residues X1X2X3...Xn where X1,X2.. are amino acids. Let these papers be represented as A1..An and the mutations they discuss be represented as XpY, which is a mutation by substitution where residue X at position p is substituted by residue Y. The papers count the positions of the mutated residues starting from the first residue on the sequence, the first residue being position one.

Our aim is to visualize all these mutations of several sequences (of the same family) on a single structure and the problem here is to select a protein P with sequence S and a 3D structure and then for each of the mutations XpY, being discussed in the papers, find the corresponding position q on this sequence. This problem is an approximate mapping problem because each sequence S1..Sn corresponding to proteins P1..Pn must be mapped to S and also each of the mutations on these sequences needs to be located on S.

This problem is approximate in nature because protein sequences undergo evolutionary changes within a family F and substitutions and gaps in sequence alignments are a common issue in bioinformatics. Also, the PubMed papers A1..An do not provide the sequences of the proteins P1..Pn but only an identifier or reference to the proteins. The papers may provide the protein name and organism name from which the protein was obtained, and a description of protein, but this information may not be sufficient to obtain the right sequence from a Sequence archive like NCBI protein database. Another problem is that the

positions of the mutations discussed in the papers do not agree with the positions of the residues on the sequence obtained from the sequence archives. Hence for each mutation XpY of each protein P, which is discussed in the papers, the real position p of the mutations needs to be found on the sequence of P, that is obtained from the sequence archives, before the mapping is done on the selected sequence S.

1.2 Requirements to solve the problem

Enormous human effort is required to keep the PMD database upto date and the PMD authors report this difficulty in their publications. Figure 1.1 shows that the last update on their database was in 2002.

File name: "pmd05Jul25"
Last update :

Current total entries are:

Year	Entry	Mutant
1910	1	2
1979	15	350
1980	15	46
1981	23	68
1982	26	61
1983	64	235
1984	139	686
1985	152	809
1986	208	962
1987	318	2057
1988	480	3752
1989	538	4689
1990	816	8745
1991	1456	12345
1992	1505	13048
1993	1865	14682
1994	2527	19698
1995	2702	7151
1996	2844	4458
1997	3396	4695
1998	3324	19109
1999	3441	22868
2000	4125	29999
2001	5213	15306
2002	4612	169
TOTAL	39805	185990

Counting date : 05.07.27

Figure 1.1 Statistics of updates on the PMD database

It is not clear if the papers on mutations published after this date have been processed by the PMD authors. This means that there is a requirement for a software system that can scan the papers with least human effort and provide the users with the mutation information in those papers, without actually having to read them. This system also needs to provide the protein name, which is being discussed in the papers along with the PubMed identifier of the papers. It also needs to provide the context information provided in the papers so that the user has an idea of the context in which the mutation is being discussed.

Protein engineers may often require to visualize the mutations, which are being discussed in the papers, but PMD does allow a visualization only for those protein sequences which have greater than 50% sequence identity with a PDB structure. Also, the visualization is only for mutations of only one single protein at a time. This means that there is no facility for a user to view mutations of a set of closely related sequences on a common structure or a structure that closely matches these set of sequences.

Hence there is a need for a software system which can take as input a set of sequences and can find a closely matching structure and can map the mutations of each of these sequences on to the structure. This enables the user to visualize all of these mutations on a common structure.

A tool is described here which has these capabilities to perform the scanning of papers, extract their mutations, retain sequences that are closely related, find the best matching structure and finally to map their mutations on to the structure and also to visualize them on a visualization tool like ProSAT.

1.3 My Work

Related work on Natural Language Processing (NLP) techniques has been successful in extracting mutation information from published PubMed papers[Wit04]. The aim of this work has been to develop a tool which uses the results of this NLP analysis and after several processing steps, outputs a visualization file, which is viewed in a special interface created for this tool in ProSAT[GHLW04]. This enables a biologist to view all mutations that pertain to a particular family on a related structure, with much less human effort in locating the structure or the mutations. A brief overview of these processing steps is illustrated in the activity diagram of Figure 1.2. The following chapters explain these steps in detail. The main requirement is to find a consensus for the sequences of interest, find the most appropriate structure that aligns well with this consensus, map the mutations on this structure and to assess the quality of the mapping.

The NLP analysis step in our work was done using the General Architecture for Text Engineering (GATE) framework[Cun02], as described in Chapter 2 and provides the initial input XML file, required by the tool for further processing. The input includes the organism name, the protein name and a list of mutations on that particular protein in predefined tags and the context in which the mutations are discussed in the paper, in which they were found. Each paper may mention more than one proteins, however it will have only one PubMed identifier (PMID) in the PubMed database[Med05]. So, the combination of PubMed identifier and the GenBank identifier (gi) of a protein can be used to identify a set of mutations that pertain to a particular protein.

The input to the tool is the output of the NLP analysis. The next step is to concatenate the organism name and the protein name in a particular way as described in chapter 2, and

to obtain the protein sequence, that is being discussed in that portion of the NLP output file. This sequence is stored in a file, called the sequence file. Like this, all the sequences discussed in the NLP file are extracted.

The next step is to store in another file, the mutations in the NLP file in a specific format. In this step, care is taken to eliminate various types of errors and redundancies in the NLP file.

The next step is to extract the conserved domains using the National Center for Biotechnology Information (NCBI) conserved domain database[GBA+02]. Domains need to be extracted because non-domain portions of the sequence influences the final consensus that is generated, thus increasing the errors. These domains are extracted in a specific method and assembled sequentially. These domains are aligned using ClustalW multiple sequence alignment program[GHT94].

Alistat[Edd05] is used to generate statistics about the alignment. A percent pairwise identity is predefined called the "threshold", below which the sequences are eliminated from the alignment. The elimination is done by removing one sequence at a time and regenerating the alignment. This step is repeated untill all the sequences below a particular threshold are removed. Once this is done, *hmmbuild* and *hmmemit* are used to build a consensus sequence for the sequences that remain[Edd05].

The most similar structure sequence to the consensus is extracted from PDB[Bio05a]. This is done using BLAST[LGM+90]. This is the structure to which all the mutations are going to be mapped.

Once the structure is obtained, each of the sequences is aligned with this structure and this alignment is used to map the protein mutations on to the structure. In order to know if

the mutations are mapped well, a score is calculated for each mapped mutation. For each mapped mutation, this score, along with the context information, is written to the final visualization file.

The final visualization file is uploaded to PROSAT[HLWG03] for visualizing the mutations.

As discussed in section 4.5, there is a possibility of error at each step of the processing including possible errors at the NLP stage.

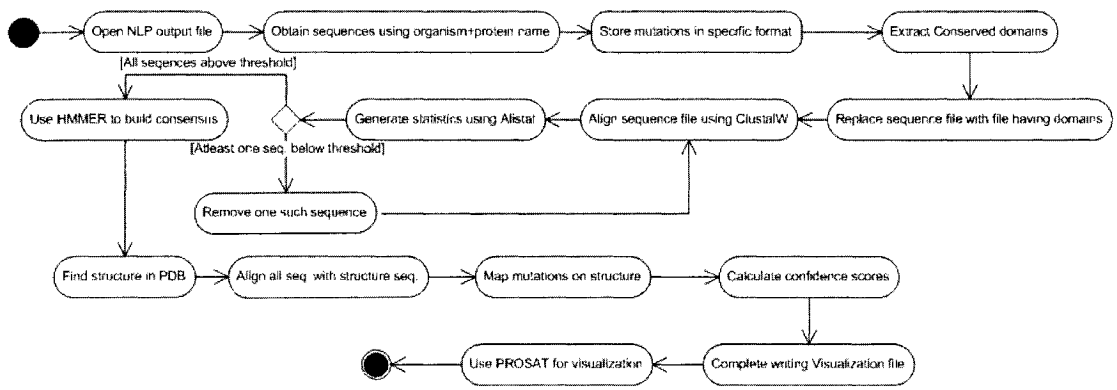

Figure 1.2 Activity Diagram Showing All the Steps from NLP Input File to the Final Visualization File

Perl[WCS96] is the programming language used for all the modules of the tool. ClustalW and Blast are currently being used for all the alignments in the tool. They have been chosen because they are global alignment algorithms, easy to use, easily available and integrate well with other tools. TCoffee[NHH00] and Dialign[Mor99b] are other alignment algorithms that can be used as alternatives to ClustalW. Smith-Waterman[SW81] and FASTA[LP85] are algorithms that could be used as alternatives to BLAST. Alistat[DEKM98] is used to generate the required statistics on the alignment because it is very well suited to

the format of the output file from ClustalW and generates the percent pairwise identity score for sequences, which is important in elimination of unwanted sequences. NorMD can be an alternative[PPR+01]. HMMER[DEKM98] is widely used for generating HMM models of alignments and generating consensus, and also readily available. A description of each of these tool is given in the Chapter 2.

1.4 Thesis Organization

The thesis is organized into an introduction which mainly discusses the problem, the choice of technology, and my work in this direction. The second chapter discusses the background work and the various tools and databases used in the tool. The third chapter discusses in detail, the implementation of the tool. The fourth chapter discusses the results obtained when the tool was tested on three families of interest, described in section 2.5 – xylanases[MS05], dehalogenases[Kol97], and biphenyl dioxygenases[JRM+99]. Finally there are conclusion, future work and references. The Appendices A through F give the actual input and output files used for the testing.

Chapter 2 Background

In this chapter, some background work related to the tool is discussed. First there is a discussion about the PubMed database, the NLP analysis required for providing inputs to this tool, and some details of three families of proteins that have wide industrial applications and which have been used for testing this tool.

2.1 Generating the NLP File

Natural Language Processing(NLP) is a field of computer science that aims to build software that understand, analyse and generate human language.

The NLP component of this tool is based on the widely used GATE [General Architecture for Text Engineering] framework[Cun02]. It has a component based architecture so new components can be added and old components can be removed. An input text content (in this case, it is a Portable Document Format — PDF document) undergoes a series of processing steps by these NLP components, which are described briefly as follows:

Preprocessing step: The input text is split up into tokens. A token can be a word or a phrase of the text. These tokens are then compared with precompiled lists. For non-biomedical information, the lists developed by the CLaC group for the newspaper domain, is used[BWK$^+$03]. For biomedical information, the list used by BioRAT system[BCJL04] is used.

Named Entity Recognition step: In this stage, the tokens are combined into named entities[BW04]. An example for a named entity is the name of a person, consisting of the first last and middle names. At this stage, the mutations are identified which can occur in several formats.

Sentence splitting step: Here the text is split into sentences and each word is associated with a part of speech tag such as noun, verb, etc.

Noun Phrase chunking step: Here the text is further analysed and "noun phrases" are identified. A noun phrase is a complex grammatical structure, which includes the head nouns, the modifiers, and the determiners. For example if a sentence says "The heat tolerance of protein for mutation N12D changed by 2%", then all the initial words of this sentence untill N12D is taken as a single noun phrase and the 2% is associated with the noun phrase rather than N12D. This helps in better annotation.

Relation detection step: This is the last step where the relation between the entities is identified. In our case, it is the relation between the protein and the mutation. Sentences containing a mutation expression are scanned to find protein names with the assumption that the mutations are done on the protein. This assumption has worked well in our case studies.

Mutation Representation in the NLP Output File: The mutations appear in the NLP output file as N10D, L20C, etc which means that a residue N at position 10 is mutated to D, a residue L at position 20 is mutated to C, and so on. Figure 2.1 shows the precision and recall values obtained when the NLP module was tested on the texts that describe the mutations on xylanase family. Precision(P) is the percentage of the chunk blocks that are correctly predicted. Recall(R) is the percentage of the true chunks that are correctly predicted. It is generally expressed in a value between 0 and 1. F-measure is the geometric mean of these two i.e 2*P*R/(P+R). It has a high value only if the precision and recall values are close enough. The fact that recall is 0.85 and precision is 1.0 shows that some of the predicted mutations are not mutations. Even the Protein/Organism recall value is

0.46 shows that several predicted proteins and organisms in the paper are not the true ones that are being discussed in the paper, but for the full paper, since F-measure is less for Protein/Organism than mutations, design improvements in the NLP module is possibly needed for better accuracy in obtaining Protein/Organism information.

	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

Figure 2.1 Precision and Recall Values of the NLP Module[BW05]

2.2 Description of Biological Terms

We provide a description of some biological terms as follows.

FASTA protein sequence format: The FASTA protein sequence format[Bio96] has two parts – the first line is a description line and the second part is the sequence. The description line starts with a > followed by a GenBank identifer which is the unique database identifier that identifies this sequence. The rest of this line is the description of the sequence. After this description is the actual amino acid sequence. The characters that can be used are those that correspond to the 20 amino acids[Che94] and also the following characters:

- gap of some length

* translation stop

X any amino acid

Figure 2.2 shows a sample FASTA sequence

```
>gi|151084|gb|AAB63425.1| biphenyl dioxygenase [Burkholderia sp. LB400]
MSSAIKEVQGAPVKWVTNWTPEAIRGLVDQEKGLLDPRIYADQSLYELELERVFGRSWLLLGHESHVPET
GDFLATYMGEDPVVMVRQKDKSIKVFLNQCRHRGMRICRSDAGNAKAFTCSYHGWAYDIAGKLVNVPFEK
EAFCDKKEGDCGFDKAEWGPLQARVATYKGLVFANWDVQAPDLETYLGDARPYMDVMLDRTPAGTVAIGG
MQKWVIPCNWKFAAEQFCSDMYHAGTTTHLSGILAGIPPEMDLSQAQIPTKGNQFRAAWGGHGSGWYVDE
PGSLLAVMGPKVTQYWTEGPAAELAEQRLGHTGMPVRRMVGQHMTIFPTCSFLPTFNNIRIWHPRGPNEI
EVWAFTLVDADAPAEIKEEYRRHNIRNFSAGGVFEQDDGENWVEIQKGLRGYKAKSQPLNAQMGLGRSQT
GHPDFPGNVGYVYAEEAARGMYHHWMRMMSEPSWATLKP
```

Figure 2.2 A Sample FASTA Sequence

Enzymes: The living cell is a site of tremendous biochemical reactions required for it proper functioning. Most of these reactions require the presence of a biological catalyst called the enzyme[Cor05]. A catalyst is a substance that accelerates a chemical reaction without itself undergoing a permanent change. Enzymes are also used in several industrial processes. Enzymes require the presence of other compounds called cofactors for their functioning.

Some enzymes catalyze a particular reaction. Others will catalyze specific groups of reactions. Hence based on specificity, they can be grouped into:

1. Absolute specificity - where the enzyme catalyses only one reaction

2. Group specificity - the enzyme acts on particular functional groups

3. Linkage specificity – the enzyme acts on a particular type of chemical bond

4. Stereochemical specificity - the enzyme acts on a particular isomer.

Enzymes are classified as follows on the basis of the type of reaction they catalyze:

1. Addition or removal of water

2. Transfer of electrons

3. Transfer of a radical

4. Splitting or forming a C-C bond

5. Changing geometry or structure of a molecule

6. Joining two molecules

More than 5000 enzymes are currently known.

2.3 Multiple Sequence Alignments

Protein sequences can be aligned across their entire length or only in certain regions. Alignment of sequences can be pairwise (two sequences) or multiple sequences at a time which is the multiple sequence alignment (MSA)[Not02][HT00][AD00]. In general, MSA are useful for the following:

1) MSA can be used to identify functionally important sites in sequences. Sequences have certain regions that have less or no mutations and are retained in evolution.

2) MSA helps us identify homologous sequences and to contrsuct phylogenetic tree.

3) MSA enables us to search for similarity among sequences of interest.

4) The number of protein sequences far outnumber the number of structures that have been discovered. Hence MSA are used to find structures that such a sequence would most likely have.

5) MSA is also useful for identifying the function of sequences.

Homologs: They are regions in protein sequences that are preserved by evolution and that play a very important role in the structure and function of a group of protein sequences.

MSA are very helpful in detecting homologous sequences. This is done by identifying elements that appear as columns with a lower level of variation than their surroundings.

Global and local alignments: If the alignment is done among sequences as a whole, it is global alignment, but if specific portions or modules of a sequence is to be aligned with other sequences, then a local alignment algorithm is to be used. Local alignment tools are also useful for detecting modules that repeat in the same sequence or occur in different positions among the sequences.

Substitutions: During evolution, some amino acids in a sequence get substituted by others. Generally such amino acids are limited to the ones having similar properties but substitutions with different properties are also possible. Multiple sequence alignment algorithms maintain a matrix having values for various naturally occuring substitutions. They use it for creating and scoring the alignments. In the ClustalW MSA algorithm, which is very popular, substitution matrices are automatically selected and varied at different alignment stages according to the sequences being aligned. Also, substitutions between two purines or two pyrimidines are more frequent than substitutions between a purine and a pyrimidine. Hence a MSA algorithm would generally weigh more heavily, a substitution between purine and pyrimidine. Some MSA algorithms introduce gap penalties — one for a gap opening and other for each gap extension in the sequence. Some algorithms are optimal and produce the best alignments but take too much time and have heavy space requirements (example is IBC MSA[Com05] algorithm). Hence they are not suitable for more than a small number of sequences.

Some examples of MSA algorithms are:

Progressive global MSA tool: ClustalW[GHT94]

Block based global MSA tool: Dialign2[Mor99a]

Motif based local MSA tool: Dialign[Mor99a]

Progressive Global Alignment: In these algorithms, the alignment score is calculated between all pairs of sequences. Using the pairwise alignment distances thus obtained, a guide tree is built. This tree is then used for performing the aligment, which begins at the leaves of the tree and ends at the tree root. ClustalW is an example of such an algorithm. This technique has the following problems:

1) Any mistakes made at an early stage in this algorithm cannot be corrected at a later stage in the algorithm.

2) When non-overlapping sequences are being aligned, these algorithms may give incorrect alignments

Block based global alignment: Sequences share common domains or blocks. Block based algorithms align the blocks occurring in these sequences, after identifying them, but blocks may not be exactly conserved. Hence these algorithms use special techniques to find the blocks. Dialign is an example of such an algorithm.

Stochastic Iterative algorithms: Stochastic iterative algorithms are based on the principle that an optimal solution can be obtained by refining an existing sub-optimal solution. Genetic algorithms[Wal05] are generally used here. Out of a given set of sequences, random alignments are formed and these random alignments constants evolve towards an optimal alignment. The changes may include the insertion or gaps, recombination of different alignments and an objective function decides the score of the alignments. Example is SAGA[NH96]. It gives good alignments and work is being done to improve its efficiency.

Non-Stochastic Iterative algorithms: In the approach of non-stochastic iterative algorithms, mistakes that occur at early stages of alignment are corrected in an iterative manner by aligning each sequence to the multiple alignment using dynamic programming.

Consistency algorithms: Consistency based algorithms align two sequences at a time using one method or a collection of methods, and do not depend a specific substitution matrix. Also, the score associated with the alignment of two residues depends on their position within the protein sequence rather than their individual nature. Given a set of independent observations, the most consistent are often closer to the truth[Not02].

Comparing different methods of alignment: If sequences are more than 50% similar, global alignment algorithms give good results. However, if two sequences among them share less identity, the result varies. If sequences have conserved domains and they are consistent among the sequences, then a block based global alignment tool like Dialign is appropriate. If sequences have conserved blocks that are not consistent, then a motif based alignment method such as MEME is used[HT00].

2.4 Tools and Databases Used

Given below is a description of the various biological tools and databases, that have been used in this project.

BLAST: Blast[LGM$^+$90] is a very useful tool to find the most similar sequences for a given query sequence, in a given database. Each Blast ouput hit has an e-value associated with it. The e-value is the number of hits one can expect to see by chance in a database of a particular size. The lesser the e-value for a hit, the better is its similarity to the query. Blast

has a variety of programs for specific types of queries and a complete manual is available online at the NCBI website[Inf05b].

BLAST works by aligning the query sequence with every sequence in the database, selected by the user. The result is reported by a ranked list of the sequences followed by the alignments of the sequences with those of the database and also scores related to the alignments. Each hit in the output has a score S, and an E-value which tells the number of hits with a score greater than or equal to this score S, in the database.

RPS-Blast: RPS-Blast(Reverse Position Specific Blast) is the reverse of the PSI Blast [LGM+90]. In the PSI Blast, a query sequence is used to search a database and the resulting alignment is used to build a PSSM (position specific score matrix). This PSSM is used to further refine the search, but RPS-Blast uses the query to search a database having pre-calculated PSSMs and reports the significant hits.

ClustalW: ClustalW[GHT94] is a multiple sequence alignment (MSA) tool, that has been used in this project along with BLAST for generating multiple sequence alignments. In this project, BLAST alignment is used to map mutations on the structure file and ClustalW is used in the initial MSA used to eliminate unrelated sequences. This because in general, ClustalW tends to perform better for sequences which are closely related or belong to the same family. ClustalW also is one of the most popular and easily available multiple sequence alignment tool.

ClustalW is a global, non-iterative, progressive multiple sequence alignment algorithm. ClustalW first calculates the alignment scores between all pairs of sequences. It next builds a tree that shows this similarity, based on these scores. Next it uses a pairwise sequence alignment algorithm to align the two sequences of each node. This is repeated starting from

the tree leaves till the root of the tree is reached.

The problems associated with this method are that the mistakes that that occur during the early alignments cannot be corrected later when new sequences are added and also the alignment is not correct when totally non-overlapping sequences are aligned.

Figure 2.3 shows a portion of the alignment generated by ClustalW. The figure shows the sequence description in FASTA format followed by the portion of sequence being aligned. The alignment also shows some symbols whose meanings are as follows:

* Symbol means that the residues in that column are identical in all the sequences.

: Symbol means conserved substitutions are observed

. Symbol means that semi-conserved substitutions are observed

```
gi|67399|CDD1|pir||WWBSXP          NT-SMTLNNGGAFSAGWNN--IGNALFRKGKKFDSTRTHHQLGNISINY-
gi|135144|CDD1|                    NT-SMTLNNGGAFSAGWNN--IGNALFRKGKKFDSTRTHHQLGNISINY-
gi|5381269|CDD1|dbj|BAA82316.1     GSGSMTLNSGGTFSAQWSN--VNNILFRKGKKFDETQTHQQIGNMSINY-
gi|21465565|CDD1|pdb|1H4G|B        GSGTMILNHGGTFSAQWNN--VNNILFRKGKKFNETQTHQQVGNMSINY-
gi|139865|CDD1|sp|P09850|XYNA_     GIVNAVNGSGGNYSVNWSN--TGNFVVGKG-------WTTGSPFRTINYN
gi|2619034|CDD1|gb|AAB84458.1|     GIVNAVNGSGGNYSVNWSN--TGNFVVGKG-------WTTGSPFRTINYN
gi|640242|CDD1|pdb|1BCX|Xylana     GIVNAVNGSGGNYSVNWSN--TGNFVVGKG-------WTTGSPFRTINYN
gi|17942986|CDD1|pdb|1HIX|B        GSVSMNLASGGSYGTSWTN--CGNFVAGKG-------WANG-ARRTVNY-
gi|6434133|CDD1|emb|CAB60757.1     SGVTYTNGAGGSFSVNWAN--SGNFVGGKG-------WNPGSSSRVINF-
gi|549461|CDD1|sp|P36217|XYN2_     GGVTYTNGPGGQFSVNWSN--SGNFVGGKG-------WQPGTKNKVINF-
gi|2851624|CDD1|sp|P55331|XYN1     LGDFTYDESAGTFSMYWEDGVSSDFVVGLG--------WTTGSSNAITYS
gi|465492|CDD1|sp|P33557|XYN3_     LADFTYDESAGTFSMYWEDGVSSDFVVGLG--------WTTGSSNAISYS
                                            .* :.  * :   .: :   *               :.:
```

Figure 2.3 A Portion of a Sample Alignment File Generated by ClustalW version 1.83

ClustaW works in three steps:

(a) The alignment scores are computed between all pairs of sequences.

(b) A guide tree is built that reflects the similarity between the sequences, using these scores

(c) Each node of the tree has an alignment associated with it and two nodes are aligned with each other. This process is continued ending with the tree root.

For example, of A1,A2..A5 are five sequences, A1 is aligned with A2, A3 is aligned with A4, then (A1,A2) is aligned with (A3,A4) and this alignment (A1,A2,A3,A4) is aligned with A5.

ALISTAT: Alistat[Edd05] is a tool that takes in an alignment in various formats and outputs simple statistics about the alignment. The statistics that alistat generates includes the name of the input format, the number of sequences, residues, sequence lengths, alignment lengths, etc.

A percentage pairwise alignment identity is defined as the fraction:

(Number of exact identities) / (minimum of the unaligned lengths of the two sequences)

From N initial sequences, *N(N-1)/2* sequence pairs are possible. Alistat also reports the average, minimum, and a maximum value of the percent pairwise identity scores. Alistat is used in this project to eliminate sequences that have a percentage pairwise identity score below a specified threshold value.

HMMER: HMMER[DEKM98] is a tool based on the hidden markov model (HMM), described in the next paragraph. The commands *hmmbuild* and *hmmemit* of HMMER are used here to obtain the consensus sequence from a set of given sequences. *hmmbuild* is a command of HMMER tool that constructs a HMM from a multiple sequence alignment. *hmmemit* uses the HMM generated by *hmmbuild* and generates the consensus sequence.

An HMM is a finite set of states, each of which has a probability distribution. Each state has its outcome based on the probability distribution. Transitions among the states are governed by transition probabilities. Only the outcomes can be seen but not the states. Hence

the name hidden.

The three problems in HMM are as follows :

(a) Given a model and a sequence of observations, what is the probability that the observations are generated by the model?

(b) Given a model and a sequence of observations, which model produced the observations?

(c) Given a model and a sequence of observations, how should the model parameters be adjusted to maximize probability that the model produces these observations?

HMMER uses profile HMM[Edd98]. For building a profile HMM, the training is done using an existing alignment. In profile HMM, for each consensus column of the alignment, a match state models the distribution of the residues in the column.

Consensus Sequence: Of a set of related protein sequences, consensus sequence[Onl05b] is the sequence that reflects the most common choice of base or amino acid at each position. The generation of consensus sequences has been subjected to intensive mathematical analysis.

Perl: The Mutation miner tool is implemented in Perl language[WCS96]. It stands for *Practical Extraction and Report Language*. Perl has powerful text manipulation and file handling functions. It is also the language of choice for many internet related applications. It is available for various operating systems and is convenient to use. The regular expressions of perl can be used for almost any text handling requirements. This is probably the reason why Perl is widely used for bioinformatics software development.

Perl takes the best features of other languages like *C*, *awk*, *sed*, *sh* and supports a variety of databases. Most bioinformatics databases have their online utilities written to facilitate

easy data extraction using Perl. An example is the EFetch utility of NCBI. Existing modules in other languages can also be run through a perl program.

Due to these reasons, Perl is the language used in the development of this tool.

ProSAT: ProSAT (Protein Structure Annotation Tool)[HLWG03] is a tool that allows automated visualization of functional annotations in 3-D structures. It can also be used for identifying functional units of a protein. New structures are always discovered and it is important to map the functional annotations of sequences on to structures for better understanding of their functions. Generally several tools need to be used in combination to obtain such information, but ProSAT alone has the capability to provide this information. Information from Prosite and Swiss-prot protein databases can be mapped to ProSAT. Mutation miner is a tool that has a capability to generate mutation information that can be visualized in ProSAT.

ProSAT has the capability to provide functional information for about 70% of the proteins using the Swiss-prot and prosite annotation. ProSAT also has the capability to interface with other tools like Rasmol and Protein Explorer[Fou96].

ProSAT uses the three databases — PQS Macromolecular Structure Database[Ins03], PROSITE Database of Protein Families and Domains [Bio05b] and Swiss-Prot Protein Knowledgebase [Sys05].

2.5 PubMed

PubMed[Med05] was developed by the NCBI[Inf05b] and the National Library of Medicine. It was designed to provide access to citations from Biomedical literature. It also has links

to other molecular biology resources. PubMed coverage includes Medline, Oldmedline and several other citations.

Medline is a premier database covering the fields of medicine, nursing, dentistry, the health care system and the pre-clinical sciences. It has more than 12 million articles from over 4800 journals published all over the world[Med05].

Oldmedline contains very old articles belonging to the 50s and 60s. These citations are not updated and do not have abstracts. There are about 2 million such citations.

PubMed also provides citations that have yet to undergo modifications before being added to Medline. This also includes publisher supplied citations. PubMed also has a citation matcher that one can use to find citations of PubMed that match their own.

The Entrez *e-utilities* [Inf05a] can be used to obtain details from PubMed just like they can extract information from any other NCBI database.

2.6 Protein Mutant Database

Each entry in the Protein Mutant Database (PMD) [KNO99] is not based on proteins but on a particular paper that describes the protein mutants. Each entry has the following attributes - JOURNAL, TITLE, CROSS-REFERENCE, PROTEIN, N-TERMINAL, CHANGE, FUNCTION, STRUCTURE, STABILITY, etc. The CROSS-REFERENCE indicates the code names of proteins, that can be cross referenced to other protein databases. CHANGE describes the mutations itself including insertions, deletions and substitutions. The mutant FUNCTION, STRUCTURE , STABILITY are described after the CHANGE. To indicate whether the mutants have differences in property with the actual protein, these

notations are used : [-],[–],[+],[++],[=],[0]. A [0] indicates that the property does not exist. A [-] indicates a negative effect of the mutation on the property. [–] indicates that there is more difference between the actual protein and the mutated version. Similarly [+] and [++] indicate an increase of a particular property due to the mutation.

A sample PMD entry is shown in Figure 2.4. This is the first hit of a query with keyword "Globin". It shows the mutations called as CHANGE-POINT and also their effects on the protein. For example, the Phe 123 Trp means the aminoacid Phenylalanine at position 123 has been mutated to aminoacid Tryptophan. The STABILITY indicates the stability property of this mutation. The ENTRY indicates the database entry, the AUTHORS indicate the authors of the journal, the MEDLINE entry corresponding to the paper is also indicated. The JOURNAL is the name of the journal. The TITLE indicates the name of the article, that speaks about these mutations, the PROTEIN, EXPRESSION-SYSTEM and SOURCE indicate the protein, the organism and the source of the organism, from which the protein was obtained. The CROSS-REFERENCE has links to Swiss-Prot and PDB databases.

```
ENTRY            A000451 - Artificial                    2618066
AUTHORS          Bondos S.E., Sligar S. & Jonas J.
JOURNAL          Biochim.Biophys.Acta (2000) 1480(1/2), 353-364
                 [LINK-TO-MEDLINE]
MEDLINE          11004573
TITLE            High-pressure denaturation of apomyoglobin.
CROSS-REFERENCE  MYG_PHYCA/1BZP
                 [LINK TO SWISS-PROT "MYG_PHYCA"]
                 [LINK TO PDB "1mlm-"]
PROTEIN          Apomyoglobin (apoMb); myoglobin
SOURCE           Sperm whale
N-TERMINAL       VLSEG
EXPRESSION-SYSTEM  Escherichia coli
CHANGE-POINT     Phe 123 Trp
STABILITY        This mutation perturbs the denaturation of apoMb to the high-
                   pressure intermediate.
CHANGE-POINT     Ser 108 Leu
STABILITY        This mutation perturbs the denaturation of apoMb to the high-
                   pressure intermediate.
CHANGE-POINT     Ser 108 Lys
STABILITY        This mutation perturbs the denaturation of apoMb to the high-
                   pressure intermediate.
CHANGE-POINT     Ala 130 Lys
STABILITY        This mutation perturbs the denaturation of apoMb to the high-
                   pressure intermediate.
...
```

Figure 2.4 A Sample PMD Entry

PMD also has a facility for the user to input an amino acid sequence and it would show the mutations for that particular sequence. Another example is shown in Figure 2.5. In this figure too, the ENTRY, AUTHORS,JOURNAL and other keywords mean the same as Figure 1.6 but here the CHANGE-POINT clearly highlights the mutation on the amino acid

sequence, which was given as the query.

```
ENTRY            B993265                                  2520649
AUTHORS          Cool R.H., Schmidt G., Lenzen C.U., Prinz H., Vogt D. &
                   Wittinghofer A.
JOURNAL          Mol.Cell.Biol. (1999) 19(9), 6297-6305
                 [LINK-TO-MEDLINE]
MEDLINE          10454576
TITLE            The Ras mutant D119N is both dominant negative and activated.
CROSS-REFERENCE  (RASH_HUMAN/6Q21)
                 [LINK TO SWISS-PROT "RASH_HUMAN"]
                 [LINK TO PDB "1ctqA"]
PROTEIN          Ras protein; transforming protein p21/H-RAS-1
SOURCE           (Human)
N-TERMINAL       MTEYK
NATIVE-SEQUENCE     1 MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG  60
                   61 QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDL 120
                  121 AARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPG 180
                  181 CMSCKCVLS                                                    189
EXPRESSION-SYSTEM
```

```
CHANGE [1]
CHANGE-POINT     Asp 119 Asn
      -SEQUENCE     1 MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG  60
                   61 QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCNL 120
                  121 AARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPG 180
                  181 CMSCKCVLS                                                    189
```

Figure 2.5 PMD Entry Showing the Mutations for a Particular Sequence

PMD uses the following databases:

1. PRF (Protein Research Foundation): based in Japan and collects information related to amino acids and proteins. It consists of articles from scientific journals, protein sequence data, data on synthetic compounds and molecular aspects of proteins.

2. PIR - Protein Information Resource: It is located at Georgetown University medical centre provides many databases and tools for protein research. PIR has joined with Swiss-Prot and TrEMBL to create Uni-Prot[Sys05].

3. Swiss-Prot: it is a curated protein database which provides a high level of annotation and provides the function of a protein, its domain structure, modifications, variants, etc. It

has minimum redundancy.

4. PDB (Protein Data Bank) : It is the major repository of 3-D structures for proteins and nucleic acids. The structures of proteins are available in standardized format files, which can be used in other visualization tools.

PMD search methodology: If the ENTRY field is known, it can be used. The mutation itself can be entered as a query: *from P to A* will list all mutations having this change. Deletion, Insertion, Termination, Extension can be indicated using radio buttons on the search interface and details about these can be obtained. A query can have two keywords with logic AND, OR between them, which can be specified by clicking the appropriate radio button on the search interface. Example is *human AND myoglobin.*

PMD has the following facilities currently:

(a) For a particular protein, it has a facility for visualizing the mutations on a tertiary structure. The mutations are showed in a different color. Chooses a structure with more than 50% identity to the given sequence.

(b) It has a facility for showing mutations on the sequence.

(c) It has a facility to retrieve a set of homologous sequences when the user specifies a similarity threshold of 30% to 100%.

(d) It has the facility to display a summary of a particular mutation for all these homologs.

ESearch utility: The ESearch utility[Inf05a] is a utility provided by NCBI for use along with other utilities like EFetch. Is used to obtain the primary ids for use in EFetch utility. It has a base URL with a facility to provide values for following parameters:

db is the database name which is to be used.

usehistory maintains results in users environment.

WebEnv is the value previously returned by ESearch.

Query–key is the value used for a history search number

Tool is the parameter that specifies which tool is using the NCBI facility and NCBI uses this information for better tracking.

Email allows user to specify the user's email so that NCBI can contact the person if required.

Retmode specifies the expected mode of output. It could also be an XML file.

Retype is the output type. "FASTA" is an example of a value that can be given to this parameter.

Term is the search term or phrase

An example is shown in Figure 2.6. This is searching for a protein whose molecular weight is 200020 atomic mass units.

```
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=protein&term=200020[molecular+weight]
```

Figure 2.6 An ESearch URL to Search a Protein with a Particular Molecular Weight

EFetch utility: The EFetch[Inf05a] utility is very useful to retrieve records of any NCBI database. Its usage is similar to ESearch for a typical biology database search, the following parameters are used:

Db is the database name like "gene", "protein", etc.

Webenv and *query–key* mean the same as in the ESearch tool.

Tool and *email* parameters are used here too and retype and retmode have the same meaning.

```
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?&extrafeatpresent=1
&ef_CDD=8&db=protein
&id=$query&rettype=gp&retmode=xml
```

Figure 2.7 Using the EFetch Utility for Each Sequence in the Sequence File

Figure 2.7 shows a sample URL for using EFetch. It sets the database as *protein*, the *retmode* is XML for an XML output, the *rettype* is set to "gp" [Inf05a] and sets *ef_CDD* as 8 to indicate that conserved domains details are to be output. *id* is the query string. Parameters like *retmode*, *rettype* can be changed to obtain outputs in different formats.

2.7 Protein Families of Interest

A description of the three families of proteins that are of interest to us is provided (because they are all having very important industrial applications) and which have been used for testing. The test results are provided in chapter 4.

Xylanases: Xylan is the most abundant polysaccharide occurring in plants and wood and can be upto 35% of their dry weight. Hence especially for the paper industry, xylan-degrading enzymes are very important. Xylanases[MS05] are these enzymes that have the capability to degrade xylan. These enzymes have wide industrial application. For example in the paper industry, xylanases are effective bio-reagents that can be used for bleaching the pulp to achieve pulp brightness of high order instead of the conventional use of chlorine, which is toxic.

The main sources of xylanase are microorganisms like bacteria and fungi. There is research going on to identify xylanase sources that can function in high alkaline ph conditions. Application of xylanase has already been done in industries with successful results.

Dehalogenases: Halogens are a group of highly reactive elements including chlorine, fluorine, bromine. Halogenated organic compounds are compounds containing halogens, which are widely present in the environment and pose a hazard to the environment because of their persistence and toxic nature. They are present in the various chemicals commonly used including fungicides, insecticides etc. Dehalogenases[Kol97] are enzymes that have the capability to effectively break down these toxic substances like PCB (Polychlorinated Biphenyl), which is an environmentally harmful compound. Genetic engineering could produce better dehalogenases that could do this more effectively. Dehalogenases find the following applications in industries: (a) Microbial dehalogenases are used as biocatalysts in industries (b) They are used in waste management and environment protection programs

Biphenyl Dioxygenases: Biphenyl dioxygenase is another family of enzymes that can effectively break down PCBs and other environmental pollutants. They can also be used for synthesizing many organic compounds, which are difficult to produce using conventional methods.

Biphenyl is a compound containing two benzene rings. Biphenyl-utilizing bacteria effectively breakdown PCB into chlorobenzoic acids, using the biphenyl-catabolic enzymes. Several such bacteria have been identified[JRM+99].

The *bph* genes of these microorganisms encode these enzymes. Genetic modifications of these genes can lead to enzymes that can be used for other special chemical reactions.

2.8 Conserved Domains

Conserved domains are distinct structural or functional units of a protein. They are independently folding units of a polypeptide chain, that carries a function. Molecular evolution has used such domains as building blocks. These domains are summarized as a multiple sequence alignment and is placed in various databases. They can be found from three major sources: The SMART database which stands for Simple Modular Architecture Research Tool[LCS+04], the PFAM database of UK a database of curated seed alignments[Ins05], the COG(Clusters of Orthologous Groups) database[TKL97].

The conserved domains can be searched either by the protein name or by its own name.

RPS-Blast, described in section 2.6, is commonly used in the search used to find conserved domains. It first uses the query to find a seed set of sequences. This is then later used to find the other similar sequences.

Searching the conserved domains: To search the conserved domains, the sequences of the query (protein) needs to be input in raw or fasta format. The e-value can be changed and some special filters can be used to change the output. NCBI provides users with online utilities like EFetch and ESearch can be used to automate the extraction of CDD information. When only conserved domains of sequences are aligned, they give a better alignment and better consensus than when entire sequences are aligned. Regions other than conserved domains sometimes align badly and lead to poorer scores and this leads to removal of sequences, which are actually required for the mapping. Inorder to obtain a better consensus sequence for more accurate mapping of mutations, the input protein sequences in the sequence file needs to be split into conserved domains. This is done by using EFetch utility and using the GenBank identifier of the proteins as the search query. A sample URL that

uses EFetch is shown in Figure 2.7. The output for this URL would be an XML file having conserved domain information corresponding to the query, which is a single protein here.

Chapter 3 Implementation Details

This chapter describes in detail each step performed by the tool starting from this XML file, to the stage where the mutations are mapped to the structure. First the processing of the NLP output XML file is explained. After this, all the steps illustrated in Figure 1.2 are explained. This includes the conserved domain processing, consensus sequence generation, structure selection and mapping of mutations.

3.1 Processing the XML file generated by NLP module

All XML files in the tool are processed using the XML::Parser module of Perl. A sample of the output from the NLP analysis is shown in Figure 3.1. The PubMed paper must be renamed by its PubMed identifier and must be fed to the NLP system. This is how the NLP module identifies the PubMed identifier of the paper it is processing. The sample portion of the text that generated the protein name and the organism name is given below:

"..In order to understand how the substitution of a single amino acid residue can modulate the pH optimum of an enzyme, we have focused our attention on the "alkaline" xylanase from Bacillus circulans (BCX). This 20.4 kDa protein has been characterized extensively using a wide range of structural, spectroscopic, and enzymatic techniques.." [JSP+00], Page 256.

The sample portion of the text that had mutation information is:

"..As predicted from sequence comparisons, substitution of this asparagine residue with an aspartic acid residue (N35D BCX) shifts its pH optimum from 5.7 to 4.6, with 20%

increase in activity.." [JSP+00], Page 255.

For each of the proteins in the NLP output file, the protein name and organism name is concatenated and made into a single search string. A sample of such a search string is *Bacillus circulans[organism] xylanase[title]* — the braces help in obtaining a better search result.

The above search string is used with the ESearch and EFetch utilities to obtain the sequence information for each protein of the NLP output file. The first hit in the search is chosen as the required sequence. Since these utilities can be incorporated into a program, the sequence retrieval process has been completely automated. The sequences thus obtained are written to a separate XML file. The PubMed identifier along with the mutation are written to another XML file. The first file is the sequence file and the second one is the mutation file.

```
<Protein>
   <Name>xylanase</Name>
   <Organisms>
      <Name>Bacillus circulans</Name>
   </Organisms>
   <pmid> 10860737 </pmid>
   <Mutations>
      <Mutation>
         <Mark>N35D</Mark>
         <Context> context information here </Context>
      </Mutation>
      <Mutation>
         <Mark>C5H</Mark>
         <Context> context information here </Context>
      </Mutation>
   </Mutations>
</Protein>
```

Figure 3.1 Sample Portion of the NLP Output XML File

The sequence file has only the sequences of the proteins whereas the mutation file has mutations for the same proteins but divided based on the PubMed identifier, which is obtained from the NLP output file and is the number that PubMed uses to uniquely identify a paper. The output file from the NLP analysis needs to be processed to remove redundancies and errors. One such redundancy is the multiple occurrence of the same mutation. The multiple occurrences of the mutations are eliminated, before the mutations are written to the mutation file. Also mutations such as +20D occur in the NLP output. These mutations are of no use in the mapping and are removed. The sequence file and mutation file are written after these redundancies are removed. The mutations are also sorted in ascending order before being written to the mutation file.

For each sequence and each PubMed identifier, one set of mutations are written into the sequence file. This is because if a particular paper is speaking about more than one protein, then the GenBank identifier would change or if two papers are speaking about same protein, then PubMed identifier will change. Thus the combination of the GenBank identifier and PubMed identifier is used to uniquely identify a set of mutations in the mutation file.

A sample of the generated sequence file and mutation file are shown in Figure 3.2 and Figure 3.3 respectively. The sequence file is used in the next stage as described in the next chapter and mutation file is used much later when the actual mapping of the mutations is done on the structure.

An activity diagram of this procedure is given in Figure 3.4.

```
<component>
<id>61222635</id>
<desc>gi|61222635|sp|P0A3G3|DHAA_RHOSO Haloalkane dehalogenase</desc>
<sequence> sequence of the protein appears here </sequence>
</component>
```

Figure 3.2 A Sample Portion of the Sequence File

```
<component>
<gi>59799356</gi>
<pmid>10100638</pmid>
<mutation>D108N</mutation>
<context> Context here </context>
<mutation>D124N</mutation>
<context> Context here </context>
<mutation>E132Q</mutation>
<context> Context here </context>
<mutation>E244Q</mutation>
<context> Context here </context>
<mutation>H272A</mutation>
<context> Context here </context>
</component>
```

Figure 3.3 A Sample Portion of the Mutation File

The algorithm for this module is:

Name of algorithm: NLP output file processing

Purpose: To process the output file generated from NLP analysis and write the sequence information to a sequence file and mutation details to a mutation file.

Assumptions: The NLP output file is complete and available for processing and is error free and has the organism names and protein names with their mutations in a predefined XML format.

Description of Inputs: The NLP output XML file

Description of Outputs: The sequence file and mutation file

Major Variables:

String protname = protein name being currently processed

String orgname = Organism name in which this protein is found

String paperid = Pubmed paper's PubMed identifier from where the information is extracted

String seq = sequence string extracted from NCBI

String con = context information extracted from the Pubmed papers.

String finalstr = concatenated value of the organism name and protein name

int pvariable= PubMed identifier

Steps:

Write the headers for the mutation file and sequence file
open(NLP file)
While(! EOF)
{
    Read protname and orgname.
    finalstr = protname + " " + orgname
    Use finalstr with ESearch and EFetch and obtain the GenBank identifier and sequence

information from NCBI

seq = obtained protein sequence.

Write this information to the sequence file.

Extract pvariable= PubMed identifier for this GenBank identifier from the NLP output file.

Write pvariable to the mutation file.

Extract con = context information from the NLP output file and write it into mutation file.

Write the mutations for this combination of GenBank identifier and PubMed identifier into the mutation file.

}

Complete writing the mutation and sequence files

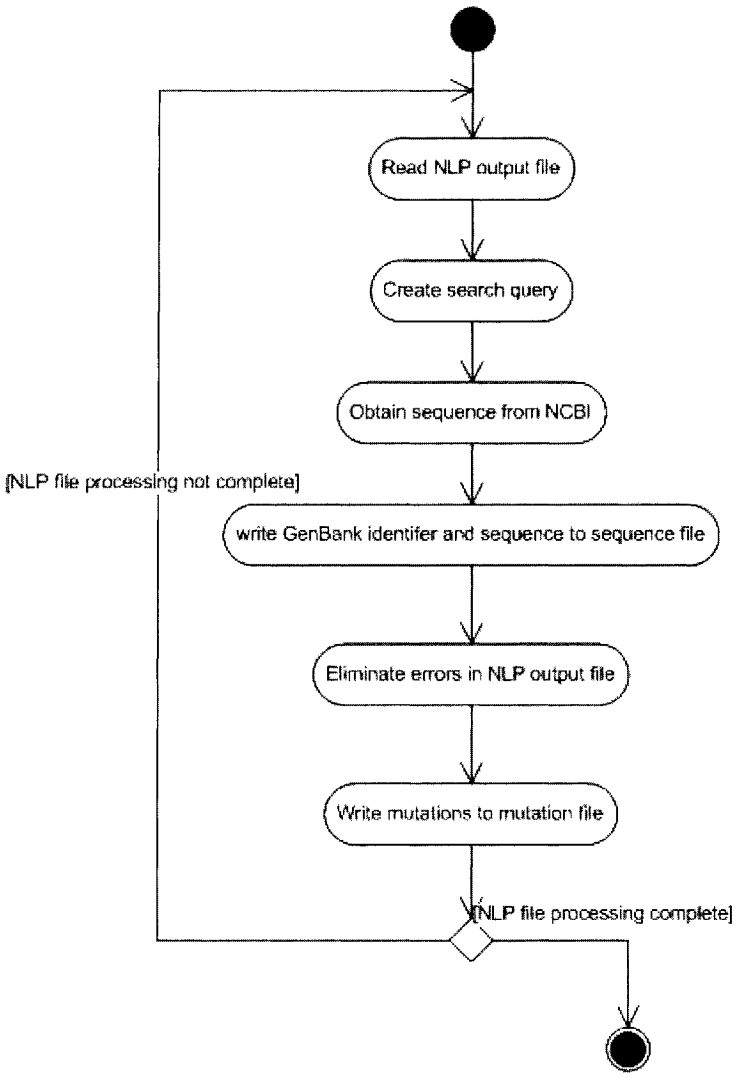

Figure 3.4 An Activity Diagram Showing the Steps Adopted to Generate the Sequence and Mutation Files from NLP Output File

3.2 Conserved Domain Extraction

The next step is to extract the conserved domains using the NCBI conserved domain database. Domains need to be extracted because non-domain portions of the sequence influences the final consensus that is generated, thus increasing the errors. These domains are extracted in a specific method and assembled sequentially. Each of the input sequences in the sequence file are subjected to the following procedure:

If a sequence has only a single conserved domain, then 20% (this is an arbitrarily chosen value) of the domain length is calculated and the residues for this length are retained around the beginning and end of the conserved domain. The rest of the residues are removed. This conserved domain with 20% residues on either side is considered as a new sequence. This is shown in Figure 3.5.

If a sequence has more than one conserved domains, the conserved domains are separated into separate sequences but here, the splitting is done in the region which is midway between the end of one domain and start of the next one. This is shown in Figure 3.6.

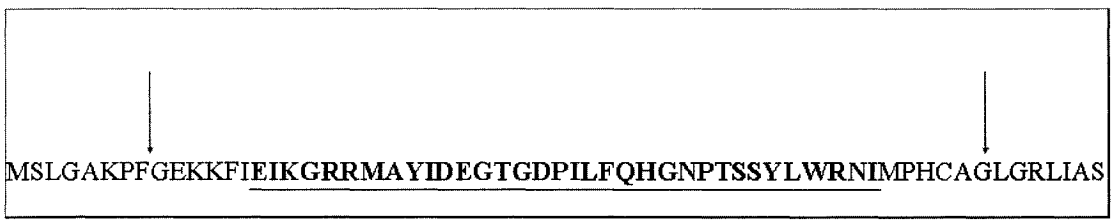

Figure 3.5 The Underlined and Blocked Residues are one Conserved Domain and is extracted according to the 20% rule

Figure 3.6 The Underlined and Blocked Residues are Conserved Domains and the Arrows Indicate the Position where the Sequences are Spliced

An activity diagram for the conserved domain processing is shown in Figure 3.7. The algorithm for this module is:

Name of algorithm: conserved domains extraction

Purpose: To extract the conserved domains of the sequences and to assemble them sequentially into a FASTA file.

Assumptions: NCBI Conserved domain database is accessible from the system. Sequence file is already written and available for processing.

Description of Inputs: The sequence file generated in earlier module

Description of Outputs: A new sequence file having only the conserved domains arranged sequentially, which overwrites the original sequence file.

Major Variables:
String gi = GenBank identifier obtained from NCBI protein database.
String cdata = Character data value of any XML tag of the sequence XML file.
int count = number of Conserved domains in the protein sequence.

Steps:

for each sequence in the sequence XML file

{

Using the GenBank identifier as query, and CDD as database, use EFetch to get XML file with domain details, from NCBI and store it in file TEMP.

Find the number of conserved domains for the sequence from TEMP.

count=0 if there are no domains, count =1 if there is a single domain, count=2 if there are more than one domains.

Switch(count)

{

case 0: Write the sequence without any change to the new sequence XML file.

case 1: Extract the conserved domain sequence along with 20% of its length before and after the sequence.

Write this extracted sequence to the new sequence XML file.

case 2: Slice the sequence between the CDDs and arrange the CDD sequences in the new sequence XML file.

}

}

Complete writing the new sequence XML file.

Replace the original sequence XML file with the new sequence XML file.

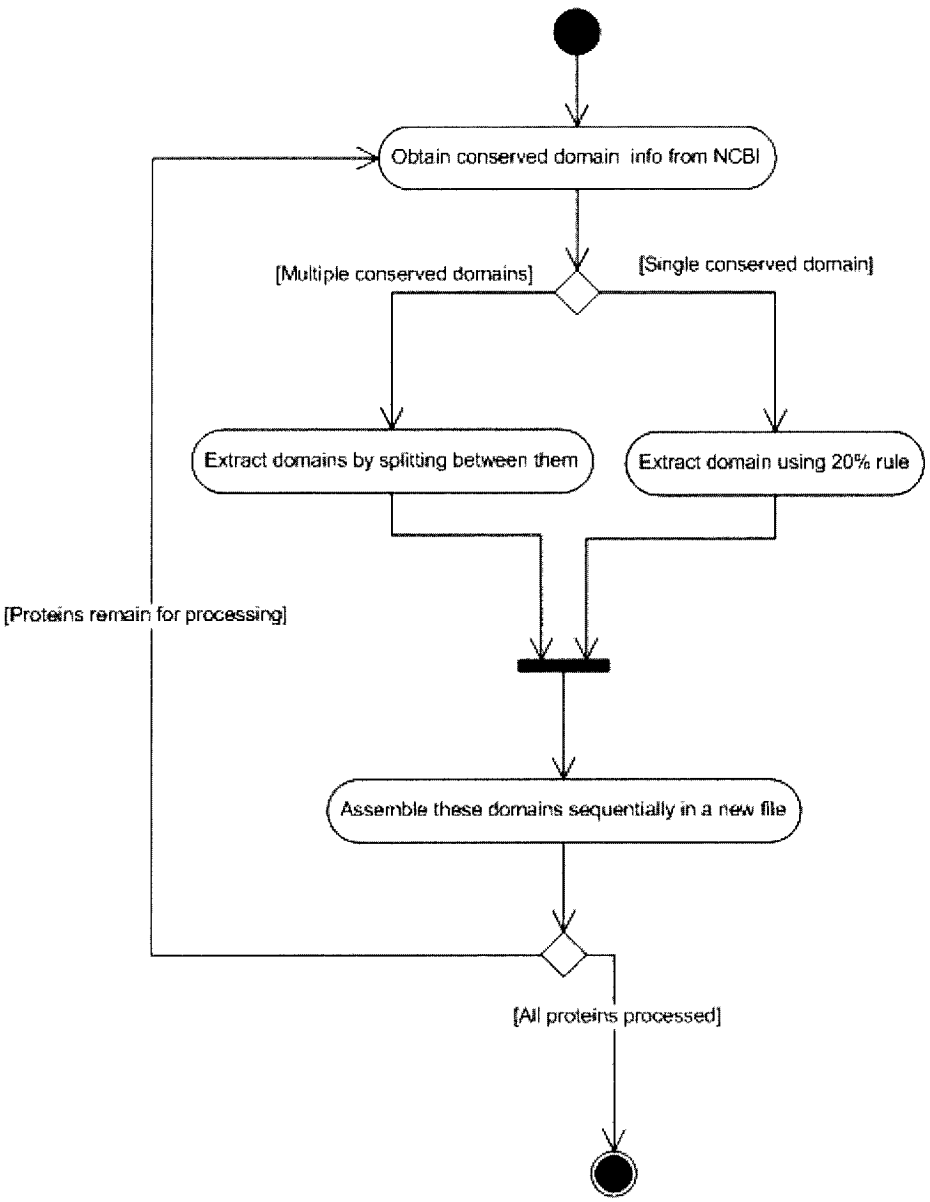

Figure 3.7 An Activity Diagram for the Extraction of Conserved Domains from Every Sequence in the Sequence File

3.3 Generating the Consensus Sequence

Alistat is used to generate statistics about the alignment. A percent pairwise identity is predefined, called the "threshold", below which the sequences are eliminated from the alignment. The elimination is done by removing each sequence at a time and regenerating the alignment. This step is repeated till all sequences below a particular threshold are removed. Once this is done, *hmmbuild* and *hmmemit* are used to build a consensus sequence for the sequences that remain. A typical alistat output is shown in Figure 3.8. The figure shows that the protein sequence with GenBank identifier 61222635 has a pairwise percent identity of 48.8% with the protein sequence with GenBank identifier 59799356.

```
gi|61222635|sp|P0A3G3|DHAA_RHO   293    48.8 gi|59799356|gb|AAX07227.1|halo
```

Figure 3.8 A Line from Alistat Output Showing Pairwise Identity Between Two Sequences

A typical Alistat output file has such a list of percent pairwise identities between the various pairs of sequences with other statistics.

The program then processes Alistat output and removes one sequence that has lowest pairwise identity in the ClustalW alignment. This removal is done by converting the FASTA file back to xml file without the sequence, which is to be eliminated. Next, this xml file is again converted back to fasta file and is made the input to ClustalW and the procedure is repeated till all sequences have a pairwise identity above a threshold value, which we call as the "Alistat threshold". Only one sequence is removed in each iteration because the alignment and pairwise identity values change in every iteration.

Regular expressions in Perl are used to extract data from all text files including alistat output file.

Once all the sequences that fall below the alistat threshold are removed, the new alignment file is input to the *hmmbuild* tool, and a HMM file is created. Then, using the *hmmemit* tool, this HMM file is then converted to a file having the consensus sequence.

The activity diagram showing the creation of the consensus file from the initial sequence XML file is shown in Figure 3.9. The algorithm for this module is described as follows:

Name of algorithm: Consensus sequence generation

Purpose: To find the consensus sequences from a set of sequences.

Assumptions: The sequence file, ClustalW, Alistat, hmmbuild, hmmemit are available

Description of Inputs: The sequence file after the conserved domain processing.

Description of Outputs: The file having the consensus sequence

Major Variables:
String cdata = Character data value of any XML tag of the sequence XML file.
TEMP= FASTA file to store sequences
int thres = alistat threshold value
int pvalue = percent pairwise identity
int currpvalue = current lowest percent pairwise identity in Alistat output
Steps:
thres= a constant K
int currpvalue=0;

```
while(currpvalue < thres)
{
  open(sequence XML file)
  for each sequence in the file
  {
    convert to FASTA and write to TEMP.
  }
```

Use ClustalW to create an alignment using TEMP as input.

Use Alistat to generate statistics on this alignment.

Find sequence GenBank identifier with least percent pairwise identity.

currpvalue=least percent pairwise identity.

Create new sequence XML file from the current sequence XML file without this sequence.

Replace old sequence XML with the new one.

}//end of while

Use HMMBUILD to generate the HMM file from the alignment file

Use HMMEMIT to generate the consensus sequence from the HMM file

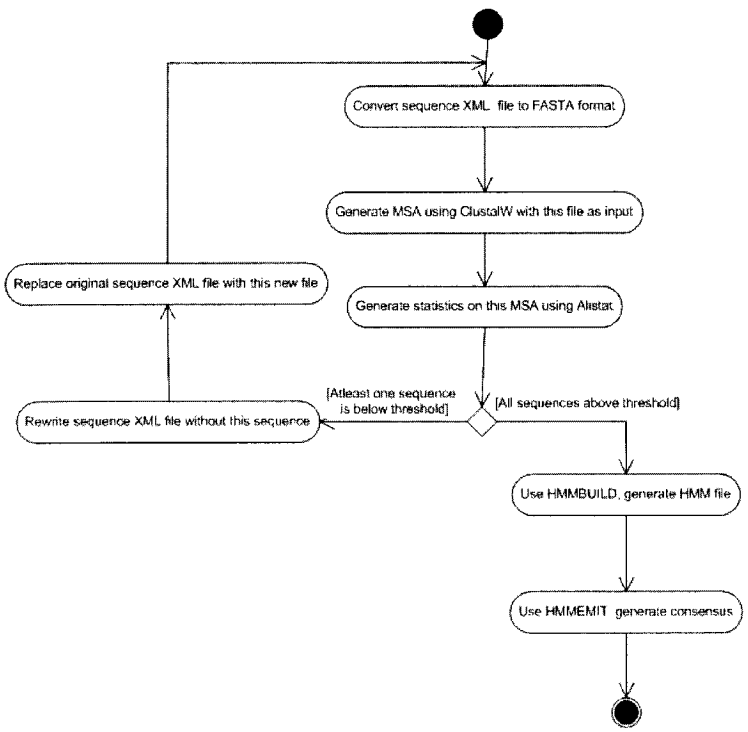

Figure 3.9 Generating the Consensus Sequence from the Sequence File

3.4 Selecting the Structure

The next step is to find an appropriate structure sequence, so that the mutations can be mapped to it. This is done using the consensus sequence generated in the step described in Section 3.3. A BLAST alignment is done on a local copy of the PDB database, with the consensus sequence as query. A sample of the output of such a BLAST search is shown in Figure 3.10. The 1BN6, 1CQW etc in the figure are the structures that are most similar to the consensus sequence. The top hit is taken as the most similar structure to the given consensus sequence. In this example, it is 1BN6A.

```
pdb|1BN6|A Chain A, Haloalkane Dehalogenase From A Rhodococcus S...   343   2e-095
pdb|1CQW|A Chain A, Nai Cocrystallised With Haloalkane Dehalogen...   343   2e-095
pdb|1K63|A Chain A, Complex Of Hydrolytic Haloalkane Dehalogenas...   296   4e-081
pdb|1MJ5|A Chain A, Linb (Haloalkane Dehalogenase) From Sphingom...   296   4e-081
pdb|1D07|A Chain A, Hydrolytic Haloalkane Dehalogenase Linb From...   296   4e-081
pdb|1G5F|A Chain A, Structure Of Linb Complexed With 1,2-Dichlor...   295   5e-081
pdb|1B6G|  Haloalkane Dehalogenase At Ph 5.0 Containing Chloride      293   3e-080
```

Figure 3.10 BLAST Output of Consensus Sequence Against the PDB Database

Using EFetch utility, the sequence corresponding to this structure is obtained from the NCBI protein database. Next, the sequence XML file is converted to a sequence FASTA file and using Formatdb[LGM+90], it is converted to a formatted database. A BLAST alignment is performed using this database with the selected structure sequence as the query. This alignment file is written to an output file in the FASTA format and another file in XML format. Both these files are required for separate purposes as explained in Section 3.5.

At this stage, all the information that is required to map the mutations is available: the

structure sequence, the alignment file, the sequence and mutation files. Next, the program checks the number of PubMed identifiers that exist for each protein sequence and writes a separate file which we call as the "mutation statistics" file, with the protein GenBank identifier, the protein sequence and the PubMed identifier. This file is used during the final mutation mapping. An activity diagram showing the steps done to obtain the structure sequence is shown in Figure 3.11.

The algorithm for this module is:

Name of algorithm: Structure selection

Purpose: To select the most appropriate structure for the set of sequences on which the mutations would be mapped

Assumptions: The consensus sequence, BLAST, Formatdb, PDB are available.

Description of Inputs: The consensus sequence, the Formatdb formatted sequence file

Description of Outputs: A file having the selected structure sequence and the mutation statistics file.

Major Variables:

TEMP is a temporary file for storing intermediate results
STRUC is the file where the selected structure sequence is stored. It is one of the output files.
ALIGN is an alignment file in Fasta format
ALIGN2 is an alignment file in XML format
String str=PDB structure name
STAT is a file that stores the gi,sequence and pmid of the sequences. It is one of the output files.
db1 is a new database created using Formatdb from the sequence FASTA file.

Steps:

Using BLAST, align the consensus sequence with PDB database.

Write result to TEMP.

str = the structure name of the top hit of the BLAST output, present in TEMP.

Use EFetch with DB=PDB and use str as query to obtain the structure sequence

Write structure sequence to STRUC.

Format the sequence FASTA file, using Formatdb and create a database called db1.

Using BLAST, align the structure sequence with the input sequences present in db1.

Write the alignment in Fasta format to ALIGN and in XML format to ALIGN2

//The following step writes the statistics file
Using the ALIGN file, write the GenBank identifiers, the sequences from sequence XML file, and PubMed identifiers from mutation file for each sequence in the alignment, into the STAT file.

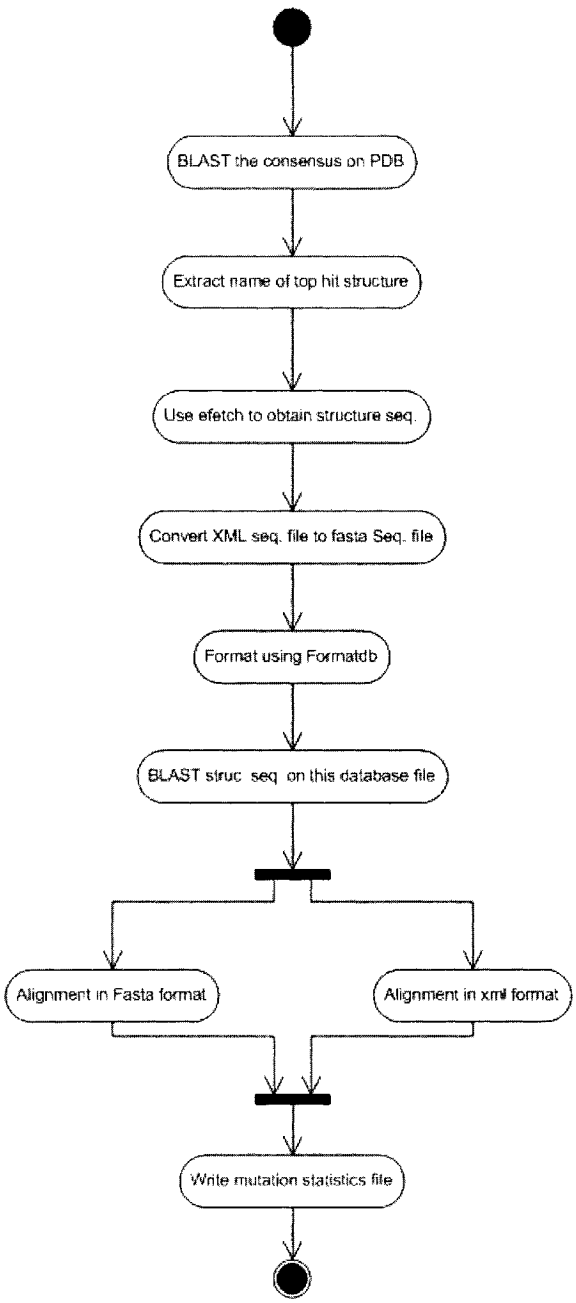

Figure 3.11 Activity Diagram Showing the Procedure Used to Select the Structure Sequence

3.5 Mapping the Mutations on the Selected Structure

One main problem while mapping the mutations to the structure was that the mutation co-ordinates obtained from the paper did not correspond to the actual positions of the residues in the sequence obtained from NCBI. For example a paper would state a mutation as N10D. However, when the actual sequence was analysed, there was no N at position 10. This was occurring frequently. The reason for this is probably due to the fact that the authors of the papers used a sequence that was slightly different from the one obtained from NCBI. The slight difference could be due to clipping of some residues at the ends of the sequences.

A new technique was devised to overcome this problem. Instead of individually taking each mutation and mapping it to the structure, the complete set of mutations per paper per protein was taken and distances between their coordinates were used to search the mutated residues on the sequences. For example, if a set of mutations are N10D, Y23L and K45R, then the distance between N and Y is 13 residues. Likewise, the distance between the Y and K is 22 residues. So the sequence is scanned repeatedly for N and Y with distance of 13 and Y and K with distance of 22. This technique works well, whatever be the number of mutations in the set as observed in the results given in Chapter 4.

The mutation statistics file helps to find all sets of mutations for all proteins in the sequence file, using this procedure. When such a set is found on the sequence, for each residue in the set, the corresponding residue and its position are found on the structure sequence (that was selected earlier) using the file having the alignment of the sequences with the structure sequences. An example follows.

Consider a sample alignment of structure sequence and input protein sequence as shown in Figure 3.12. Let the mutations be M10D, H27N. The distance between them is 17 and

the the M and H is found at actual positions 22 and 39 respectively. Once this is done, the alignment file is analysed to find the actual positions on the structure. In the example of Figure 3.12, the actual positions on structure that correspond to M and H are 21 and 36 respectively.

```
Query: 7    PFGEKKFIEIKGRRMAYIDEGT--GDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLI 64
            PF +  ++E+ G RM Y+D G   G P+LF HGNPTSSYLWRNI+PH A   R IA DLI
Sbjct: 9    PF-DPHYVEVLGERMHYVDVGPRDGTPVLFLHGNPTSSYLWRNIIPHVAPSHRCIAPDLI 67
```

Figure 3.12 Sample Alignment of a Part of Input Sequence (Sbjct) with the Structure Sequence (Query)

Confidence Scores: Once the mapping of mutations is done, a confidence score is calculated to assess the quality of the mapping and it is done as follows — the alignment file in fasta format has been used for the mapping of mutations but the alignment file in xml format is used for calculating the confidence score. In the alignment, if there is a perfect match in the residue at the point of mutation, in both the sequence and structure, then a score of 2 is given for that position. If they are similar, but not identical, then a score of 1 is given. This is seen in the alignment by a + between the lines of sequence and structure. If they are completely different, as indicated by a blank, then that position receives a 0. Such a score is calculated for 5 positions on either side of a point of mutation and added. This sum is divided by 22 (which is the maximum score for the point of mutation and 5 positions around it) , which gives us the confidence score in terms of a percentage. Hence the units of a confidence score is a percent.

All this information i.e the GenBank identifier, the PubMed identifier, the context of each

mutation, the actual position on structure, the confidence score are all written using specific predefined tags into a visualization file, which is in XML format. The activity diagram for this whole procedure is shown in Figure 3.13. The algorithm for this module is:

Name of algorithm: Visualization file creation

Purpose: To use the alignment of structure with sequences and map the mutations on the structure and create a final visualization file.

Assumptions: The alignment file having the alignment of file with the structure is available

Description of Inputs: The mutation statistics file — STAT, the mutation file and the alignment file having the structure-sequence alignments

Description of Outputs: The visualization file in XML format.

Major Variables:

InputArray - a six dimensional perl array for storing input mutation details. The first dimension will have the residue that was mutated. The second will have the PubMed identifier of the paper, the third will have the distance between this residue and the next in the set of mutations, The fourth will have the residue to which this residue was mutated. The fifth will have the context information and the sixth will have the position of the mutated residue.

Outputarray - a six dimensional perl array for storing mapped mutation details. The description of the dimensions is the same as in InputArray

int len = sequence length in terms of the number of aminoacids in the sequence

int count=0 This is to keep a count of the number of mutations found given a set of mutations for a particular protein in a particular paper.

int N = total number of mutated residues per sequence

Steps:

```
for(each entry in the statistics file STAT)
{
   Parse mutation file
   {
   Split each mutation into its components  residues, the position, distance between this
   position and that of next mutation, the context information and store in input array.
   for (each amino acid of the sequence)
   {
      Find if it is a mutated resdiue and if the next mutated residue in mutation file can be
      found on the sequence.
      if(amino acid is a mutated residue) count=count+1;
   }
   If (count=N-1)
   {
      Find equivalent residues on the structure.
      For each mutation
      {
         calculate confidence score.
         write the context information to the visualization file and also the position and
         equivalent residue on the structure sequence.
      }
   }
   }
}
```

Write the closing tags for the tags opened earlier, in the visualization file.

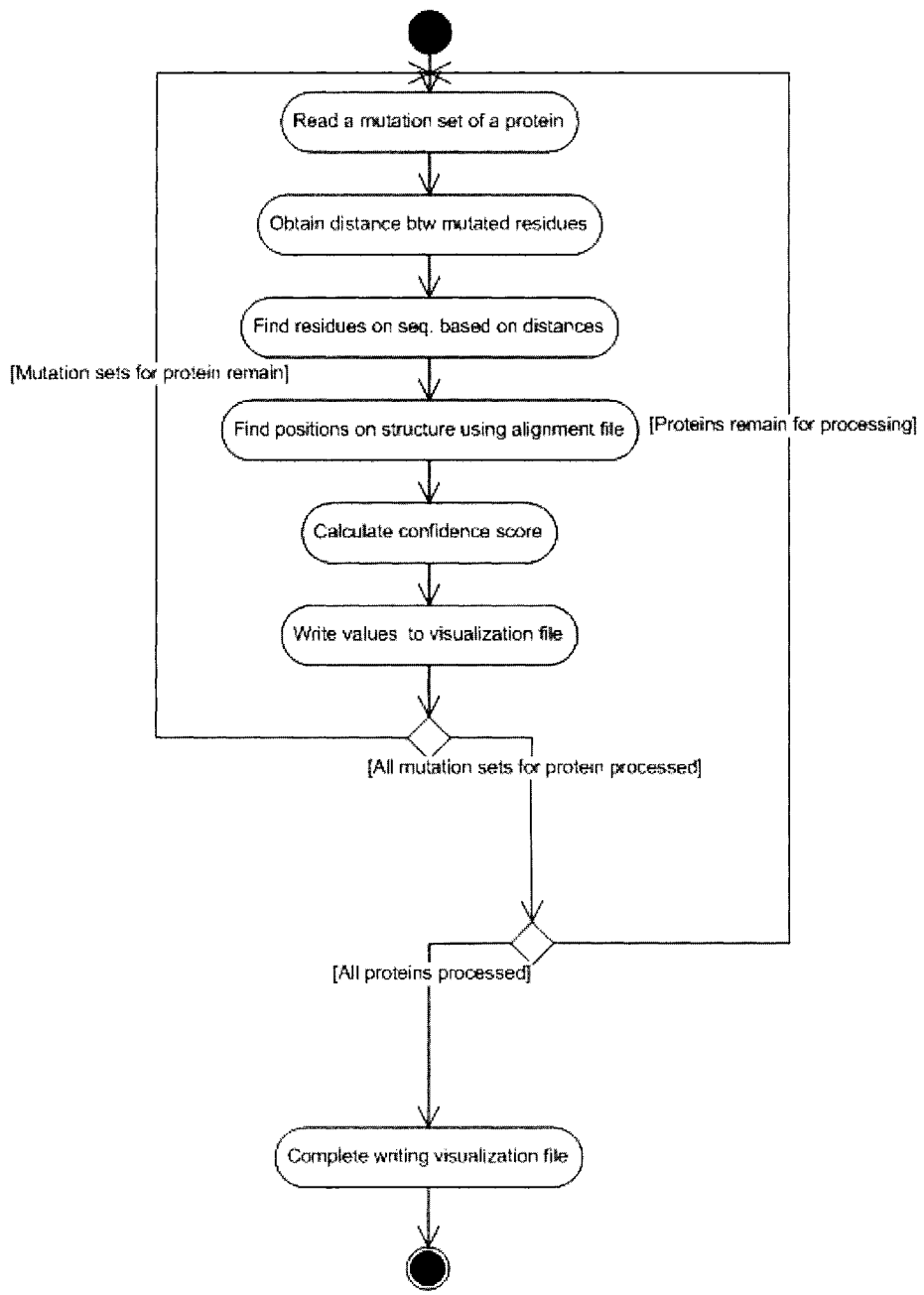

Figure 3.13 The Activity Diagram Showing the Procedure Used to Obtain the Final Visualization File

For the final visualization, a separate interface[GHLW04] in PROSAT protein visualization tool has been created by the European Media Laboratory, Germany, which enables the uploading of the visualization files generated by this tool along with the protein structure file (structure selected by the tool for mapping) obtained from PDB database. A sample of such a visualization is given in Figure 3.14. To the left of the screen, the 3D strcutre of the Chain A and Chain B of the 1EDE structure is seen. The highlights on the structure are the mutations mapped to it. In this case, it is the xylanase family mutations, mapped to Chain A of the structure. To the right of the screen, is a list of mutations that are being visualized. The two numbers separated by the ":" are the PubMed identifier and GenBank identifers respectively. For example 8019418:139865 means the mutations are for PubMed identifer 8019418 and for GenBank identifier 139865. The mutations are shown in the format $E->D, 100$. In this case, residue E is mutated to residue D with a 100% confidence score.

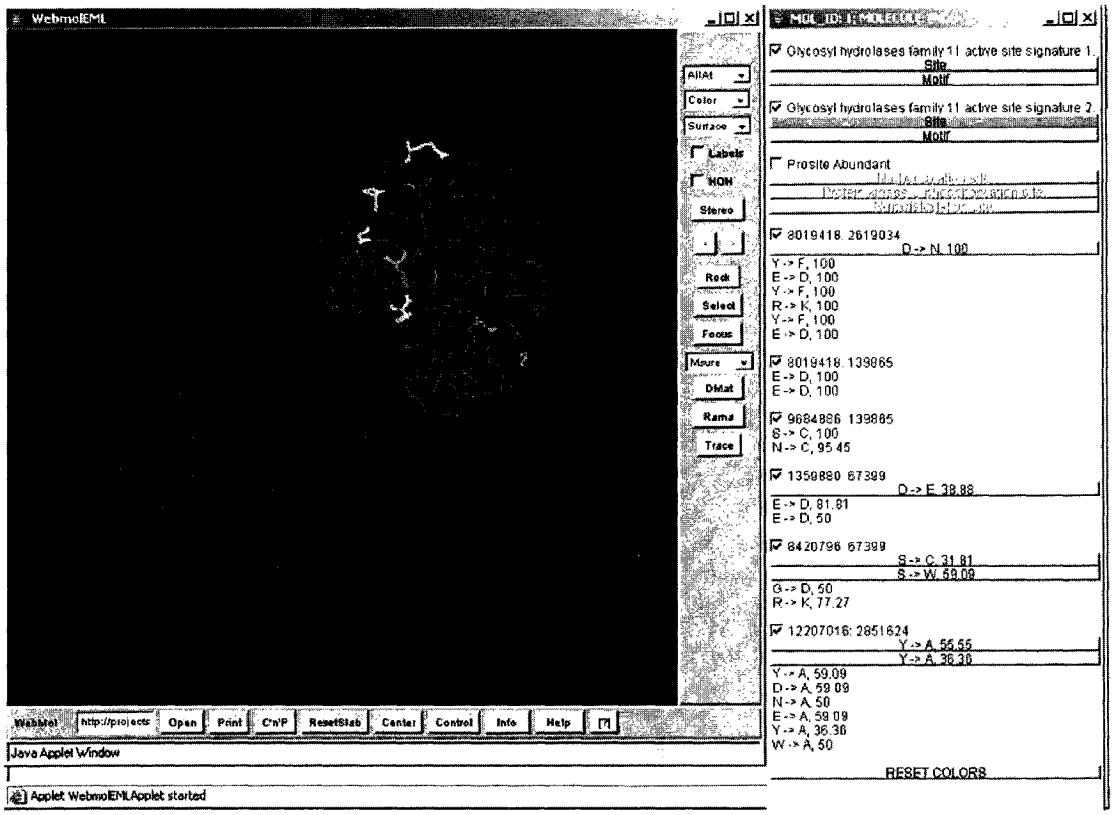

Figure 3.14 The Final Visualization in PROSAT Tool

Chapter 4 Test Results

This tool has been tested on three families of proteins xylanases, dehalogenases, and biphenyl dioxygenases. All of these enzymes have important industrial applications as described in Section 2.5. The meaning of a "Confidence score" and the procedure to calculate it is described in Section 3.5. The test results are provided in this order of the families – dehalogenase, xylanase and biphenyl dioxygenase.

Whether the mutations XpY of each protein in the protein sequences P1..Pn have been mapped to the right position on the selected sequence S, has to be verified manually by a scientist on a case-by-case basis. The results provided here have not been verified by a scientist. Hence we are not aware of whether the results provided here are correct or not. There is no benchmark with which the results can be compared. However, the validation of the results was done by checking the results of the tool at each step. For example, in the structure selection from consensus step, the E-values of the structures chosen were noted to check if the most similar structure has been chosen. Also, to validate the mapping, the alignment file of structure with sequences was checked for each mutation.

4.1 Dehalogenase Test Results

Aim of the test: The aim of this test is to detect how many mutations map to the structure, from a set of input mutations extracted from the papers.

Inputs: Two XML files are input. One file has the sequence details, called the sequence file. The other one has mutations for these sequences and is called the mutation file and this file is shown in appendix A.

Description of the Inputs: It is an XML file referred to as the mutation file in this thesis.

The *gi* tag value is the GenBank identifer of the protein sequence. The *pmid* refers to the PubMed identifier of the paper being discussed. The *mutation* tag has the value of the mutation referred to in this paper for a particular protein. The *context* tag provides more details about the mutation.

Expected output: All the input mutations are expected to map (100% mapping) but the results are usually lower than this value due to the reasons discussed in section 4.4

Actual Output: The actual output is a final visualization file and is shown in Appendix B

Description of the Output: This is the file that can be directly uploaded onto the PROSAT tool. The *label* tag indicates the PubMed identifer and the GenBank identifer. There are two *link* tags. The first one is the link to the paper and the second one is the link to the protein being discussed. The *range* tag indicates the position on the structure where this mutated residue occurs. If it is 155:A, then it means in chain A, the 155th residue is mutated. The *label* tag indicates more details on the mutation. For example if the mutation is

```
A F172 ->L  40.0
```

then it means that residue F at position 172 is mutated to L and the corresponding residue on the structure is A and the confidence score is 40%.

Procedure and results: The sequence file has details of three sequences of the dehalogenase family, whose GenBank identifiers are 61222635, 59799356 and 279969. There were no conserved domains found for these sequences. This testing was independent of the NLP component and so the sequence file and mutation files were prepared for providing the input to the tool. In the first trial, the Alistat threshold was set to 20%. At this threshold, the tool selected 1BN6 chain A as the structure file for mapping. The results are in Table 1. So 84% of the mutations were mapped to the structure as this threshold. The confidence scores

are distributed as shown in Figure 4.1. In the next two trials, the threshold was changed to 30% and 40% and the tool selected 1MJ5A as the structure. At this threshold, 62% of the mutations were mapped.

The thresholds could not be increased further since the three sequences had a percent pairwise identity of about 48%, according to Alistat statistics of their alignment.

At 30% threshold, the sequence with GenBank identifier 279969 gets removed, but this sequence had maximum mutations in the input file. The remaining two sequences generate a consensus that find 1MJ5 as their best structure (different from the one at 20%).

At threshold values 20%, 30% and 40%, the GenBank identifier 61222635 has all mutations mapped. The PubMed identifier values 14525993 and 12676719 are the only two values that appear for the protein sequence with GenBank identifier 59799356. This means that for this particular sequence, PubMed identifier values 12450392 and 10100638 do not appear in any of the thresholds. This is the reason for a greater loss of mutations and reduction in percentage of mutations at 30% and 40% threshold values.

The three trials were repeated for three other structures of the dehalogenase family. Their PDB identifier are 1CQW, 1CV2 and 1EDE. For 1CQW, the test results were identical to the earlier table, in terms of the number of mutations that mapped at various threshold values. In the case of the strcutre with PDB identifier 1CV2, at 20% threshold, about 79% of mutations were mapped to the strucure. Figure 4.2 shows the distribution of the confidence scores for 1CV2. For 1EDE, there was a small increase in the number of mutations that were mapped as is seen from Table 3. The confidence scores were also much better than for the structure selected by the tool and better than the values obtained for 1CQW and 1CV2. The confidence scores for 1EDE are shown in Figure 4.3.

Analysis of the results: The results show the number of input mutations and number of output mutations and this percentage. Since the expected is 100%, any value that is high or close to 100% is a good result and any value that is closer to 0% is a bad result.

In the 1EDE confidence score distribution graph, the number of occurrences of higher scores is larger compared to other structures, which indicates a better mapping.

The confidence score maps, shown in following figures, have the ability to tell us whether the overall mapping was good or bad but this does not tell us whether a particular mapping was useful to the user or not. For example, in Figure 4.2, the number of mutations with 20% exceed the number of mutations mapped with 60%. If a user is interested only in those mutations only that mapped in the 60% range, then this is a very good mapping for him. If the user is studying the mutations which have appeared at the 20% range, he would not find this particular mapping useful. He would then have to eliminate some of the sequences and then use the tool again to get a better mapping for the mutations that are of interest to him. For example, the confidence score chart of Figure 4.1 for 1CV2 structure shows that a large number of mutations were having confidence scores above 50%

Conclusion: If we consider that the result is good if the tool maps more than 50% of the input mutations, then the tool produced good results in this case. It was also observed that the default structure may not be the best possible selection since 1EDE gave better confidence scores than the default structure 1BN6. The reasons for non-mapping of mutations, is given in Section 4.4

Sensitivity of the tool: For this family of proteins, when the threshold was increased from 20% to 30%, the percentage of mutations mapped reduced from 84% to 62%. When the threshold is increased by 2 to 5 percentages, there is no change since a change occurs only

when sequences are eliminated and this is also the reason why the percentage mapped does not change between 30% to 40% thresholds. Hence sensitivity depends on the number and family of sequences that are being tested.

Alignment Threshold	Sequences Remaining	Mutations at Start	Multiple-mutations after alignment	Mutations in output file	Percentage of Mutations Mapped	Structure sequence
40	2	119	45	28	62.2	1MJ5A
30	2	119	45	28	62.2	1MJ5A
20	3	119	119	100	84.03	1BN6A

Table 1 Test Results When No Specific Structure Was Specified By the User (Structure Selection By Tool Itself)

Alignment Threshold	Sequences Remaining	Mutations at Start	Multiple-mutations after alignment	Mutations in output file	Percentage of Mutations Mapped	Structure sequence
40	2	119	45	28	62.2	1CQW
30	2	119	45	28	62.2	1CQW
20	3	119	119	100	84.03	1CQW

Table 2 Test Results When 1CQW Structure Was Specified By the User

Alignment Threshold	Sequences Remaining	Mutations at Start	Multiple-mutations after alignment	Mutations in output file	Percentage of Mutations Mapped	Structure sequence
40	2	119	45	28	62.2	1CV2
30	2	119	45	28	62.2	1CV2
20	3	119	119	94	79	1CV2

Table 3 Test Results When 1CV2 Structure Was Specified By the User

Alignment Threshold	Sequences Remaining	Mutations at Start	Multiple-mutations after alignment	Mutations in output file	Percentage of Mutations Mapped	Structure sequence
40	2	119	45	28	62.2	1EDE
30	2	119	45	28	62.2	1EDE
20	3	119	119	101	84.87	1EDE

Table 4 Test Results When 1EDE Structure Was Specified By the User

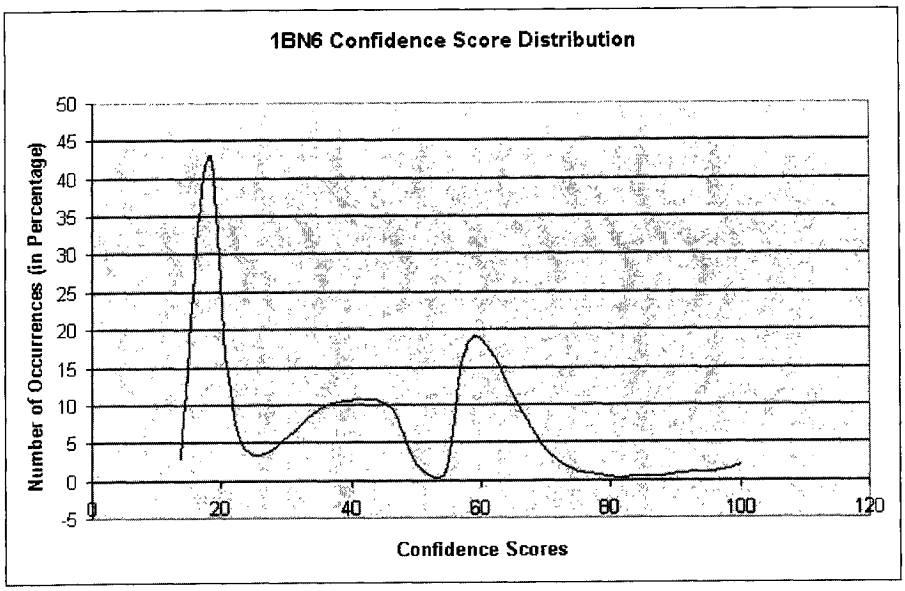

Figure 4.1 Confidence Scores For the 1BN6 Structure For Dehalogenase Mutations at 20% Threshold

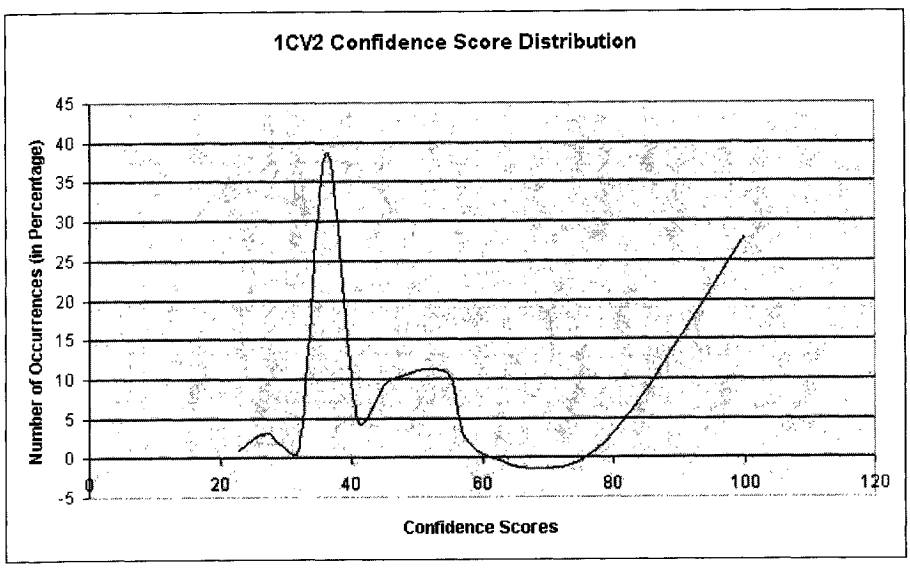

Figure 4.2 Confidence Scores For the 1CV2 Structure For Dehalogenase Mutations at 20% Threshold

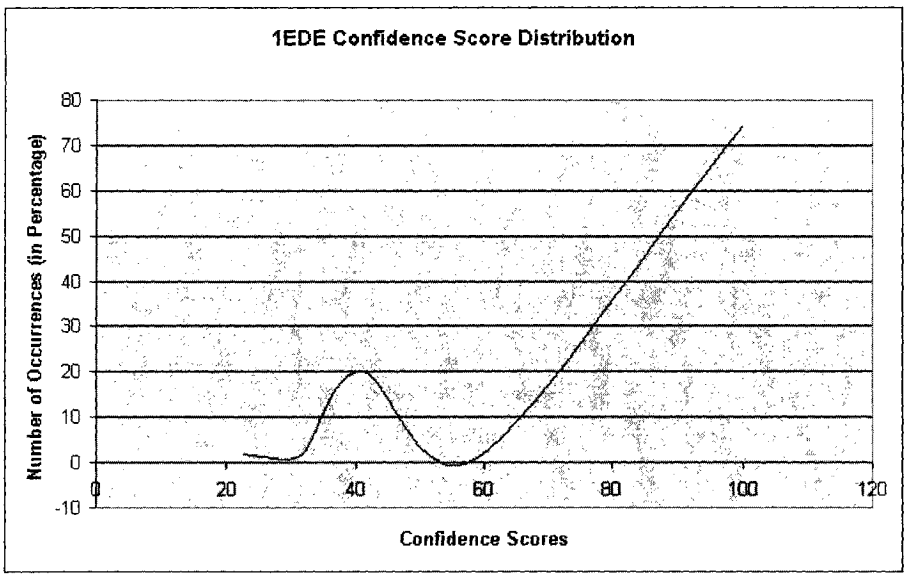

Figure 4.3 Confidence Scores For the 1EDE Structure For Dehalogenase Mutations at 20% Threshold

4.2 Xylanase Test Results

The test results for the xylanase family, another protein family of interest is described.

Aim of the test: The aim is the same as the previous test, to see how many mutations from an input file of mutations for the xylanase family map to the structure.

Input: Two input files were used - one having the sequences (sequence file), the other with the mutations (mutation file). The mutation file is shown in Appendix C.

Expected output: All the input mutations are expected to map (100% mapping) to the structure sequence.

Actual Output: The actual output is a final visualization file and is shown in Appendix D

Procedure and results: The sequence file had 19 protein sequences of the xylanase family.

The Conserved domain extraction module sliced these and produced 33 sequences, which formed the new sequence file. The Alistat threshold was varied from 20% to 90%. It was found that the number of mutations that were displayed increased as the threshold was reduced. It was maximum at 20%. Table 5 shows the list of 19 protein sequence gi numbers from the NCBI protein database, along with details of which ones are retained by Alistat at each threshold. The second column indicates the total number of mutations the sequence had in the beginning. Numbers in the cells corresponding to the various thresholds indicate the number of mutations that appeared in the final visualization file. Cells that are empty means that the sequence was not in the input sequences for that threshold after alignment. Cells that have a 0 means that 0 mutations was seen in the visualization for the sequence though such sequences were in the input sequences after the alignment. The details of each trial are shown in Table 6. A graph plotting the frequency of each confidence score (at 20% threshold) is in Figure 4.4.

The confidence score graph shows that this mapping was a good one. This can be found because most of the mutations have a confidence score of more than 50%. The assumption here is that a confidence score of 50% or above indicates a good map, though the exact value is dependent on the requirements of the biologist who uses the tool.

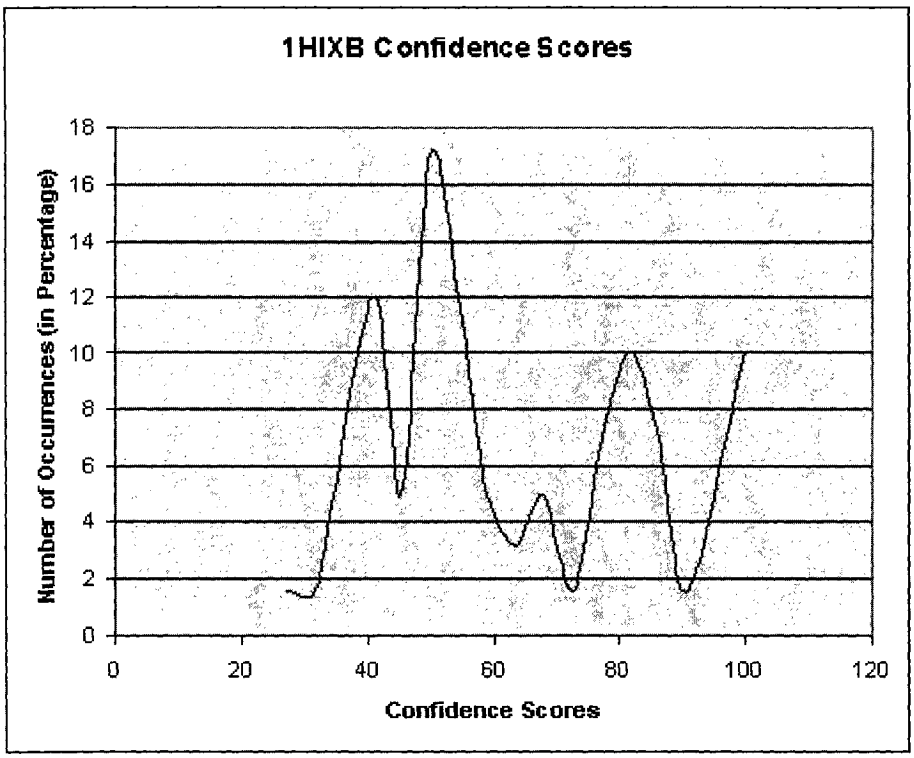

Figure 4.4 Confidence Scores For the 1HIXB Structure For Xylanase Mutations at 20% Threshold

Sequence GI number	No. of Mutations	Thresholds for alistat							
		20%	30%	40%	50%	60%	70%	80%	90%
67399	13	13	13	13	13	13	13	13	12
77715	14								
121856	7								
139865	9	4	4	4	4	4	4	4	4
139886	1								
538957	6								
2506385	5								
2619034	11	11	11	11	11	11	11	11	11
2851624	10	8	8	8	8	8	10	10	10
4514624	1								
5381269	4	4	4	4	4	4	4	4	
6226911	33								
6434133	6	6	6	6	6	6	6		
17942986	7	7	7	7	7	7			
21465565	1	0	0	0	0	0	0	0	
1351447	6								
465492	1	0	0	0	0	0	0	0	0
549461	19	18	18	18	18	18	19		
640242	2	0	0	0	0	0	0	0	0

Table 5 Sequence Details For the Xylanase Protein Family

Alignment Threshold	Sequences Remaining	Mutations at Start	Multiple-mutations after alignment	Mutations in output file	Percentage of Mutations Mapped	Structure sequences
90	6	143	45	36	80	1AXKB
80	8	143	50	41	82	1H4GB
70	10	143	64	55	85.9	1XND
60	11	143	70	58	82.8	1HIXB
50	11	143	70	58	82.8	1HIXB
40	11	143	70	58	82.8	1HIXB
30	11	143	70	58	82.8	1HIXB
20	11	143	70	58	82.8	1HIXB

Table 6 Output Results For Xylanase Family

4.3 Biphenyl Dioxygenase Test Results

The test results for the biphenyl dioxygenase family, another protein family of interest follows.

Aim of the test: The aim is the same as the previous tests, to see how many mutations from an input file of mutations for the biphenyl dioxygenase family map to the structure.

Input: The input file is in Appendix E.

Expected output: All the input mutations are expected to map (100% mapping) to the structure sequence.

Actual Output: The actual output is a final visualization file and is shown in Appendix F

Procedure and results: Two sequences with GenBank identifiers 151084 and 3023415 were chosen for the testing. Alistat reported them as having a percent pairwise identity of 95%. Hence for all alistat thresholds between 20% and 90%, the same percentage of mutations mapped and are shown in Table 7. The structure selected by the tool was 1O7H - chain A. Almost all mutations except one deletion mutation were mapped to the selected structure. A graph showing the confidence scores obtained is shown in Figure 4.5.

Alignment Threshold	Sequences Remaining	Mutations at Start	Multiple-mutations after alignment	Mutations in output file	Percentage of Mutations Mapped	Structure sequence
20% To 90%	2	48	48	47	98	1O7HA

Table 7 Biphenyl Dioxygenase Test Results

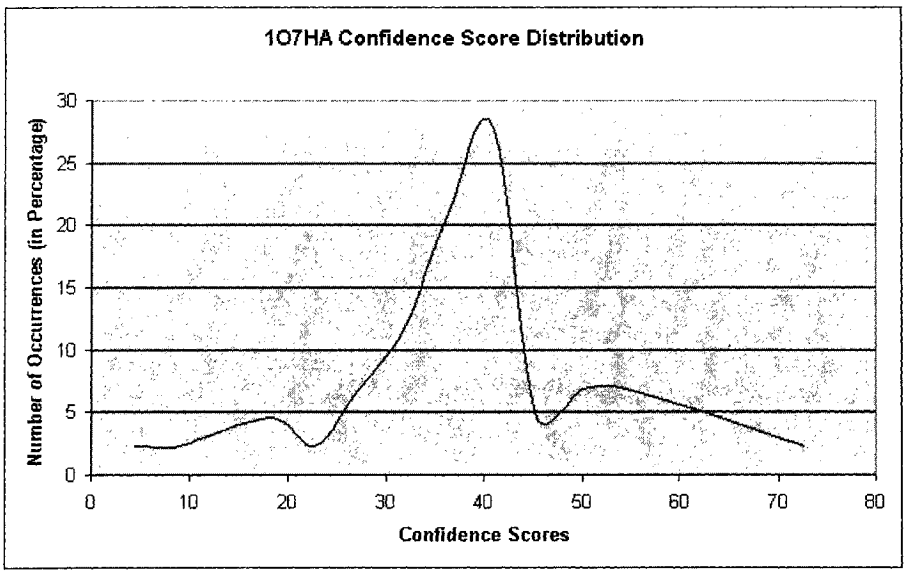

Figure 4.5 Confidence Scores For 1O7HA Structure for Biphenyl Dioxygenase Mutations at 20% Threshold

Analysis of the results: The results shows that about 98% of the mutations from the input file mapped to the structure. This is a good result because it is very close to the expected 100%, but the confidence score chart shows that most of the mappings had a score of about 40%. This means that if 50% confidence scores are required, then this mapping is not a good one. This happens because there is no structure detected for this consensus sequence that could provide a better mapping.

The tables shown in the earlier sections show the number of mutations that were input to the tool and the number of mutations that finally showed up in the visualization file. These numbers indicate how the tool responded to a particular family that was being tested but these results cannot be generalized and hence vary on a family to family basis. The aim

of the tool is to maximize the number of mutations that it can pick from the input and map to an automatically chosen structure. Hence the percentage of mutations mapped would be an ideal measure of how the tool is performing for a particular family. For example in the biphenyl dioxygenases test case, 47 out of 48 mutations mapped. This is a good mapping.

Chapter 5 Conclusions and Future Work

The issue to begin with was that PMD database has a huge backlog of unprocessed papers and hence access to latest mutation data was difficult. In order to address this issue, a tool was developed that has the capability to process the papers and extract mutations and finally map them on to the structure. In all the three enzyme families on which this tool was tested, a high percentage of mutations could be mapped to the structure. At 20% threshold, for the xylanase test case, 83% of mutations could be mapped. For the biphenyl dioxygenase case, 98% of mutations could be mapped. For the dehalogenase test case, 84% of mutations could be mapped. The tool was also successful in choosing the most appropriate structure for the given set of sequences. The display not only has the mutations mapped to structure but also the two URLs: one links to the paper in which the mutations are being discussed and the other one links to the sequence that is being discussed. The display also gives confidence scores for each mapped mutation, which enables a biologist to decide whether the mutation can be used for his purpose or it can be discarded. This means that the tool is ready for use for protein engineering and research.

The tool makes an assumption that the file from NLP analysis is consistent and correct but the NLP techniques are still being improved. It also uses the NCBI formats and depends on the availability of its databases. The tool uses the following information from the NLP analysis: the protein name, the organism name, the context, and mutation information. Using this information, it extracts the sequence information, conserved domain information from NCBI databases. Using this, it reformats the sequences by slicing them into the conserved domains and arranges them sequentially. Then this file is aligned using

a popular global alignment algorithm ClustalW and then using Alistat, sequences below a specified threshold are removed one by one. A consensus is generated for the final set of sequences using HMMER tools and the best possible structure for this consensus is found using BLAST tool. After a structure is found, the mutations are mapped onto the structure and the positions are found and written to a special output XML file. To assess the quality of the mapping, in a region of five aminoacids, on either sides of the mutated residue, a score is calculated for the quality of the alignment in that region. This is the confidence score, that allows the user to either trust the mapping or discard the result. All this information is written to the final visualization file, which can be uploaded along with the PDB structure file corresponding to the selected structure, in a special interface in PROSAT too, for the visualization

A low confidence score can occur due to the following reasons:

1. No structure was found that gave a good alignment with the consensus sequence. So, the best available is used.

2. A good consensus sequence for the set of sequences could not be found due to large non-overlapping regions in the sequences. This generally happens at lower alistat thresholds since non-overlapping sequences are not fully removed at lower thresholds.

There are certain areas where improvement is also possible such as better NLP analysis so that more accurate data can be obtained, a better method to obtain the consensus sequence so that more mutations can be mapped. Another significant improvement would be to use the Distributed, Environment-Centered Agent Framework (DECAF)[DGM03] to develop agents, which can perform various tasks. Our initial experiments with DECAF agents for information extraction were successful. Intelligent algorithms can be embedded in an agent

to perform various special tasks like periodic update of the locally stored PDB database, used by the tool. This would improve the performance of the tool.

For example, an agent can be programmed to choose another data source if one source fails or the required data cannot be obtained from a particular source. It is also possible to program an agent to make decisions on the choice of the data source to be used and hence a decision of the e-utilities to be used depending on network traffic, size of the response for a query, etc. DECAF agent framework can be easily used to create such agents and the agents can be given control of the Perl modules of this tool. An example of such a system, that uses agents is BIOMAS[DKM+02]. Other improvements are also possible.

The NLP output file currently provides an organism name and a protein name for identifying the protein sequence but this information is not accurate enough to fetch the right sequence that is being spoken about in the file. More information needs to be provided such as an accession sequence or the gi number itself using other information from the PubMed papers. This would improve the mutation mapping capability of the tool.

Further work on the NLP component is required to eliminate zip codes and false mutations like H2O. Also, the context information that is extracted from the papers needs to be filtered and special characters and other irrelevant characters need to be removed. It is also preferable to include some inference information too using knowledge bases.

Our strategy for dealing with single mutation was as follows if for example a single mutation N20C occurs, then a check is made if an N occurs at 20. If so, it is considered as the mutation that is being discussed about and a confidence score is calculated. This mutation is not mapped if N is not found at position 20.

The multiple mutations problem is solved by using the first set of mutations that occur on

the sequence. The displayed context information enables the user to decide if the mapping is to be used or not.

Another possibility is to use more than just one structure to map the mutations, i.e., if a number of structures are very similar to the consensus but not identical, then mutations can be mapped on all of them and the confidence scores displayed, so that the user can make a choice of which one he would want to retain.

Currently, the tool can map only mutations that are substitutions one residue being replaced by another. Future versions of the tool would also include capabilities to handle deletions of residues in sequences and insertions of residues in sequences.

5.1 Limitations of the Tool

In this section, we discuss the limitations of the tool. They relate to the inaccuracies of the NLP processing and the problem of identifying the correct position of a mutation in the sequence obtained from NCBI. A final limitation is that we only deal with mutations by substitution.

Note that we have not verified the correctness of any of our mapping results.

The Natural Processing techniques used in this tool need some improvement for producing more accurate data. Currently, the following open problems exist in the NLP analysis part:

Insufficient data: The NLP output file provides the organism name and protein name to identify the protein, on which it lists the mutations, but this data is sometimes not sufficient to identify the exact sequence for mapping the mutations. Due to this, many mutations cannot be mapped due to lack of information for getting the right sequence.

False mutations: In a zip code occurring in the paper like H3T 1E8, the H3T is reported as a mutation. Similarly H2O, which is water is reported as a mutation.

By improved NLP techniques, these problems can be solved in the future versions of this tool. Some other problems in the sequence processing part of the tool are as follows:

Single Mutations: Single mutations means that a paper talks about a single mutation on a protein like N23D. Mapping such mutations on the structure is a problem because in most cases the mutated residue is not found at the specified position i.e. in this example, N is not found at position 23. Since there is no way to ascertain the right mutated residue, the single mutation is mapped only if the residue is found at the exact position specified.

Multiple mutations: Since distances are used to find mutations on the sequences and not the positions, in some cases there exist more than one set of residues that have the same distances. In this case, the first set that occurs on the sequence is chosen.

Mutations not found: In the xylanase and dehalogenase test cases, some of the mutations obtained from the papers could not be found on the sequence obtained from NCBI. Due to this reason, some mutations cannot be displayed, but such sequences contribute to the consensus and this may affect the mapping of mutations of other sequences. For example in xylanase test case, 8% mutations did not map due to this reason at 20% threshold.

Deletions and additions: The current version of this tool cannot map deletions and additions of residues. It would be an interesting additional feature in the future versions of this tool.

Sources of Errors in The Tool:

In case the Organism name or Protein name in the NLP analysis is inaccurate, then there is a loss of mutations. For example if protein name is "xylanase" and organism name is

"bacillus circulans", then the tool attempts to find the sequence that is being discussed but since several possibilities exist, it is likely that there could be errors in choosing the right sequence. This problem can be overcome by searching the mutations on all possible sequences.

In the next stage, where conserved domains are found, it is possible that there could be some mutations outside the domains, although we never came across such a case. This means they cannot be found and due to the distance method of finding mutations, the complete set of mutations for that protein cannot be found. This is one source of error. Another could be that the conserved domains we obtain from NCBI may not be accurate enough and correspond to the ones used by the authors of the paper while they studied the mutations on the sequences. For example, in the following sequence :

NGNSYLTVYGWTVDP*LVEFYIVDSWGTPKGTINVD*GGTYQIYETTRYNQPSIKGTA

the italicized portion is a conserved domain and the blocked residues are mutations. If only the conserved domains are extracted, then there is a chance that the mutation "G" is lost and hence the whole set of mutations cannot be found.

Since distances between mutations are used to find the sequences, it is possible that the set of mutations we are using may not be the right one. There is no way to know since the numbers used by the authors are not standardized. For example, in the following sequence and mutation set M5D, S11F there are two sets of residues M and S with same distance of six.

D*M*WGTYR*S*TGAYKGSFYAD*M*GTYDI*S*YETTRVNKQYWSVRQTKRT

We use a progressive global alignment algorithm (ClustalW) for aligning the sequences. This alignment algorithm does not guarantee an optimal alignment. Progressive global

alignments also have a problem that errors induced initially in the alignment remain in the alignment. Also alignment is unpredictable in case of non-overlapping sequences.

In the stage where sequences are eliminated based on the user specified threshold, it is not possible to state that a particular threshold would be appropriate to retain a family of sequences and remove the rest. It is likely that sequences may remain that do not belong to the family of interest. This would lead to errors in the consensus sequence and an inappropriate structure could be used to map the mutations, thus reducing the reliability of the output.

The next step is where a consensus sequence is generated. Although consensus sequence represents the set of sequences that was used to generate it, there may not be perfect match between the consensus sequence and any of the sequences. This means the structure selected tries to represent the group as a whole and not any particular sequence. This may reduce the confidence scores of some mutations that map to structures that do not represent the sequence accurately. For example, the structure 1XND aligns very well with GenBank identifier 6434133 but not with GenBank identifier 139865 in the alignment of Figure 5.1. This happens more frequently at lower alistat thresholds and rarely at higher alistat thresholds.

```
>gi 6434133
Query: 61  NFSGSYNPNGNSYLSIYGWSRNPLIEYYIVENFGTYNPSTGATKLGEVTSDGSVYDIYRT 120
           NFSGSYNPNGNSYLS+YGWS+NPLIEYYIVENFGTYNPSTG TKLGEVTSDGSVYDIYRT
Sbjct: 86  NFSGSYNPNGNSYLSVYGWSKNPLIEYYIVENFGTYNPSTGTTKLGEVTSDGSVYDIYRT 145

>gi 139865
Query: 6   GTGYSNGYYYSYWNDGHAGVTYTNGGGGSFTVNWSNSGNFVAGKGWQPGTKNKVINFSGS 65
           GT  S+G  Y  +          +G   +FT  WS        +  +P   N  I F+
Sbjct: 124 GTVKSDGGTYDIYTTTRYNAPSIDGDRTTFTQYWS------VRQSKRPTGSNATITFTNH 177
```

Figure 5.1 Sample Alignment of Structure With Sequences

Even after the consensus sequence is found, there is no guarantee that the best structure can be found for the given consensus sequence. The structure is found here using BLAST on the PDB database, and the accuracy of the structure selected depends on the availability of a closely matching structure for the consensus sequence. If no structure closely matches the consensus sequence, then BLAST reports what is available and this would affect the qualilty of the output.

After mapping the mutations, for calculating a confidence score, we use a window of eleven residues. This is a sample of the region surrounding the mutation in the alignment, but the confidence score changes if the window is increased or decreased. This means that this window size may not be the best choice to evaluate the alignment at a region and hence can introduce errors in the confidence score that is output. For example, consider the window of size 11 shown in Figure 5.2. The confidence score for the residue T here is 95.45%. If a window size of 15 is used, the confidence score becomes 83%.

```
Query: 121 QRVNQPSIIGTATFYQYWSVRRNHRSSGSVNTANHFNAWASHGLTLGTMDYQIVAVEGYF 180
           QRVNQPSIIGTATFYQYWSVRRNH  +         NAW + GLTLGT+DYQI+AVEGYF
Sbjct: 146 QRVNQPSIIGTATFYQYWSVRRNHAPAARSRLRTTSNAWRNLGLTLGTLDYQIIAVEGYF 205
```

Figure 5.2 Window Size For Calculating Confidence Scores

As discussed in Chapter 4, the correctness of the results produced by this tool have not been checked by a scientist or compared with results of other similar tools, but the results produced by the tool have been analysed for accuracy at each step.

Bibliography

[AD00] S. Abdeddaim and L. Duret. *Multiple alignments for structural, functional, or phylogenetic analyses of Homologous sequences*, chapter 3. Oxford University Press, 2000. This is a chapter of [HT00].

[BCJL04] Bernard F. Buxton, David P. A. Corney, David T. Jones, and William B. Langdon. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213, 2004.

[Bio96] USC Computational Biology. FASTA Format. *http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html*, 1996.

[Bio05a] Research Collaboratory For Structural Bioinformatics. Protein data bank. *http://www.rcsb.org/pdb/*, 2005.

[Bio05b] Swiss Institute Of Bioinformatics. PROSITE Database of Protein Families and Domains. *http://ca.expasy.org/prosite/*, 2005.

[BW04] Christopher J. O. Baker and Rene 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) / IBM CASCON*, 2004.

[BW05] Christopher J. O. Baker and Rene Witte. Combining Biological Databases and Text Mining to support New Bioinformatics Applications. In *10th International Conference on Applications of Natural Language to Information Systems, NLDB*, number 3513, pages 310–321. Springer, 2005.

[BWK+03] Sabine Bergler, Rene Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz. Using knowledge-poor coreference resolution for text summarization. In *Document Understanding Conferences*, pages 43–50. National Institute Of Science And Technology, 2003.

[Che94] Institute Of Chemistry. Amino Acids. *http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html*, 1994.

[Com05] Institute For Biomedical Computing. MSA. *http://softlib.rice.edu/msa.html*, 2005.

[Cor05] Worthington Biochemical Corporation. Enzymes and Life Processes. *http://www.worthington-biochem.com/introBiochem/lifeProcesses.html*, 2005.

[Cun02] H. Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254, 2002.

[DEKM98] R. Durbin, Sean R. Eddy, A. Krogh, and G. J. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[DGM03] Keith S. Decker, John R. Graham, and Michael Mersic. DECAF - A Flexible Multi Agent System Architecture. *Autonomous Agents and Multi-Agent Systems*, 7:7–27, 2003.

[DKM+02] Keith S. Decker, Salim Khan, Ravi Makkena, Dennis Michaud, Carl Schmidt, and Gang Situ. BioMAS: A Multi-Agent System for Genomic Annotation. *International Journal of Cooperative Information System*, 11:265–292, 2002.

[Doi05] A. J. Doig. Protein Engineering Introduction and Lecture Summaries. *http://www.bi.umist.ac.uk/users/mjfajdg/2PAB/default.asp*, 2005.

[Edd98] Sean R. Eddy. Profile Hidden Markov Models. *Bioinformatics*, 14:755–763, 1998.

[Edd05] Sean R. Eddy. HMMER: Sequence Analysis using Profile Hidden Markov Models. *http://hmmer.wustl.edu/*, 2005.

[Fou96] National Science Foundation. RasMol home page. *http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html*, 1996.

[GBA+02] L. Y Geer, S. H Bryant, Marchler-Bauer A, Panchenko A. R, B. A Shoemaker, and P. A Thiessen. CDD: a database of conserved domain alignments with links to domain three dimensional structure. *Nucleic Acids Research*, 30(1):281– 283, 2002.

[GHLW04] Razif R. Gabdoulline, Ren Hoffmann, Florian Leitner, and Rebecca C. Wade. ProSAT - Protein Structure Annotation Tool. *http://projects.villa-bosch.de/dbase/pdba2/*, 2004.

[GHT94] Toby J. Gibson, D. G. Higgins, and J. D. Thompson. CLUSTALW: 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.

[HLWG03] Ren Hoffmann, Florian Leitner, Rebecca C. Wade, and Razif R. Gabdoulline. ProSAT: functional annotation of protein 3D structure. *Bioinformatics*, 19(13):1723–1725, 2003.

[HT00] Des Higgins and Willie Taylor. *Bioinformatics: Sequence, Structure, and Databanks : A Practical Approach.* Oxford University Press, 2000.

[Inf05a] National Center For Biotechnology Information. Entrez Programming Utilities. *http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html*, 2005.

[Inf05b] National Center For Biotechnology Information. National Center For Biotechnology Information Homepage. *http://www.ncbi.nlm.nih.gov/*, 2005.

[Ins03] European Bioinformatics Institute. Macromolecular Structure Database. *http://pqs.ebi.ac.uk/*, 2003.

[Ins05] Sanger Institute. Pfam home page. *http://www.sanger.ac.uk/Software/Pfam/*, 2005.

[JRM+99] Y Jouanneau, D Rodarie, C Meyer, N Hugo, and D Tropel. Catalytic Role of Dioxygenases in the Biodegradation of Aromatic Hydrocarbons. *Proceedings of the Biotechnology in Environmental Protection BIODEPOL 99, France*, October 26–27 1999.

[JSP+00] Manish D. Joshi, Gary Sidhu, Isabelle Pot, Gary D. Brayer, Stephen G. Withers, and Lawrence P. McIntosh. Hydrogen Bonding and Catalysis: A Novel Explanation for How a Single Amino Acid Substitution Can Change the pH Optimum of a Glycosidase. *Journal Of Molecular Biology*, 299:255–279, 2000.

[KNO99] Takeshi Kawabata, Ken Nishikawa, and Motonori Ota. The protein mutant database. *Nucleic Acid Research*, 27(1):355–357, 1999.

[Kol97] Craig Koltes. Haloalkane Dehalogenase. *http://www.chem.uwec.edu/Chem406/Webpages97/craig/begin.htm*, 1997.

[LCS+04] I. Letunic, R.R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, C. P. Ponting, and P. Bork. SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, 32:142–144, January 2004.

[LGM+90] David J. Lipman, Warren Gish, Webb Miller, Stephen F. Altschul, and EugeneW Meyers. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–310, 1990.

[LP85] David J. Lipman and W. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

[Med05] National Library Of Medicine. PubMed overview. *http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html*, 2005.

[Mor99a] B. Morgenstern. DIALIGN 2: improvement of the segment–to–segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.

[Mor99b] B. Morgenstern. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.

[MS05] V. Meenakshi and M. C. Srinivasan. Microbial xylanases for paper industry. *Current Science*, 89:Online, 2005.

[NH96] Cedric Notredame and Desmond G. Higgins. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Research*, 24:1515–1524, 1996.

[NHH00] C. Notredame, D. Higgins, and J. Heringa. A novel method for multiple sequence alignments. *Journal of Molecular Biology*, 302:205–217, 2000.

[Not02] Cedric Notredame. Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics*, 3:131–144, 2002.

[Onl05a] Biology Online. Definition from Biology-online.org. *http://www.biology-online.org/dictionary/protein*, 2005.

[Onl05b] Biology Online. Definition from Biology-online.org. *http://www.biology-online.org/dictionary/consensus_sequence*, 2005.

[PPR+01] Frederic Plewniak, Oliver Poch, Raymond Ripp, Jean-Claude Thierry, and Julie D. Thompson. Towards a reliable objective function for multiple sequence alignments. *Journal of Molecular Biology*, 314:937–951, 2001.

[SW81] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[Sys05] Expert Protein Analysis System. UniProt - The Universal Protein Resource. *http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html*, 2005.

[TKL97] R.L. Tatusov, E.V. Koonin, and David J. Lipman. A genomic perspective on protein families. *Nucleic Acids Research*, 278:631–637, October 1997.

[Wal05] Matthew Wall. Introduction to Genetic Algorithms. *http://lancet.mit.edu/ mbwall/presentations/IntroToGAs/*, 2005.

[WCS96] Larry Wall, Tom Christiansen, and Randal L. Schwartz. *Programming Perl*. O'Reilly, 2nd edition, 1996.

[Wit04] Rene Witte. An Integration Architecture for User-Centric document creation, retrieval, and analysis. *Proceedings of the VLDB Workshop on Information Integration on the Web (IIWeb-2004), Toronto, Canada*, pages 141–144, August 2004.

Appendix A

Dehalogenase Test Input Mutation File

```
<?xml version="1.0"?>
<main>
<component>
<gi>61222635</gi>
<pmid>12089046</pmid>
<mutation>C176Y</mutation>
<context> ok </context>
<mutation>Y273F</mutation>
<context> ok </context>
</component>

<component>
<gi>59799356</gi>
<pmid>12450392</pmid>
<mutation>N38F</mutation>
<context> ok </context>
<mutation>W109L</mutation>
<context> ok </context>
<mutation>F151L</mutation>
<context> ok </context>
<mutation>F169L</mutation>
<context> ok </context>
<mutation>W175Y</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>8110757</pmid>
<mutation>D124A</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>7705355</pmid>
<mutation>W125Q</mutation>
<context> ok </context>
<mutation>W175Q</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>7828730</pmid>
<mutation>D124N</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>7737973</pmid>
<mutation>H289Q</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>8855957</pmid>
<mutation>F172W</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>9236003</pmid>
<mutation>N148D</mutation>
<context> ok </context>
<mutation>D260N</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>9051734</pmid>
<mutation>V226G</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>10099367</pmid>
<mutation>F164A</mutation>
<context> ok </context>
<mutation>D170A</mutation>
<context> ok </context>
<mutation>F172A</mutation>
<context> ok </context>
<mutation>W175A</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>9790663</pmid>
<mutation>W175Y</mutation>
<context> ok </context>
</component>

<component>
<gi>59799356</gi>
<pmid>10100638</pmid>
<mutation>D108N</mutation>
<context> ok </context>
<mutation>D124N</mutation>
<context> ok </context>
<mutation>E132Q</mutation>
<context> ok </context>
<mutation>E244Q</mutation>
<context> ok </context>
<mutation>H272A</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>10231528</pmid>
```

```
<mutation>W175Y</mutation>
<context> ok </context>
</component>

<component>
<gi>59799356</gi>
<pmid>14525993</pmid>
<mutation>L177A</mutation>
<context> ok </context>
</component>

<component>
<gi>59799356</gi>
<pmid>12676719</pmid>
<mutation>W109L</mutation>
<context> ok </context>
<mutation>F143W</mutation>
<context> ok </context>
<mutation>Q146A</mutation>
<context> ok </context>
<mutation>D147V</mutation>
<context> ok </context>
<mutation>L177A</mutation>
<context> ok </context>
<mutation>I211L</mutation>
<context> ok </context>
<mutation>L248I</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>11937643</pmid>
<mutation>F172L</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>11932489</pmid>
<mutation>D16C</mutation>
<context> ok </context>
<mutation>A201C</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>10963662</pmid>
<mutation>W125Q</mutation>
<context> ok </context>
<mutation>W175Y</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>10585505</pmid>
<mutation>N148D</mutation>
<context> ok </context>
<mutation>D260N</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>9862209</pmid>
<mutation>F172H</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>9579656</pmid>
<mutation>T175I</mutation>
<context> ok </context>
</component>

<component>
<gi>279969</gi>
<pmid>8021255</pmid>
<mutation>P168S</mutation>
<context> ok </context>
<mutation>D170H</mutation>
<context> ok </context>
</component>

</main>
```

Appendix B

Dehalogenase Test Output Visualization File

```
 <menue>
<status=on>
<label>PMID: 12089046, GI : 61222635</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=12089046</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=61222635</link>
<item>
<range>128:A 128:A</range>
<color=yellow>
<status=on>
<label>C C176 -> Y, 100.0</label>
<context> ok </context></item>
<item>
<range>225:A 225:A</range>
<color=yellow>
<status=on>
<label>Y Y273 -> F, 100.0</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 14525993, GI : 59799356</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=14525993</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=59799356</link>
<item>
<range>176:A 176:A</range>
<color=yellow>
<status=on>
<label>C L177 -> A, 59.09</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 12676719, GI : 59799356</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=12676719</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=59799356</link>
<item>
<range>112:A 112:A</range>
<color=yellow>
<status=on>
<label>G W109 -> L, 95.45</label>
<context> ok </context></item>
<item>
<range>141:A 141:A</range>
<color=yellow>
<status=on>
<label>W F143 -> W, 31.81</label>
<context> ok </context></item>
<item>
<range>144:A 144:A</range>
<color=yellow>
<status=on>
<label>F Q146 -> A, 36.36</label>
<context> ok </context></item>
<item>
<range>145:A 145:A</range>
<color=yellow>
<status=on>
<label>A D147 -> V, 36.36</label>
<context> ok </context></item>
<item>
<range>175:A 175:A</range>
<color=yellow>
<status=on>
<label>K L177 -> A, 59.09</label>
<context> ok </context></item>
<item>
<range>210:A 210:A</range>
<color=yellow>
<status=on>
<label>P I211 -> L, 36.36</label>
<context> ok </context></item>
<item>
<range>246:A 246:A</range>
<color=yellow>
<status=on>
<label>L L248 -> I, 50.0</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 8110757, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=8110757</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>109:A 109:A</range>
<color=yellow>
<status=on>
<label>S D124 -> A, 45.45</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 7705355, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=7705355</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>109:A 109:A</range>
<color=yellow>
<status=on>
<label>S W125 -> Q, 45.45</label>
<context> ok </context></item>
<item>
<range>157:A 157:A</range>
<color=yellow>
```

```
<status=on>
<label>V W175 -> Q, 18.18</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 7828730, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=7828730</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>109:A 109:A</range>
<color=yellow>
<status=on>
<label>S D124 -> N, 45.45</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 7737973, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=7737973</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>274:A 274:A</range>
<color=yellow>
<status=on>
<label>L H289 -> Q, 22.72</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 8855957, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=8855957</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>155:A 155:A</range>
<color=yellow>
<status=on>
<label>A F172 -> W, 13.63</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 9236003, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9236003</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>130:A 130:A</range>
<color=yellow>
<status=on>
<label>E N148 -> D, 36.36</label>
<context> ok </context></item>
<item>
<range>244:A 244:A</range>
<color=yellow>
<status=on>
<label>G D260 -> N, 13.63</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 9051734, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9051734</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>205:A 205:A</range>
<color=yellow>
<status=on>
<label>F V226 -> G, 27.27</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 10099367, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=10099367</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>146:A 146:A</range>
<color=yellow>
<status=on>
<label>R F164 -> A, 27.27</label>
<context> ok </context></item>
<item>
<range>152:A 152:A</range>
<color=yellow>
<status=on>
<label>F D170 -> A, 18.18</label>
<context> ok </context></item>
<item>
<range>154:A 154:A</range>
<color=yellow>
<status=on>
<label>T F172 -> A, 18.18</label>
<context> ok </context></item>
<item>
<range>157:A 157:A</range>
<color=yellow>
<status=on>
<label>V W175 -> A, 18.18</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 9790663, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9790663</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>158:A 158:A</range>
<color=yellow>
<status=on>
<label>G W175 -> Y, 18.18</label>
<context> ok </context></item>
```

```
</menue>
<menue>
<status=on>
<label>PMID: 10231528, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=10231528</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>158:A 158:A</range>
<color=yellow>
<status=on>
<label>G W175 -> Y, 18.18</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 11937643, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=11937643</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>155:A 155:A</range>
<color=yellow>
<status=on>
<label>A F172 -> L, 13.63</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 11932489, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=11932489</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>179:A 179:A</range>
<color=yellow>
<status=on>
<label>R D16 -> C, 54.54</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 10963662, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=10963662</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>109:A 109:A</range>
<color=yellow>
<status=on>
<label>S W125 -> Q, 45.45</label>
<context> ok </context></item>
<item>
<range>157:A 157:A</range>
<color=yellow>
<status=on>
<label>V W175 -> Y, 18.18</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 10585505, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=10585505</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>130:A 130:A</range>
<color=yellow>
<status=on>
<label>E N148 -> D, 36.36</label>
<context> ok </context></item>
<item>
<range>244:A 244:A</range>
<color=yellow>
<status=on>
<label>G D260 -> N, 13.63</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 9862209, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9862209</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>155:A 155:A</range>
<color=yellow>
<status=on>
<label>A F172 -> H, 13.63</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 8021255, GI : 279969</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=8021255</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=279969</link>
<item>
<range>75:A 75:A</range>
<color=yellow>
<status=on>
<label>P P168 -> S, 72.72</label>
<context> ok </context></item>
<item>
<range>77:A 77:A</range>
<color=yellow>
<status=on>
<label>L D170 -> H, 63.63</label>
<context> ok </context></item>
</menue>
```

Appendix C
Xylanase Test Input Mutation File

```
<?xml version="1.0"?>
<main>

<component>
<gi>67399</gi>
<pmid>1359880</pmid>
<mutation>D21E</mutation>
<context> ok </context>
<mutation>E93D</mutation>
<context> ok  </context>
<mutation>E182D</mutation>
<context> ok </context>
</component>

<component>
<gi>67399</gi>
<pmid>8420796</pmid>
<mutation>S12C</mutation>
<context> ok </context>
<mutation>S26W</mutation>
<context> ok </context>
<mutation>G38D</mutation>
<context> ok </context>
<mutation>R48K</mutation>
<context> ok </context>
<mutation>T126S</mutation>
<context> ok </context>
</component>

<component>
<gi>77715</gi>
<pmid>9006940</pmid>
<mutation>E43A</mutation>
<context> ok </context>
<mutation>N44A</mutation>
<context> ok </context>
<mutation>M46A</mutation>
<context> ok </context>
<mutation>K47A</mutation>
<context> ok </context>
<mutation>N126A</mutation>
<context> ok </context>
<mutation>E127G</mutation>
<context> ok </context>
<mutation>N182A</mutation>
<context> ok </context>
</component>

<component>
<gi>77715</gi>
<pmid>9211898</pmid>
<mutation>D256A</mutation>
<context> ok </context>
<mutation>N261A</mutation>
<context> ok </context>
<mutation>D262A</mutation>
<context> ok </context>
</component>

<component>
<gi>121856</gi>
<pmid>7910761</pmid>
<mutation>E127A</mutation>
<context> ok </context>
</component>

<component>
<gi>121856</gi>
<pmid>8855954</pmid>
<mutation>D123A</mutation>
<context> ok </context>
<mutation>E127A</mutation>
<context> ok </context>
<mutation>E233A</mutation>
<context> ok </context>
</component>

<component>
<gi>139865</gi>
<pmid>8019418</pmid>
<mutation>E78D</mutation>
<context> ok </context>
<mutation>E172D</mutation>
<context> ok </context>
</component>

<component>
<gi>139865</gi>
<pmid>9684886</pmid>
<mutation>S100C</mutation>
<context> ok </context>
<mutation>N148C</mutation>
<context> ok </context>
</component>

<component>
<gi>139865</gi>
<pmid>9047328</pmid>
<mutation>E172Q</mutation>
<context> ok </context>
</component>

<component>
<gi>139865</gi>
<pmid>10860737</pmid>
<mutation>N35D</mutation>
<context> ok </context>
</component>

<component>
<gi>139865</gi>
<pmid>10220321</pmid>
<mutation>Y69F</mutation>
<context> ok </context>
</component>

<component>
<gi>139886</gi>
<pmid>11601976</pmid>
<mutation>S172A</mutation>
```

```
<context> ok </context>
</component>

<component>
<gi>538957</gi>
<pmid>8376336</pmid>
<mutation>D537N</mutation>
<context> ok </context>
<mutation>E541Q</mutation>
<context> ok </context>
<mutation>H572N</mutation>
<context> ok </context>
<mutation>E600Q</mutation>
<context> ok </context>
<mutation>D602N</mutation>
<context> ok </context>
<mutation>D645N</mutation>
<context> ok </context>
</component>

<component>
<gi>2506385</gi>
<pmid>10767281</pmid>
<mutation>Y87A</mutation>
<context> ok </context>
<mutation>W313A</mutation>
<context> ok </context>
<mutation>L314A</mutation>
<context> ok </context>
</component>

<component>
<gi>2619034</gi>
<pmid>8019418</pmid>
<mutation>D11N</mutation>
<context> ok </context>
<mutation>Y69F</mutation>
<context> ok </context>
<mutation>E78D</mutation>
<context> ok </context>
<mutation>Y80F</mutation>
<context> ok </context>
<mutation>R112K</mutation>
<context> ok </context>
<mutation>Y166F</mutation>
<context> ok </context>
<mutation>E172D</mutation>
<context> ok </context>
</component>

<component>
<gi>2851624</gi>
<pmid>12207016</pmid>
<mutation>Y6A</mutation>
<context> ok </context>
<mutation>Y10A</mutation>
<context> ok </context>
<mutation>Y89A</mutation>
<context> ok </context>
<mutation>D113A</mutation>
<context> ok </context>
<mutation>N117A</mutation>
<context> ok </context>
<mutation>E118A</mutation>
<context> ok </context>
<mutation>Y164A</mutation>
<context> ok </context>
<mutation>W172A</mutation>
<context> ok </context>
</component>

<component>
<gi>4514624</gi>
<pmid>12557547</pmid>
<mutation>I156A</mutation>
<context> ok </context>
</component>

<component>
<gi>5381269</gi>
<pmid>14510507</pmid>
<mutation>G32V</mutation>
<context> ok </context>
<mutation>Y176S</mutation>
<context> ok </context>
<mutation>E177Q</mutation>
<context> ok </context>
</component>

<component>
<gi>6226911</gi>
<pmid>9194164</pmid>
<mutation>N127D</mutation>
<context> ok </context>
</component>

<component>
<gi>6226911</gi>
<pmid>10235626</pmid>
<mutation>W85A</mutation>
<context> ok </context>
<mutation>Y172A</mutation>
<context> ok </context>
<mutation>W266A</mutation>
<context> ok </context>
<mutation>W274A</mutation>
<context> ok </context>
</component>

<component>
<gi>6226911</gi>
<pmid>7764794</pmid>
<mutation>F155Y</mutation>
<context> ok </context>
<mutation>R156E</mutation>
<context> ok </context>
<mutation>N173D</mutation>
<context> ok </context>
</component>

<component>
<gi>6226911</gi>
<pmid>9681873</pmid>
<mutation>H86A</mutation>
<context> ok </context>
</component>

<component>
<gi>6226911</gi>
<pmid>9201919</pmid>
<mutation>H81R</mutation>
<context> ok </context>
```

```
<mutation>H207E</mutation>
<context> ok </context>
</component>

<component>
<gi>6226911</gi>
<pmid>7915112</pmid>
<mutation>D124E</mutation>
<context> ok </context>
<mutation>E128Q</mutation>
<context> ok </context>
<mutation>E236Q</mutation>
<context> ok </context>
</component>

<component>
<gi>6434133</gi>
<pmid>10517825</pmid>
<mutation>E86D</mutation>
<context> ok </context>
<mutation>E177D</mutation>
<context> ok </context>
</component>

<component>
<gi>17942986</gi>
<pmid>10752608</pmid>
<mutation>T11Y</mutation>
<context> ok </context>
<mutation>S33P</mutation>
<context> ok </context>
<mutation>A82R</mutation>
<context> ok </context>
<mutation>F168H</mutation>
<context> ok </context>
<mutation>N169D</mutation>
<context> ok </context>
<mutation>Y170D</mutation>
<context> ok </context>
</component>

<component>
<gi>21465565</gi>
<pmid>11526340</pmid>
<mutation>E94A</mutation>
<context> ok </context>
</component>

<component>
<gi>1351447</gi>
<pmid>1359880</pmid>
<mutation>D21E</mutation>
<context> ok </context>
<mutation>E93D</mutation>
<context> ok </context>
<mutation>E182D</mutation>
<context> ok </context>
</component>

<component>
<gi>465492</gi>
<pmid>8376336</pmid>
<mutation>D37N</mutation>
<context> ok </context>
</component>

<component>
<gi>549461</gi>
<pmid>15260499</pmid>
<mutation>T2C</mutation>
<context> ok </context>
<mutation>T28C</mutation>
<context> ok </context>
</component>

<component>
<gi>549461</gi>
<pmid>11917150</pmid>
<mutation>T26R</mutation>
<context> ok </context>
<mutation>Q34R</mutation>
<context> ok </context>
<mutation>S40R</mutation>
<context> ok </context>
</component>

<component>
<gi>549461</gi>
<pmid>11377763</pmid>
<mutation>S110C</mutation>
<context> ok </context>
<mutation>N154C</mutation>
<context> ok </context>
<mutation>Q162H</mutation>
<context> ok </context>
</component>

<component>
<gi>640242</gi>
<pmid>9731776</pmid>
<mutation>E127A</mutation>
<context> ok </context>
<mutation>H250N</mutation>
<context> ok </context>
</component>

</main>
```

Appendix D

Xylanase Test Output Visualization File

```
<menue>
<status=on>
<label>PMID: 10752608, GI : 17942986</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=10752608</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=17942986</link>
<item>
<range>10:A 10:A</range>
<color=yellow>
<status=on>
<label>T -> Y, 100</label>
<context> ok </context>
</item>
<item>
<range>32:A 32:A</range>
<color=yellow>
<status=on>
<label>S -> P, 100</label>
<context> ok </context></item>
<item>
<range>81:A 81:A</range>
<color=yellow>
<status=on>
<label>A -> R, 100</label>
<context> ok </context></item>
<item>
<range>167:A 167:A</range>
<color=yellow>
<status=on>
<label>F -> H, 100</label>
<context> ok </context></item>
<item>
<range>168:A 168:A</range>
<color=yellow>
<status=on>
<label>N -> D, 100</label>
<context> ok </context></item>
<item>
<range>169:A 169:A</range>
<color=yellow>
<status=on>
<label>Y -> D, 100</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 8019418, GI : 2619034</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=8019418</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=2619034</link>
<item>
<range>21:A 21:A</range>
<color=yellow>
<status=on>
<label>D -> N, 68.18</label>
<context> ok </context></item>
<item>
<range>77:A 77:A</range>
<color=yellow>
<status=on>
<label>Y -> F, 86.36</label>
<context> ok </context></item>
<item>
<range>86:A 86:A</range>
<color=yellow>
<status=on>
<label>E -> D, 81.81</label>
<context> ok </context></item>
<item>
<range>88:A 88:A</range>
<color=yellow>
<status=on>
<label>Y -> F, 77.27</label>
<context> ok </context></item>
<item>
<range>120:A 120:A</range>
<color=yellow>
<status=on>
<label>R -> K, 77.27</label>
<context> ok </context></item>
<item>
<range>173:A 173:A</range>
<color=yellow>
<status=on>
<label>Y -> F, 54.54</label>
<context> ok </context></item>
<item>
<range>179:A 179:A</range>
<color=yellow>
<status=on>
<label>E -> D, 59.09</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 8019418, GI : 139865</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=8019418</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=139865</link>
<item>
<range>86:A 86:A</range>
<color=yellow>
<status=on>
<label>E -> D, 81.81</label>
<context> ok </context></item>
<item>
<range>179:A 179:A</range>
<color=yellow>
<status=on>
<label>E -> D, 59.09</label>
<context> ok </context>
</item>
```

```
</menue>
<menue>
<status=on>
<label>PMID: 9684886, GI : 139865</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9684886</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=139865</link>
<item>
<range>108:A 108:A</range>
<color=yellow>
<status=on>
<label>S -> C, 90.90</label>
<context> ok </context></item>
<item>
<range>155:A 155:A</range>
<color=yellow>
<status=on>
<label>N -> C, 54.54</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 15260499, GI : 549461</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract
&list_uids=15260499</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=549461</link>
<item>
<range>29:A 29:A</range>
<color=yellow>
<status=on>
<label>T -> C, 31.81</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 11917150, GI : 549461</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=11917150</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=549461</link>
<item>
<range>27:A 27:A</range>
<color=yellow>
<status=on>
<label>T -> R, 40.90</label>
<context> ok </context></item>
<item>
<range>35:A 35:A</range>
<color=yellow>
<status=on>
<label>Q -> R, 27.27</label>
<context> ok </context></item>
<item>
<range>41:A 41:A</range>
<color=yellow>
<status=on>
<label>S -> R, 50</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 11377763, GI : 549461</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=11377763</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=549461</link>
<item>
<range>110:A 110:A</range>
<color=yellow>
<status=on>
<label>S -> C, 63.63</label>
<context> ok </context></item>
<item>
<range>152:A 152:A</range>
<color=yellow>
<status=on>
<label>N -> C, 54.54</label>
<context> ok </context></item>
<item>
<range>160:A 160:A</range>
<color=yellow>
<status=on>
<label>Q -> H, 72.72</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 14510507, GI : 5381269</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=14510507</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=5381269</link>
<item>
<range>33:A 33:A</range>
<color=yellow>
<status=on>
<label>G -> V, 63.63</label>
<context> ok </context></item>
<item>
<range>168:A 168:A</range>
<color=yellow>
<status=on>
<label>Y -> S, 54.54</label>
<context> ok </context></item>
<item>
<range>169:A 169:A</range>
<color=yellow>
<status=on>
<label>E -> Q, 50</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 1359880, GI : 67399</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=1359880</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=67399</link>
<item>
<range>21:A 21:A</range>
<color=yellow>
<status=on>
<label>D -> E, 40.90</label>
<context> ok </context></item>
```

```
<item>
<range>91:A 91:A</range>
<color=yellow>
<status=on>
<label>E -> D, 77.27</label>
<context> ok </context></item>
<item>
<range>176:A 176:A</range>
<color=yellow>
<status=on>
<label>E -> D, 40.90</label>
<context> ok </context>
</item>
</menue>
<menue>
<status=on>
<label>PMID: 8420796, GI : 67399</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=8420796</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=67399</link>
<item>
<range>12:A 12:A</range>
<color=yellow>
<status=on>
<label>S -> C, 59.09</label>
<context> ok </context></item>
<item>
<range>26:A 26:A</range>
<color=yellow>
<status=on>
<label>S -> W, 50</label>
<context> ok </context></item>
<item>
<range>38:A 38:A</range>
<color=yellow>
<status=on>
<label>G -> D, 50</label>
<context> ok </context></item>
<item>
<range>48:A 48:A</range>
<color=yellow>
<status=on>
<label>R -> K, 50</label>
<context> ok </context></item>
<item>
<range>119:A 119:A</range>
<color=yellow>
<status=on>
<label>T -> S, 68.18</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 11526340, GI : 21465565</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=11526340</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=21465565</link>
<item>
<range>93:A 93:A</range>
<color=yellow>
<status=on>
<label>E -> A, 86.36</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 10517825, GI : 6434133</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=10517825</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=6434133</link>
<item>
<range>86:A 86:A</range>
<color=yellow>
<status=on>
<label>E -> D, 81.81</label>
<context> ok </context></item>
<item>
<range>175:A 175:A</range>
<color=yellow>
<status=on>
<label>E -> D, 50</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 12207016, GI : 2851624</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=12207016</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=2851624</link>
<item>
<range>96:A 96:A</range>
<color=yellow>
<status=on>
<label>Y -> A, 68.18</label>
<context> ok </context></item>
<item>
<range>120:A 120:A</range>
<color=yellow>
<status=on>
<label>Y -> A, 54.54</label>
<context> ok </context></item>
<item>
<range>124:A 124:A</range>
<color=yellow>
<status=on>
<label>Y -> A, 45.45</label>
<context> ok </context></item>
<item>
<range>125:A 125:A</range>
<color=yellow>
<status=on>
<label>D -> A, 45.45</label>
<context> ok </context></item>
<item>
<range>169:A 169:A</range>
<color=yellow>
<status=on>
<label>N -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>177:A 177:A</range>
<color=yellow>
<status=on>
<label>E -> A, 54.54</label>
```

```
<context> ok </context></item>
</menue>
```

Appendix E

Biphenyl Dioxygenase Test Input Mutation File

```
<?xml version="1.0"?>
<main>

<component>
<gi>151084</gi>
<pmid>12081948</pmid>
<mutation>A267S</mutation>
<context> ok </context>
<mutation>V292I</mutation>
<context> ok </context>
<mutation>T335A</mutation>
<context> ok </context>
<mutation>N377T</mutation>
<context> ok </context>
</component>

<component>
<gi>151084</gi>
<pmid>9251194</pmid>
<mutation>I247M</mutation>
<context> ok </context>
<mutation>Y277F</mutation>
<context> ok </context>
<mutation>T335A</mutation>
<context> ok </context>
<mutation>F336L</mutation>
<context> ok </context>
<mutation>N338T</mutation>
<context> ok </context>
<mutation>I341T</mutation>
<context> ok </context>
<mutation>N377T</mutation>
<context> ok </context>
</component>

<component>
<gi>151084</gi>
<pmid>15342624</pmid>
<mutation>T335A</mutation>
<context> ok </context>
<mutation>F336M</mutation>
<context> ok </context>
<mutation>I341V</mutation>
<context> ok </context>
</component>

<component>
<gi>151084</gi>
<pmid>10397810</pmid>
<mutation>T335A</mutation>
<context> ok </context>
<mutation>F336I</mutation>
<context> ok </context>
</component>

<component>
<gi>151084</gi>
<pmid>9190809</pmid>
<mutation>E303D</mutation>
<context> ok </context>
<mutation>V320F</mutation>
<context> ok </context>
<mutation>T325S</mutation>
<context> ok </context>
<mutation>I326V</mutation>
<context> ok </context>
<mutation>T335A</mutation>
<context> ok </context>
<mutation>F336I</mutation>
<context> ok </context>
<mutation>N338T</mutation>
<context> ok </context>
</component>

<component>
<gi>3023415</gi>
<pmid>11312272</pmid>
<mutation>V172I</mutation>
<context> ok </context>
<mutation>T376V</mutation>
<context> ok </context>
</component>

<component>
<gi>3023415</gi>
<pmid>9190809</pmid>
<mutation>M283S</mutation>
<context> ok </context>
<mutation>S324T</mutation>
<context> ok </context>
<mutation>V325I</mutation>
<context> ok </context>
<mutation>T340I</mutation>
<context> ok </context>
<mutation>T376N</mutation>
<context> ok </context>
</component>

<component>
<gi>3023415</gi>
<pmid>11514531</pmid>
<mutation>H255Q</mutation>
<context> ok </context>
<mutation>V258I</mutation>
<context> ok </context>
<mutation>G268A</mutation>
<context> ok </context>
<mutation>F277Y</mutation>
<context> ok </context>
</component>

<component>
<gi>3023415</gi>
<pmid>12057964</pmid>
<mutation>F227V</mutation>
<context> ok </context>
<mutation>L332A</mutation>
<context> ok </context>
```

```
<mutation>I335F</mutation>
<context> ok </context>
<mutation>T376N</mutation>
<context> ok </context>
<mutation>F377L</mutation>
<context> ok </context>
<mutation>F383L</mutation>
<context> ok </context>
</component>

</main>
```

Appendix F

Biphenyl Dioxygenase Test Output Visualization File

```
<menue>
<status=on>
<label>PMID: 12081948, GI : 151084</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=12081948</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=151084</link>
<item>
<range>252:A 252:A</range>
<color=yellow>
<status=on>
<label>V A267 -> S, 40.90</label>
<context> ok </context></item>
<item>
<range>277:A 277:A</range>
<color=yellow>
<status=on>
<label>Q V292 -> I, 31.81</label>
<context> ok </context></item>
<item>
<range>307:A 307:A</range>
<color=yellow>
<status=on>
<label>L T335 -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>349:A 349:A</range>
<color=yellow>
<status=on>
<label>Q N377 -> T, 40.90</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 9251194, GI : 151084</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9251194</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=151084</link>
<item>
<range>222:A 222:A</range>
<color=yellow>
<status=on>
<label>S I247 -> M, 45.45</label>
<context> ok </context></item>
<item>
<range>262:A 262:A</range>
<color=yellow>
<status=on>
<label>S Y277 -> F, 36.36</label>
<context> ok </context></item>
<item>
<range>307:A 307:A</range>
<color=yellow>
<status=on>
<label>L T335 -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>308:A 308:A</range>
<color=yellow>
<status=on>
<label>T F336 -> L, 36.36</label>
<context> ok </context></item>
<item>
<range>310:A 310:A</range>
<color=yellow>
<status=on>
<label>S N338 -> T, 31.81</label>
<context> ok </context></item>
<item>
<range>313:A 313:A</range>
<color=yellow>
<status=on>
<label>F I341 -> T, 31.81</label>
<context> ok </context></item>
<item>
<range>349:A 349:A</range>
<color=yellow>
<status=on>
<label>Q N377 -> T, 40.90</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 15342624, GI : 151084</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=15342624</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=151084</link>
<item>
<range>307:A 307:A</range>
<color=yellow>
<status=on>
<label>L T335 -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>308:A 308:A</range>
<color=yellow>
<status=on>
<label>T F336 -> M, 36.36</label>
<context> ok </context></item>
<item>
<range>313:A 313:A</range>
<color=yellow>
<status=on>
<label>F I341 -> V, 31.81</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 10397810, GI : 151084</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=10397810</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
```

```
/viewer.fcgi?db=protein&val=151084</link>
<item>
<range>307:A 307:A</range>
<color=yellow>
<status=on>
<label>L T335 -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>308:A 308:A</range>
<color=yellow>
<status=on>
<label>T F336 -> I, 36.36</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 9190809, GI : 151084</label>
<link>http://www.ncbi.nlm.nih.gov/entrez/
query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9190809</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=151084</link>
<item>
<range>288:A 288:A</range>
<color=yellow>
<status=on>
<label>R E303 -> D, 18.18</label>
<context> ok </context></item>
<item>
<range>305:A 305:A</range>
<color=yellow>
<status=on>
<label>S V320 -> F, 27.27</label>
<context> ok </context></item>
<item>
<range>297:A 297:A</range>
<color=yellow>
<status=on>
<label>N T325 -> S, 9.09</label>
<context> ok </context></item>
<item>
<range>298:A 298:A</range>
<color=yellow>
<status=on>
<label>C I326 -> V, 18.18</label>
<context> ok </context></item>
<item>
<range>307:A 307:A</range>
<color=yellow>
<status=on>
<label>L T335 -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>308:A 308:A</range>
<color=yellow>
<status=on>
<label>T F336 -> I, 36.36</label>
<context> ok </context></item>
<item>
<range>310:A 310:A</range>
<color=yellow>
<status=on>
<label>S N338 -> T, 31.81</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 11312272, GI : 3023415</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=11312272</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=3023415</link>
<item>
<range>117:A 117:A</range>
<color=yellow>
<status=on>
<label>V V172 -> I, 77.27</label>
<context> ok </context></item>
<item>
<range>314:A 314:A</range>
<color=yellow>
<status=on>
<label>K T376 -> V, 50.0</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 9190809, GI : 3023415</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=9190809</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=3023415</link>
<item>
<range>266:A 266:A</range>
<color=yellow>
<status=on>
<label>V M283 -> S, 22.72</label>
<context> ok </context></item>
<item>
<range>307:A 307:A</range>
<color=yellow>
<status=on>
<label>L S324 -> T, 40.90</label>
<context> ok </context></item>
<item>
<range>299:A 299:A</range>
<color=yellow>
<status=on>
<label>T V325 -> I, 27.27</label>
<context> ok </context></item>
<item>
<range>314:A 314:A</range>
<color=yellow>
<status=on>
<label>K T340 -> I, 50.0</label>
<context> ok </context></item>
<item>
<range>350:A 350:A</range>
<color=yellow>
<status=on>
<label>R T376 -> N, 50.0</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 11514531, GI : 3023415</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=11514531</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
```

```
/viewer.fcgi?db=protein&val=3023415</link>
<item>
<range>230:A 230:A</range>
<color=yellow>
<status=on>
<label>N H255 -> Q, 27.27</label>
<context> ok </context></item>
<item>
<range>233:A 233:A</range>
<color=yellow>
<status=on>
<label>L V258 -> I, 27.27</label>
<context> ok </context></item>
<item>
<range>243:A 243:A</range>
<color=yellow>
<status=on>
<label>T G268 -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>260:A 260:A</range>
<color=yellow>
<status=on>
<label>V F277 -> Y, 36.36</label>
<context> ok </context></item>
</menue>
<menue>
<status=on>
<label>PMID: 12057964, GI : 3023415</label>
<link>http://www.ncbi.nlm.nih.gov/entrez
/query.fcgi?cmd=Retrieve&db=pubmed
&dopt=Abstract&list_uids=12057964</link>
<link>http://www.ncbi.nlm.nih.gov/entrez
/viewer.fcgi?db=protein&val=3023415</link>
<item>
<range>202:A 202:A</range>
<color=yellow>
<status=on>
<label>F F227 -> V, 63.63</label>
<context> ok </context></item>
<item>
<range>306:A 306:A</range>
<color=yellow>
<status=on>
<label>M L332 -> A, 40.90</label>
<context> ok </context></item>
<item>
<range>309:A 309:A</range>
<color=yellow>
<status=on>
<label>C I335 -> F, 40.90</label>
<context> ok </context></item>
<item>
<range>350:A 350:A</range>
<color=yellow>
<status=on>
<label>R T376 -> N, 50.0</label>
<context> ok </context></item>
<item>
<range>351:A 351:A</range>
<color=yellow>
<status=on>
<label>T F377 -> L, 50.0</label>
<context> ok </context></item>
<item>
<range>357:A 357:A</range>
<color=yellow>
<status=on>
<label>F F383 -> L, 45.45</label>
<context> ok </context></item>
</menue>
```