AD_____

Award Number: DAMD17-03-1-0520

TITLE: A System for Discovering Bioengineering Threats by Knowledge Base Driven Mining of Toxin Data

PRINCIPAL INVESTIGATOR: Subramanyam Swaminathan, Ph.D.
I.V. Ramakrishnan
M. Kifer
H. Davulcu

CONTRACTING ORGANIZATION: Brookhaven National Laboratory
Upton, NY 11973-5000

REPORT DATE: August 2005

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-08-2005	Annual	1 Aug 2004 – 31 Jul 2005

4. TITLE AND SUBTITLE
A System for Discovering Bioengineering Threats by Knowledge Base Driven Mining of Toxin Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-03-1-0520

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Subramanyam Swaminathan, Ph.D.
I.V. Ramakrishnan
M. Kifer
H. Davulcu
E-mail: swami@bnl.gov

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Brookhaven National Laboratory
Upton, NY 11973-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The overall goal of this project is to establish a Toxin Knowledge Base (TKB) – a bioinformatics resource primarily focused on molecular information about toxins and other virulence factors that are the natural products of *biological and potential biological warfare* (*BW and PBW*) agents. The resource will be mined to assimilate, synthesize, analyze and disseminate genomic and structural information on BW and PBW genes and their products. The TKB will be designed and developed to serve as a powerful analysis and decision system tool for the intelligence community. In this first annual report we describe in detail the TKB system that we have developed that can be used to identify homologs of toxins. The TKB contains molecular, biological and structural information of about 1000 toxins and is still being expanded. TKB will also use innovative computer methods to parse the literature available in public resources (web sites) to identify new and emerging toxins to be included in the database.

15. SUBJECT TERMS
Toxin data base, toxin homologs, bioengineered threat, text mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	33	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

A System for Discovering Bioengineered Threats by Knowledge Base Driven Mining of Toxin Data

Annual Report for the Period ending July 2005

Introduction

The overall goal of this project is to establish an easy to use database *viz.* a Toxin Knowledge Base (TKB) which will populate itself and expand using machine learning techniques. It will be a bioinformatics resource primarily focused on molecular information about toxins and other virulence factors that are the natural products of biological and potential biological warfare (BW and PBW) agents. The resource will be mined to assimilate, synthesize, analyze and disseminate genomic and structural information on BW and PBW genes and their products. Using advanced machine learning and data mining the TKB will be mined to look for motifs, to design new experiments and also to predict structure and function of molecules (including putative chimeras) for which these data are not available. Knowledge learned from this and similar analysis will be encoded as rules in an expert system. Both the TKB and its front-end expert system will be used for analyzing genomic data to compare pathogenic and non-pathogenic viral, bacterial and plant genomes in order to identify specific regions that encode factors that contribute to virulence. TKB will also use innovative computer methods to parse the literature available in public resources (web sites) to identify new and emerging toxins to be included in the database.

Body

A. Design and implementation of a highly curated Toxin Knowledge Base:

In the last two years we have modified, improved and expanded the previously existing database for storing, managing and accessing molecular information on known as well as potential biological toxins. A description was presented in last year's report. Here we are presenting a more complete description with new additions and examples. This section describes our work, progress and deliverables as given in Specific Tasks 1, 2

and 3. As of now, the system is ready to be tested by interested people and we are requesting permission from the Army to open this to interested scientists. We have proper procedures in place to keep the data and the site secure.

System Architecture

The primary motivation for developing TKB was to address the need to establish an infrastructure resource that will aid studies in (1) developing methods for identifying potential bio-warfare agents, (2) identifying and developing counter measures such as anti-toxins, vaccines, and inhibitors, and (3) developing a better understanding of the mode of actions of these toxins at the cellular, sub-cellular, and molecular levels. TKB also focuses on correlating known and predicted 3-dimensional structures for these toxins with sequence, function, and biological activity. In order to develop a system that satisfies all these aims, we have developed a comprehensive architecture that accommodates the needs of a growing system.

TKB is comprised of two major components: (1) A powerful data-acquisition/ administration system for direct deposition of data related to toxins and (2) an ad-hoc query and reasoning system to access and to analyze information. Figure 1 shows the system architecture of TKB showing the querying and reasoning subsystem and the data acquisition subsystem. It also shows the architecture of the system, from the users' perspective.

Toxin Knowledge Base (TKB) is used to store biological information about various kinds of toxins. It stores homologs and active site information for each toxin and models for the homologs. It provides two interfaces to the user namely:

1. **Query and Reasoning Interfaces**: This facilitates the following:
    a. Toxin Search - Selective retrieval of toxin information.
    b. Homology Search - Finding toxins that are homologous to a given protein sequence.
    c. MuToxin - Determining whether a protein can be transformed into a toxin.
2. **Administrative Interfaces**: This interface is accessible only to a user with administrative rights. The toxin knowledge base can be updated in two ways:
    a. User-initiated: This involves updating the knowledge base with newly identified toxin information and related active site information.

b. Automated: This involves updating the knowledge base with new homologs and their models for the toxins on a periodic basis, so as to keep the toxin knowledge base up-to-date.

c. User Approval: This allows a new user's identity to be verified and approved for use of the TKB.

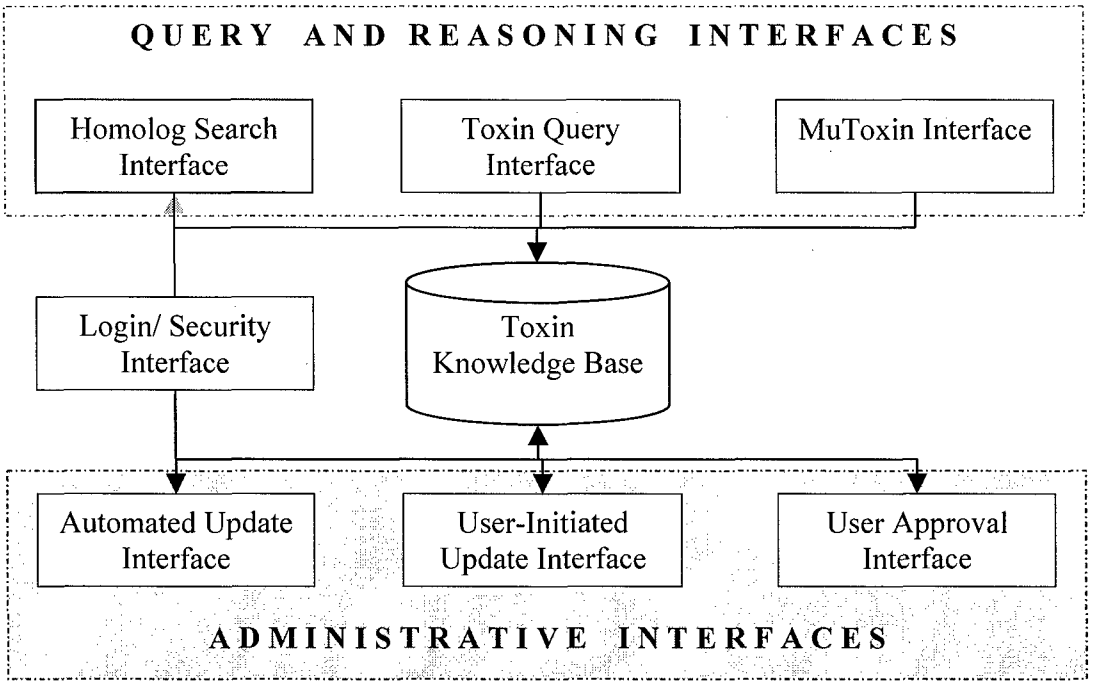


Figure 1: System Architecture: The above figure represents a concise architecture of the system which has been developed by Brookhaven National Laboratory and Stony Brook University. The toxin knowledge base essentially is a data source which provides two kinds of interfaces to the user – one used to query the knowledge base and the other used to update the information in the knowledge base.

TKB integrates several publicly available tools that were developed for various unrelated purposes, but which are engineered into a workflow for identifying potential mutated toxins. This is a part of the query and reasoning interface, and is explained further in the section on the Query and Reasoning Interfaces. The system's architecture and description are organized as follows. First, the powerful data acquisition system is presented, along with the workflow used for the same. Second, we present the logical query and inference system that has been developed to identify a protein that can be converted into a toxin. The implementation details of the system is finally presented, with

a report on the current status of the knowledge base, and followed by a section on the results obtained so far using the system.

Data Acquisition System

Public databases of biological information are popular research tools in the biological community. While providing wealth of information, they offer little help in analyzing, assimilating, and collecting data related to a particular topic (like toxins). As a result, the user is forced to search through multiple data sources and correlate the data manually. TKB fills a sorely needed gap. In particular it is an integrated tool for collecting, aggregating, and analyzing toxin data from different *data sources*. The sources that we currently use in our data acquisition process are PUBMED, SWISS-PROT, and RCSB.

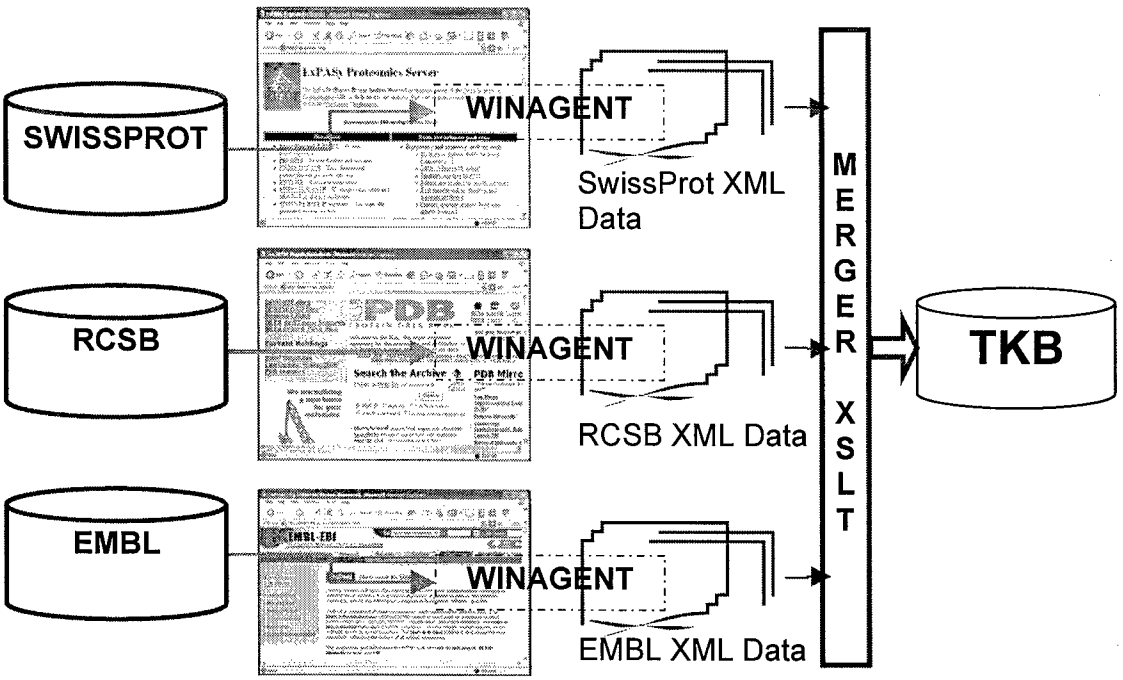


Figure 2: Data Acquisition Workflow. The data acquisition process consists of integrating data from disparate sources (publicly available bioinformatics databases) into a common data repository such that the information can be collected and assembled so that it can be queried.

In order to acquire data from vastly disparate sources like the RCSB, SWISSPROT and EMBL, an information extraction tool was built using an inbuilt tool known as WinAgent. This tool can be used to mine data from various web data sources. WinAgent is a software robot that learns to extract data from the Web by observing a

user's navigation activity. By training the agent on the websites of interest, the user can easily teach the tool to acquire relevant data. In order to overcome the problem of data incompatibility among the different sites, a merger XSLT tool was built that compiles all the data into a single unified schema (explained further in this section) and stores it into the TKB. Most of the data acquisition is automated, except for situations in which a new data source has been identified, and the WinAgent has to be trained to extract information from such a data source. This process is also made easy by the fact that the training process is just a few clicks on the mouse and showing how the user would want to navigate the new data source. The data acquisition system includes algorithms that make it scalable in lieu of the rapidly growing amounts of data.

The data thus acquired is stored in a single unified schema, shown in figure 3. The schema has been developed with great care in order to include all relevant information a user might want to learn about a toxin. The sample schema shown here includes the details presented when the user looks at a particular toxin. In this example we present the schema when the user has selected botulinum neurotoxin type E as his choice. As shown in the figure the list of schema headings (left hand side of the table) is quite comprehensive. Certain fields have more information – providing a summary on the activity and mechanism of reaction (if it is available in the data sources or through literature search) of a toxin, where as certain fields do not have a lot of information, but represented using a hyperlink. This indicates that the information was collected from a different public source and can be then obtained by clicking on that relevant hyperlink.

An exception to this case is the capability of the system to provide the user with the structure information, if it is available. Using an internet plug-in called Chime (http://www.mdli.com/chime), the user can directly link to the structure and perform various operations on the 3 dimensional structures using mouse buttons and key board buttons. This allows for increased interactivity with the system, and hence improves the overall experience for the user while using the TKB. This usage has been further shown in the usage section of the report.

The MuToxin (Trans-toxin) Workflow

A critical part of the system is dedicated to the derivation of new knowledge from the existing knowledge. Hence as part of the powerful query and reasoning system, we

have included an engineered workflow that allows the end user to determine whether a given protein, (1) resembles a toxin at its active site and (2) whether residue substitutions at specific locations on the protein, can modify the protein into a toxin (minimally the active site). The workflow is illustrated in Figure 4.

As shown in figure 4, the workflow integrates three separate off-the shelf bioinformatics and structural biology resources into a neat workflow. When the user provides an input protein sequence through the user interface, the homologs of the input sequence is collected and based on the homologs, if a structure exists within our structure database, a reasonable model is built using the Modeller program. Based on the active site information available, it is then provided as input to the SPASM program, which superposes the built model against a database of active site templates and compares them for some match using a customized substitution matrix score. This provides the end user with a reasonable estimate as to whether the input protein resembles a toxin in some fashion.

Another output from the workflow is a table of substitution scores and positions at which possible residue substitutions need to be made such that the active site resembles the target toxin. This provides information whether the protein can be a potential chimera (to hide a potentially toxic active site into a benign protein). All these are illustrated in the usage section of this report.

Toxin Name	Botulinum neurotoxin type E [Precursor]
category	Bacteria
synonyms	EC, 3.4.24.69, BoNT/E, Bontoxilysin E,
Organism_scientific_name	*Clostridium botulinum*
swiss_prot_entry	*Q00496*
GeneBank_accession_no	*X62089*
rcsb_pdb_entry	1E1H
gene_locus	CBNEUTOXE 4017 bp DNA linear BCT 18-APR-2005
strain	
gene	None
length	1250 AA [This is the length of the unprocessed precursor]
molecular_weight	143713 Da [This is the MW of the unprocessed precursor]
calculated_pI	6.19
structure_representation	1E1H
PDBSum	1E1H
amino_acid	*Amino Acid*
gene_sequence	*X62089*
related_structures	Belongs to the peptidase M27 family [view classification] .
metal_cofactor	Binds 1 zinc ion per subunit
inhibitors	
target	
ab_toxin	
pore_former	
enzyme_catalyzed	
EC_number	*3.4.24.69*
mode_of_action	Botulinum toxin acts by inhibiting neurotransmitter release. It binds to peripheral neuronal synapses, is internalized and moves by retrograde transport up the axon into the spinal cord where it can move between postsynaptic and presynaptic neurons. It inhibits neurotransmitter release by acting as a zinc endopeptidase that catalyzes the hydrolysis of the 180-Arg-|-Ile-181 bond in SNAP-25.
mechanism	Botulinum toxin acts by inhibiting neurotransmitter release. It binds to peripheral neuronal synapses, is internalized and moves by retrograde transport up the axon into the spinal cord where it can move between postsynaptic and presynaptic neurons. It inhibits neurotransmitter release by acting as a zinc endopeptidase that catalyzes the hydrolysis of the 180-Arg-|-Ile-181 bond in SNAP-25.
biochemical_information	Catalytic Activity : Limited hydrolysis of proteins of the neuroexocytosis apparatus, synaptobrevins, SNAP25 or syntaxin. No detected action on small molecule substrates. Cofactor : Binds 1 zinc ion per subunit
biomedical_information	
reference	[1] NUCLEOTIDE SEQUENCE Pubmed, Medline [2] NUCLEOTIDE SEQUENCE Pubmed, Medline [3] NUCLEOTIDE SEQUENCE OF 1-251 [4] PROTEIN SEQUENCE OF 1-13 Pubmed, Medline [5] PROTEIN SEQUENCE OF 419-426 Pubmed, Medline [6] NUCLEOTIDE SEQUENCE OF 615-981 Pubmed, Medline [7] IDENTIFICATION OF SUBSTRATE Pubmed, Medline [8] IDENTIFICATION OF SUBSTRATE Pubmed, Medline
keywords	Botulinum neurotoxin type E [Precursor], EC, 3.4.24.69, BoNT/E, Bontoxilysin E, Bacteria, Firmicutes, Clostridia, Clostridiales, Clostridiaceae, Clostridium

Figure 3: Overall schema representation for Botulinum neurotoxin type E [precursor]

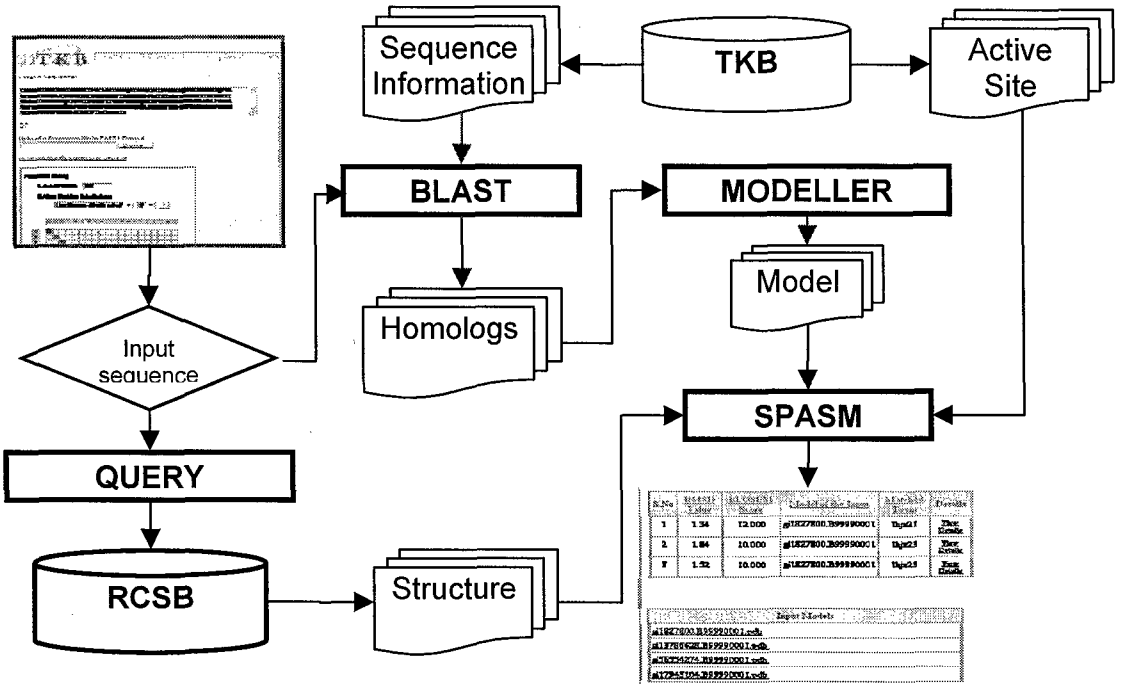


Figure 4: The Mu-Toxin (Trans-Toxin) Workflow.

System Implementation

The system has been developed entirely using Java, Java Server Pages (JSP), HTML and XML/ XSLT technologies. Essentially organized into three layers (based on the Model View Controller design pattern), the front end (view) of the system consists of interacting JSP which are kept extremely functional. All the aspects of user views and definitions are made using XSLT, which allows for a very flexible front end to be developed. In fact the user is usually unaware of the existence of the XSLT since the code generated on the front end is very dynamic in nature.

The controller objects are developed as Java Servlets, with ability to handle multiple sessions, control opening and closing of new windows as and when required, pass session control to JSP and retain information for further processing of user commands. The controller objects do not generate any HTML artifacts except for some administrative logs that are stored at the server end for monitoring the status of the system.

The model (back end) is implemented using Oracle 10G as the primary database, with extensive support using XML. The database schema is very flexible in order to

accommodate periodic changes that may be necessary because of the ever expanding knowledge within the field of toxicology.

We also provide here the current status of the database and the various statistics as an estimate of the size of the database tables.

Total Number of Toxins	1009
Number of Toxins with Structures	539
Total number of Homologs	79,658 (79 Homologs / Toxin)
Total size of Toxin database	1.64 GB
Total number of indices used	14

Usage Scenarios

In this section a detailed look into the system implementation is provided. We provide information about the various interfaces, how they are organized, how a typical user will navigate and use the system (both a normal user and an administrator) and also various screen shots of the system.

Logging into the System

An important step towards using the system is to have some form of authentication of the end users, so that the system is not compromised. To this effect all users need to log into the system. The login interface is a simple authentication mechanism, which verifies the user (through a suitable user-name/ password) and also provides access to the user interface that a user is privileged to use (Figure 5). This means that a user just needs to type in his user name and password – and the system then recognizes the privileges of the user and allows access to only those pages that he/she can use. This improves the security of the system by providing a single point of entry into the system.

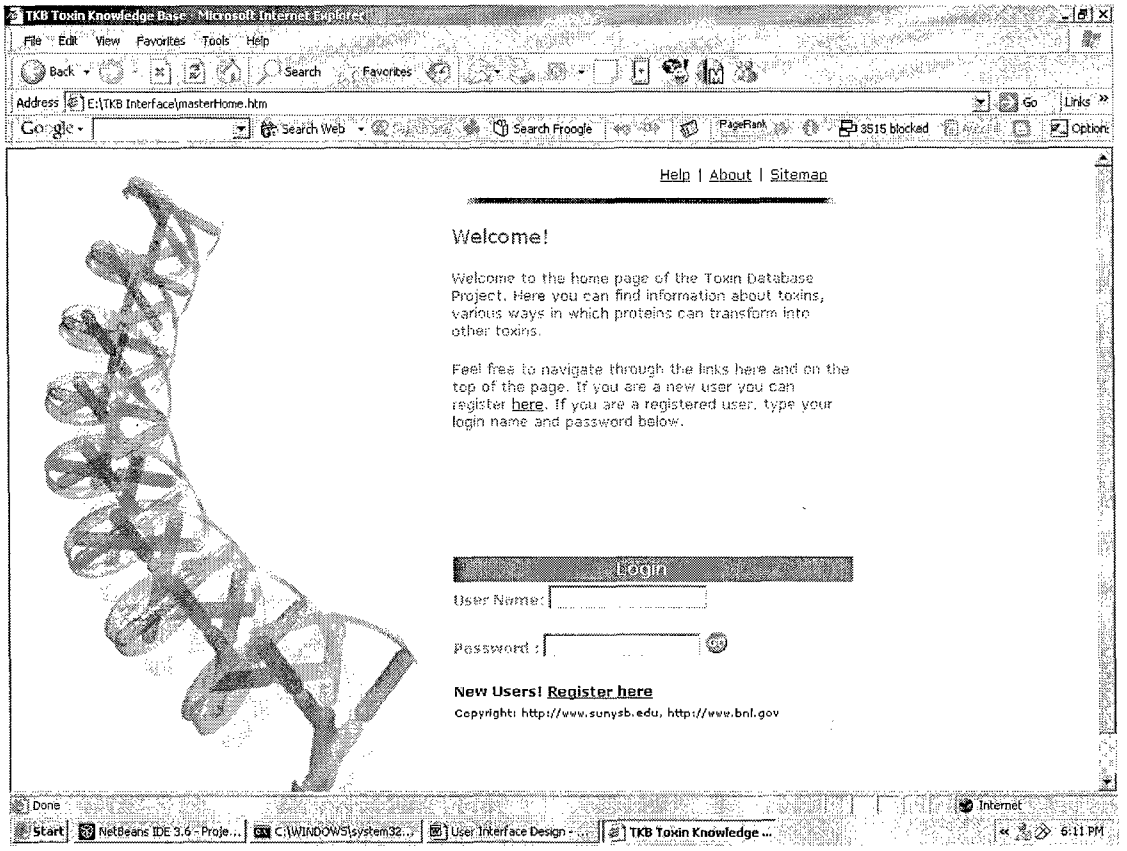


Figure 5: Login Screen: The login screen is a simple welcome page that allows the user to enter his user name and password so that he can login to the system. There is a separate form for new users who want to register and start using the system. However, new users can register only after they have been screened for security purposes.

Query and Reasoning Interface

The query interface provides facilities to the user to query and use the information stored in the knowledge base in different ways, described in the following sections.

Homology Search

- This interface helps in finding homologs of a given protein sequence from the TKB using PSI-BLAST (Position specific iterative BLAST), which is the NCBI tool for homology search.
- The "Homology search" interface (as shown below in Figure 6) accepts two forms of input from the user:
    1. A protein FASTA sequence (or a list of such sequences)
    2. An accession number (or a list of accession numbers)

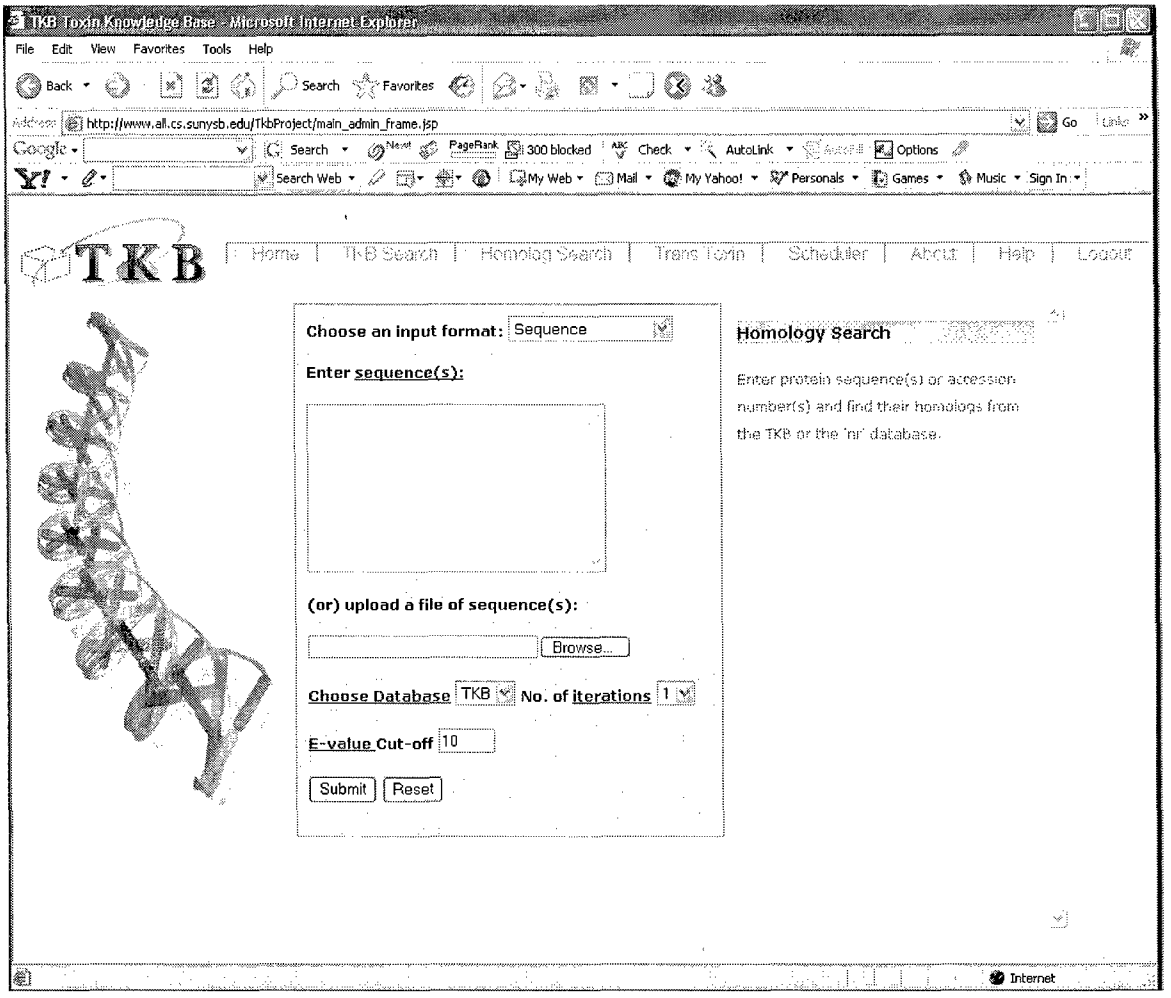


Figure 6: Homolog Search Interface: The search interface helps user to find a homolog given either a sequence or an accession number. The right hand pane shows how a help page can be dynamically loaded based on the links on the form. This allows all users with minimal knowledge of the system to understand the terms and use the system with minimal training.

- The following are the options provided for the homology search:
    1. Database: Provides a choice of database to be BLASTed against. The two options that are currently provided are the "TKB" and "nr"(non-redundant database)
    2. Number of iterations: PSI-BLAST uses the results of each "iteration" to refine the profile. This iterative searching strategy results in increased sensitivity.
    3. E-value cut-off: Lets the user define the "expect" threshold for the homology search.

- When the user submits the required inputs and options, PSI-BLAST is used to search against the specified database and the results are presented to the user in a concise yet comprehensive fashion, as the Figure 7 depicts.

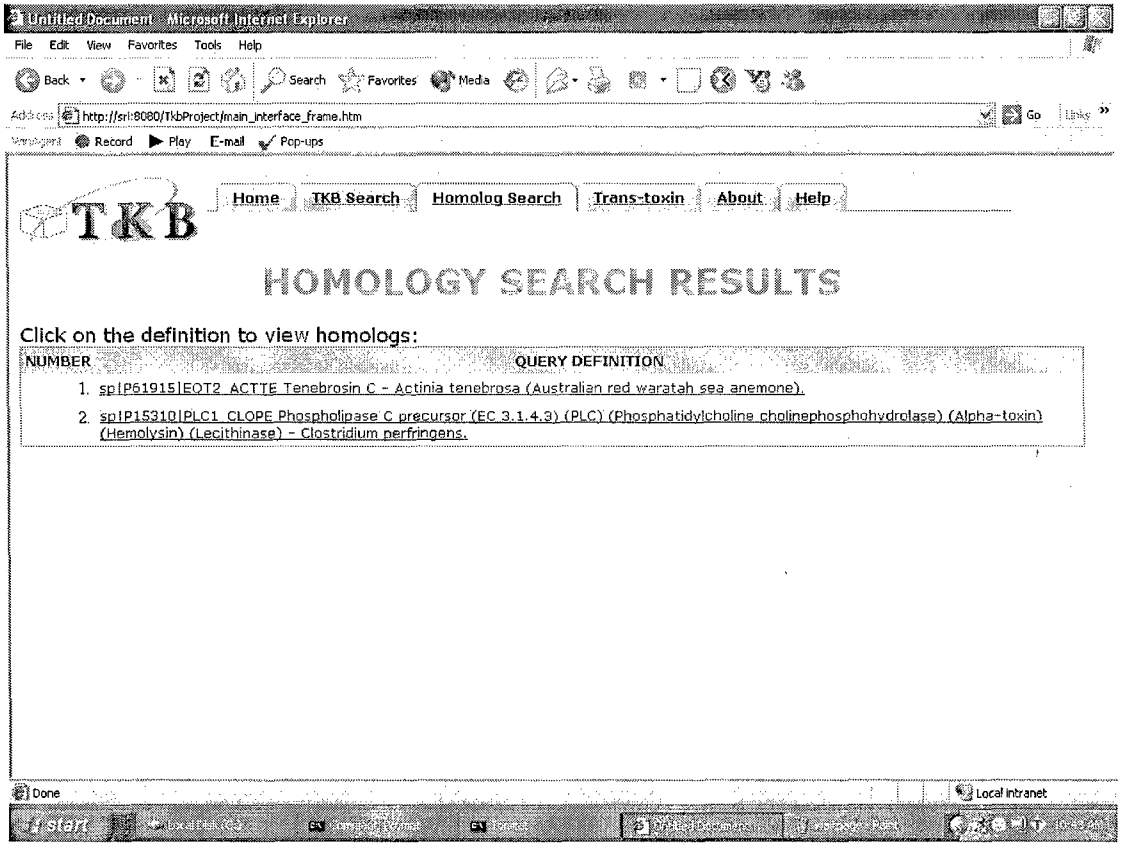


Figure 7: Results for homology search: The initial results list the query sequences (if more than one query has been submitted) which are links which take the user to the results page with the list of actual homologs.

- The list of homologs in a page-wise format, iteration by iteration, is displayed as shown below in Figure 8.

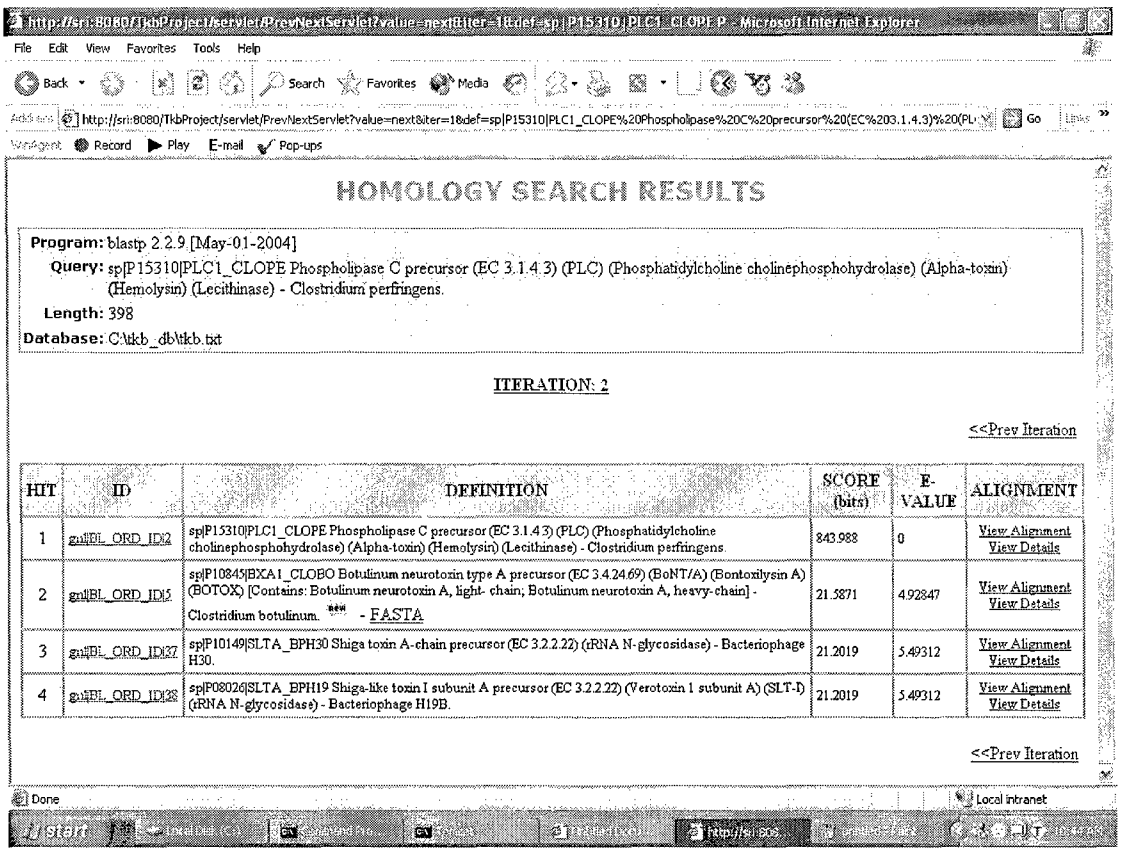


Figure 8: Results for Homology Search: The results of the each query are reproduced in a easy-to-use tabular format that shows all the required information, along with the new homologs identified in subsequent iterations, that are highlighted as shown here.

- Each homolog has the following information:
    1. A link to the NCBI/Swiss-Prot entry depending on whether the homologs are from "nr" or "TKB" respectively.
    2. New homologs in subsequent iterations are identified.
    3. Each homolog has the following information:
        - Score for each homolog
        - E-value for each homolog
        - Pair-wise alignment of the query and homolog sequence, showing the positives, identities and gaps. (shown in Figure 9 below)
        - Alignment details (shown in the Figure 9 below)

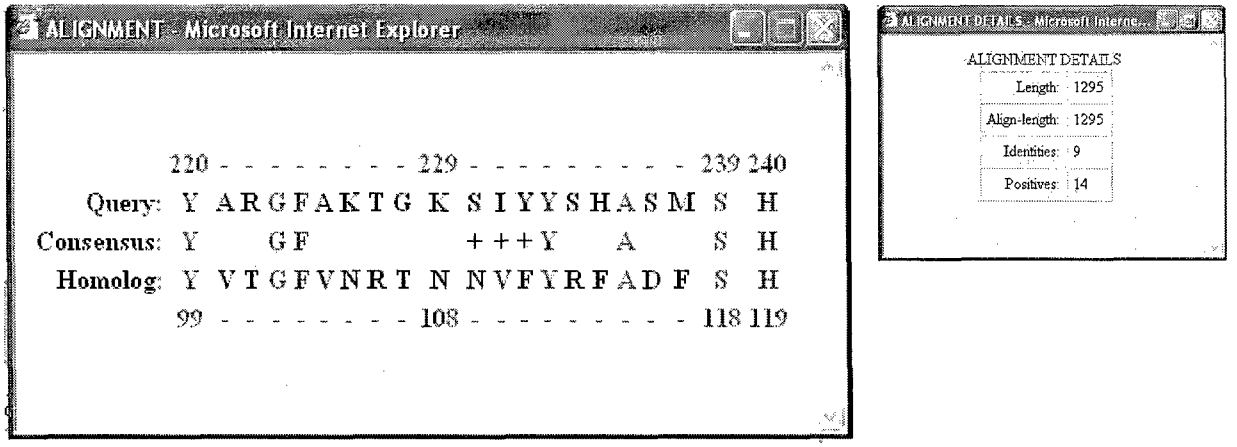


Figure 9: Pair-wise alignment and Details: The pair-wise alignment shows the identities(the residues marked red) and the positives (the residues marked blue). The alignment details are also provided.

TKB Search

- The TKB Search interface is useful to view the toxin data stored in the TKB.
- The user can browse through the toxins alphabetically. The user can also search for particular toxins by specifying certain filter criteria as shown in the Figure 10 below.

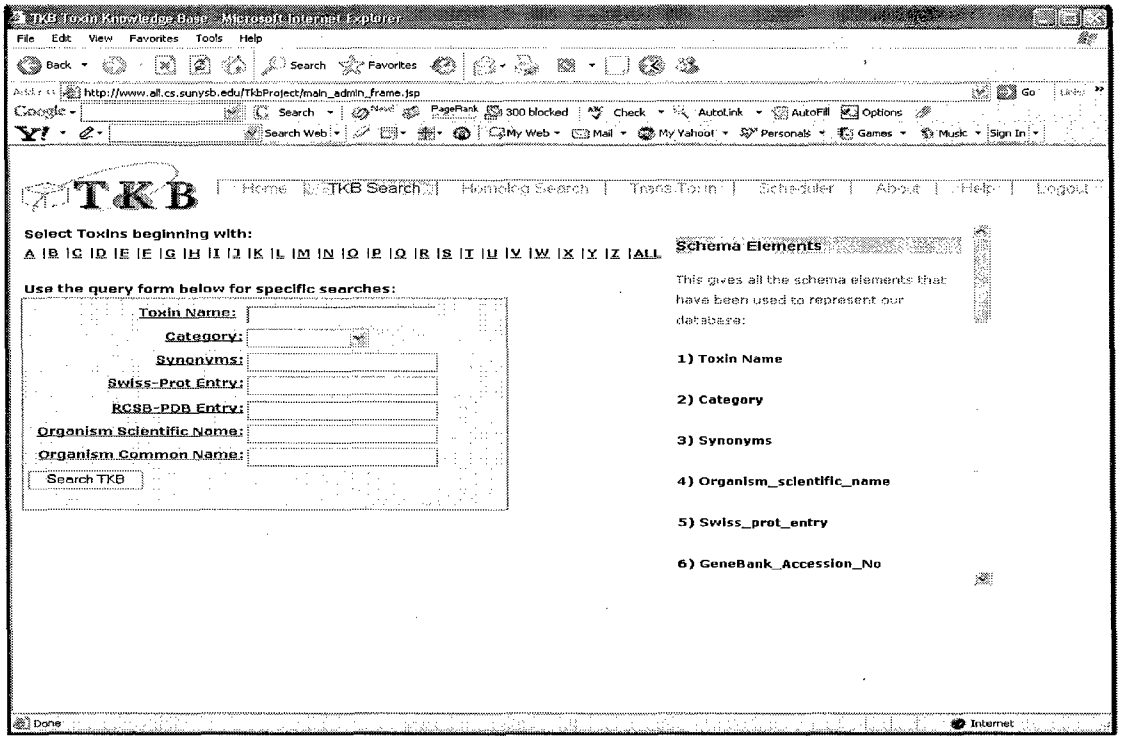


Figure 10: TKB Search Interface: The above figure displays the search options

The search results containing the names of the toxins satisfying the chosen filter criteria are displayed as shown in the following figures (Figures 11).

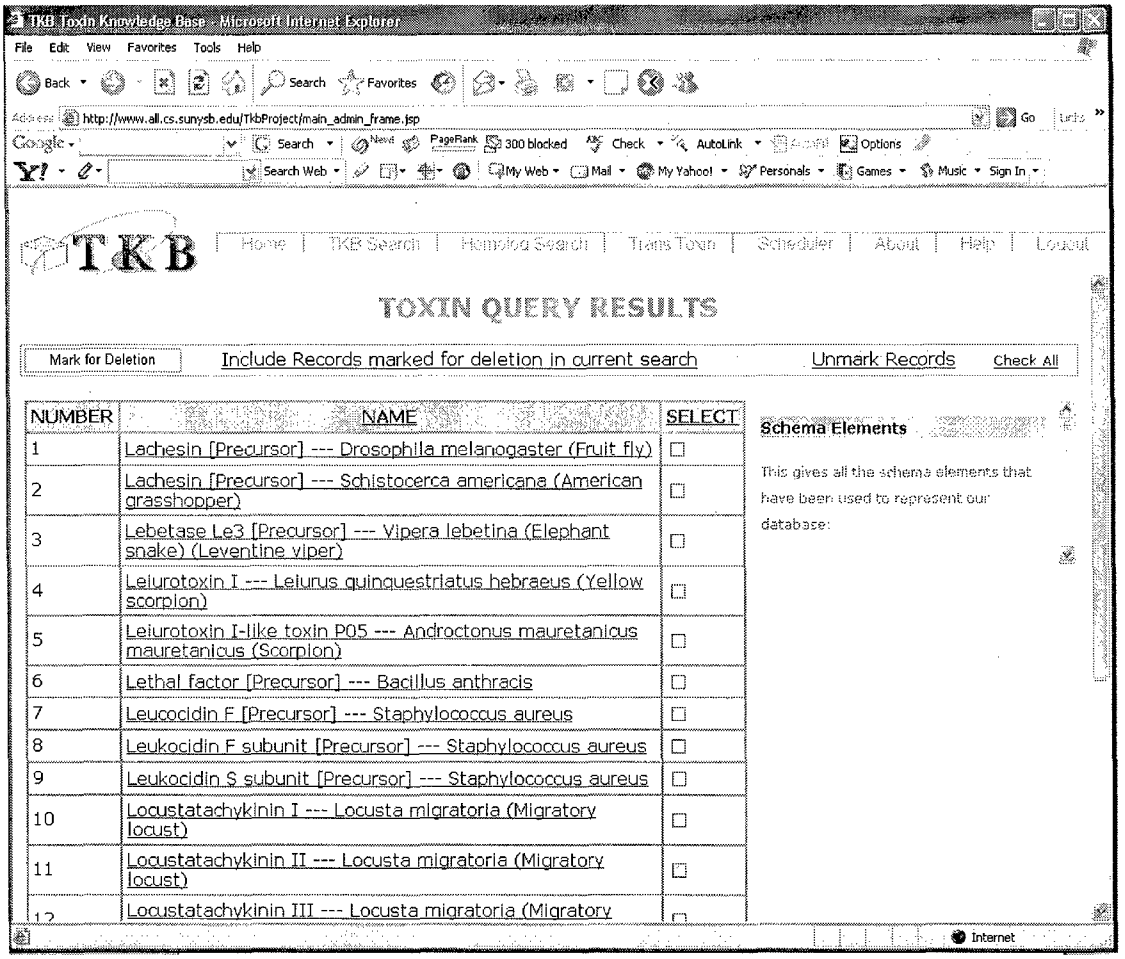


Figure 11: TKB Search Results: The above are the results that are displayed when the user searched for toxins starting with alphabet 'L'

- The details of the toxin can be seen by clicking on that particular toxin name in the search results. This was illustrated in the previous section on the system design (figure 3).

Trans-Toxin (Mu-Toxin)

- This interface lets the user investigate whether a protein can be transformed into a toxin.

- It accepts the protein sequence from the user, either as file or a text string. The user interface is shown in Figure 12.

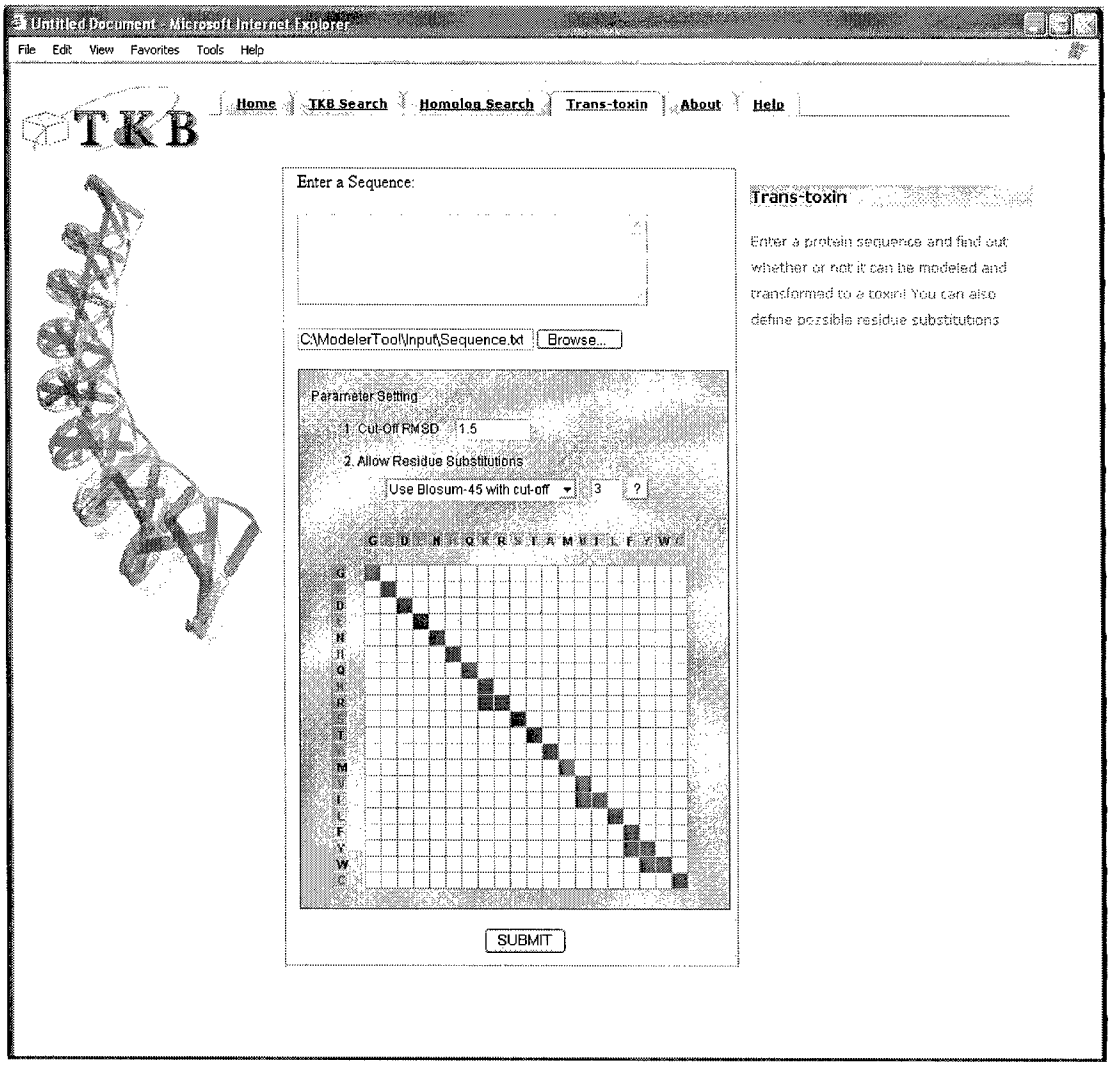
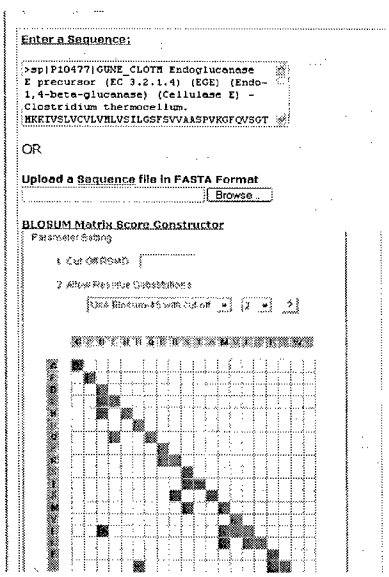


Figure 12: MuToxin (Trans-toxin) Interface.

A sample output from the Mutoxin (trans-toxin) interface is shown in the figure below. It provides two results. One is a tabulation of all the possible matches of the input protein against the templates of active sites in the knowledge base with the RMSD values, BLOSUM scores and model output by Modeller to get these results. It also provides a detailed view (highlighted in the picture), where in the user can see the matching active site residues, and the corresponding BLOSUM scores (if they were selected by the user;

otherwise the customized score value is calculated from the input as highlighted by the user).



S.No	RMSD Value	BLOSUM Score	Model of the Input	Matched Toxin	Details
1	1.34	12.000	gi1827800.B99990001	1hjx21	View Details
2				23	View Details
3				23	View Details

Matched Toxin	Model of the Input	BLOSUM Score
GLY C 181	GLY 184	7.0
LYS C 182	ARG 181	3.0
THR C 184	SER 183	2.0

Input Models
gi1827800.B99990001.pdb
gi13786628.B99990001.pdb
gi56554274.B99990001.pdb
gi17943104.B99990001.pdb

Administrative interfaces

The update interface lets the Administrator maintain a consistent and up-to-date knowledge base. The updates to the toxin knowledge base can be done in two ways:

- Automated: This includes the following two tasks:
  - Update TKB
  - Update NR
- User-initiated

These tasks are explained in detail in the following sections.

Update TKB

The TKB has to be updated whenever new proteins are added to the NR database. New entries in the NR database could mean additional homologs for the toxins in the TKB. An update on the TKB is performed by blasting each toxin against the most recent additions to the NR database.

Given a sequence from the TKB, the following steps in this task have been automated.

i. Blast the Query sequence against the most recently available updates to the NR database.

ii. Process the list of homologs to obtain the list of homologs that are relevant to the input to be given to the modeler. This processing involves filtering based on e-values and identity cut-offs.

iii. Store the desired set of homologs in the TKB.

Update NR

A copy of the NR database is being maintained. Sequences are added to the NR database whenever new proteins are released. This addition of sequences involves matching new sequences to existing ones and appending and/or inserting new entries in the NR. Throughout the process, caution is taken to maintain the non-redundant property of the NR database.

User initiated update of TKB

This is the step when a new toxin has been identified and hence an entry is made in the TKB along with the following information:

- PDB ID: Using WinAgent, Swiss-Prot is searched with the toxin's accession number (for example, P10844 for Botulinum neurotoxin type B), and then the PDB ID's for the toxin are extracted.
- Active site information: Using WinAgent, PDB, PDBSUM, and LPC databases are searched using the PDB ID of a toxin, and then the active site information, if there is any is extracted.
- Models: For each homolog, the toxin to which it is homologous is known. If the toxin structure information is available, a model is built for the homolog based on the alignment of the toxin and the homolog as well as the structure of the toxin using MODELLER. MODELLER is a well-known comparative modeling tool. It has its own script language to control the modeling process. Using a program to generate the script, the modeling process is automated.

Scheduler

- The above-mentioned update tasks are long-running tasks, to be performed on a periodic basis. A Scheduler application has been developed to handle the updates.
- The Administrator is provided with a facility to schedule these tasks to be executed at a specified time and interval. It is the responsibility of the Scheduler application to execute the tasks thereafter, at the predefined time and interval on a regular basis.

The user interface for the scheduler is shown in Figure 13 below:

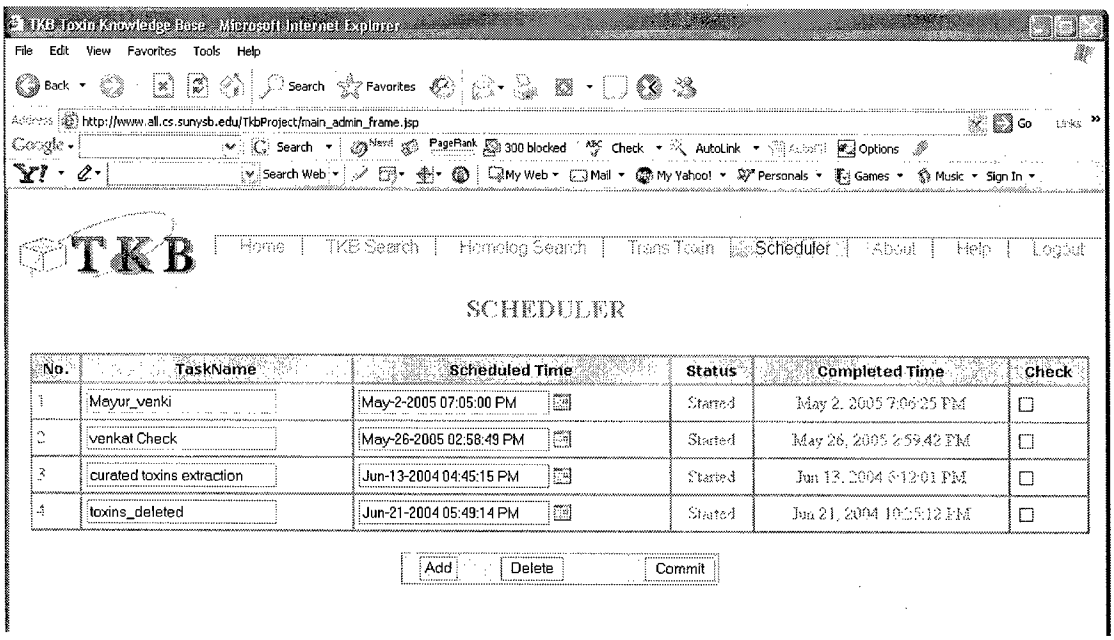


Figure 13: The Update Interface: The administrator can schedule the Update Task by just selecting the Task from a menu. The administrator decides the start date for Scheduling an event.

- The information about the status of the execution of these scheduled tasks is stored in the database. The Administrator will be provided with an interface to see the status of these scheduled tasks and also update the task information. This will include the time at which the task is to be executed or the interval between two successive executions of the task. The user will also be able to delete a pre-defined task.

Help interface

Help pages are provided for each module of the interface as shown in the screen shot (figure 14). The help consists of hierarchical tree structure to help users to navigate and also a separate section on help for all information regarding the project and the site. Apart from this an interesting and user-friendly feature of the help pages is the ability to

display information "inline" – especially when forms are being used to gather user-typed input. This improves the functionality as well as interactivity of the user.

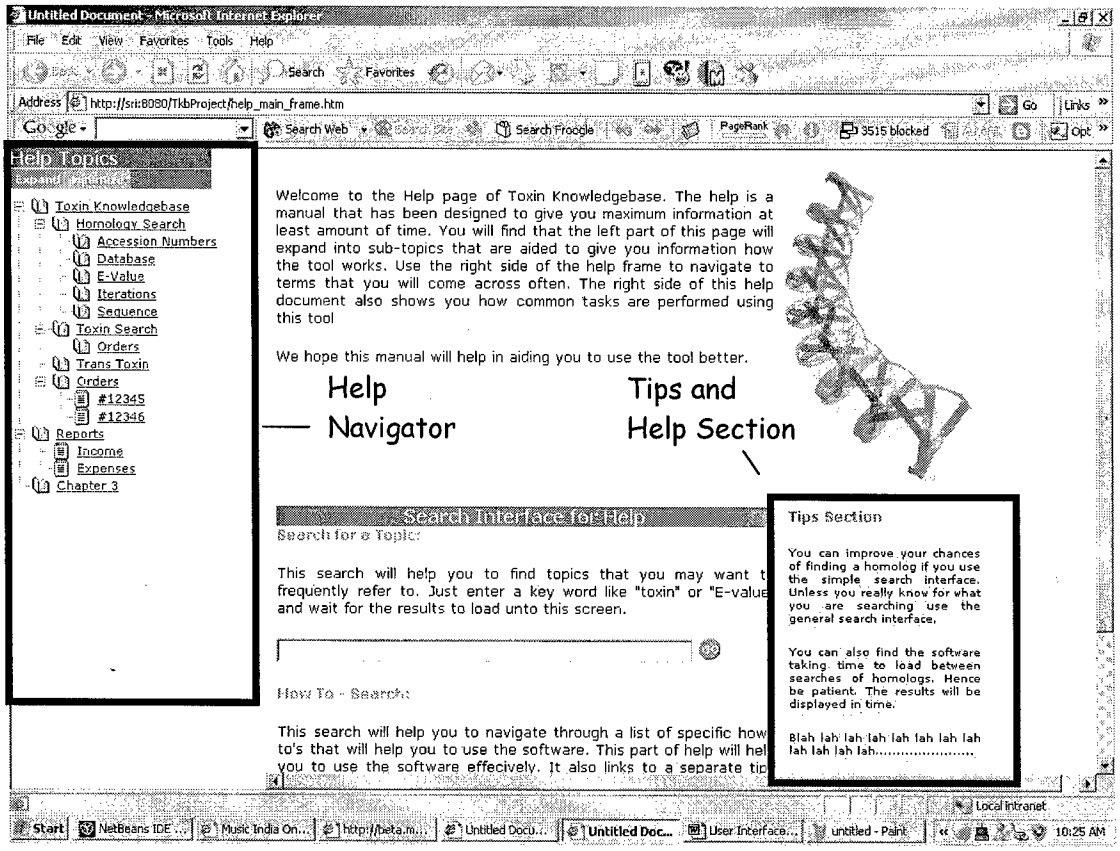


Figure 14: Help Interface Design: The help interface is designed to be user friendly having a tree that helps the user to navigate easily through out the help. The help interface also gives useful tips and strategies to use the site as a whole.

Case Studies

This section provides briefly some observed results validated by biochemical experiments, as well as some results that do not yet have a thorough biochemical validation. First we present the result of comparing Thermolysin, Neprilysin and Botulinum neurotoxin type E, of which Thermolysin and Neprilysin are non neurotoxic proteins, whereas Botulinum neurotoxin type E is a potent neurotoxin. Next we present the results of comparing Endoglucanase, with Chitinase of which Chitinase is a known toxin. For the former set of proteins, several studies have focused on the similarities of all the three proteins; however, more focus has been laid on the similarity of Thermolysin and Neprilysin which are known zinc binding proteases. The latter set, although a couple

of studies have been published, there has been no conclusive evidence through biochemical experiments that these two proteins are indeed similar.

Similarity of the Reaction Mechanism in Thermolysin, Neprilysin, and Botulinum neurotoxin Type E

Thermolysin, Neprilysin and Botulinum share a same motif HEXXH + E and it is speculated that they have similar reaction mechanism. The functional similarity of Thermolysin and Neprilysin has long been recognized due to a relatively significant sequence homology between the two proteins, as is shown in Figure 15-a. The statistics of the alignment is shown in the following table.

Length	333
Number of identical matches	37
Number of positive matches	89

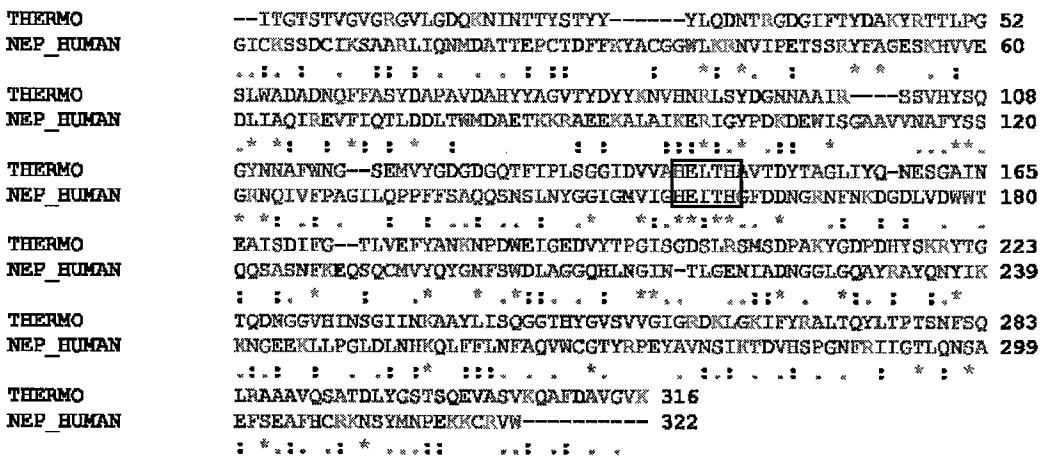


Figure 15-a. Homology between Thermolysin and Neprilysin

In contrast, the sequence homology between Thermolysin, Neprilysin and Botulinum is low, which can be seen from the alignment between Thermolysin and Botulinum neurotoxin serotype E in Figure 15-b. The statistics of the alignment is listed below.

Length	421
Number of identical matches	56

Number of positive matches	88

Because the alignment between Thermolysin and Botulinum does not indicate that they are closely related, one may conclude that they may not share any significant structural or functional similarity. Structural alignment of the full structures of Thermolysin, Neprilysin and Botulinum also fails to reveal the functional similarity between them. By concentrating on the active sites, the Trans-toxin (Mutoxin interface) is able to find that the active sites of Thermolysin and Neprilysin are similar to that of Botulinum, as shown in Figure 15-c, and therefore predicts that non-toxic proteins Thermolysin and Neprilysin might be mutated to function like Botulinum neurotoxin.

```
BoNT/E    PKINSFNYNDPVNDRTILYIKPGGCQEFYKSFNIMKNIWIIPERNVIGTTPQDFHPPTSL 60
THERMO    ------------------------------ITGTSTVGVGRGVLGDQ---------- 17

BoNT/E    KNGDSSYYDPNYLQSDEEKDRFLKIVTKIFNRINNNLSGGILLEELSKANPYLGNDNTPD 120
THERMO    KNITTTYSTYYYLQDN--------------TRGDGIFTYDAKYRTTLPGSLWADADN--- 60

BoNT/E    NQFHIGDASAVEIKFSNGSQDILLPNVIIMGAEPDLFETNSSNISLRNNYMPSNHRFGSI 180
THERMO    QFFASYDAPAVDAHYYAGVTYDYYKNVHNR-------LSYDGNNAAIRSSVHYS---QGYN 111

BoNT/E    AIVTFSPEYSFRFNDNCMNEFIQDPALTLMHELIHSLHGLYGAKGITTKYTITQKQNPLI 240
THERMO    NAFWNGSEMVYGDGDGQTFIPLSGGIDVVAHELTHAVTDYTAGLIYQNE---SGAINEAI 168

BoNT/E    TNIRGTNIEEFLTFGGTDLNIITSAQSNDIYTNLLADYKKIASKLSKVQVSNPLLNPYKD 300
THERMO    SDIFGTLVEFYANK------------NPDWEIGEDVYTPGISGDSLRSMSDPAKYGDPD 215

BoNT/E    VFEAKYGLDKDASGIYSVNINKFNDIFKKLYSFTEFDLRTKFQVKCRQTYIGQYKYFKLS 360
THERMO    HYSKRYTGTQDNG---GVHINSGIINKAAYLISQGGTHYGVSVVGIGRDKLGKIFYRALT 272

BoNT/E    NLLNDSIYNISEGYNINNLKVNFRGQNANLNPRIITPITGRGLVKKIIRFCKNIVSVKGI 420
THERMO    QYLTPTSNFSQLRAAAVQSATDLYGSTSQEVASVKQAFDAVGVK---------------- 316

BoNT/E    R 421
THERMO    -
```

Figure 15-b. Remote homology between Thermolysin and Botulinum E

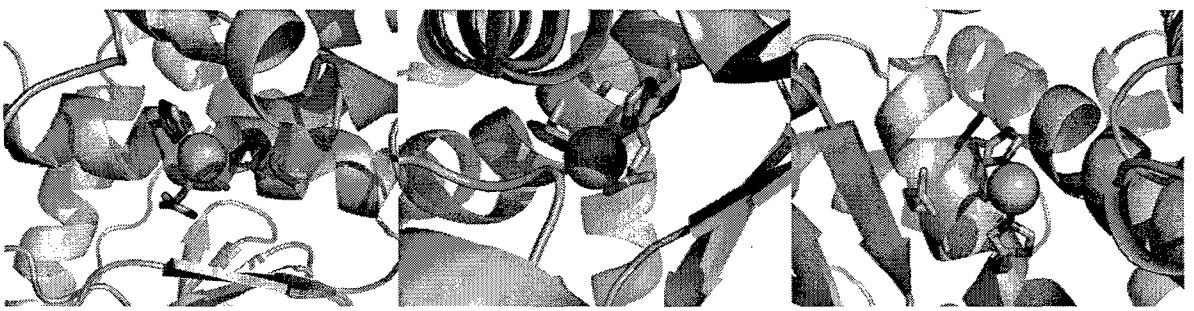


Figure 15-c(1): Active site of Botulinum neurotoxin E light chain (PDB:1T3A), residues shown H211, E212 and H215

Figure 15-c(2): Active site of human Neprilysin (PDB:1DMT), residues shown H583, E584 and H587

Figure15-c(3): Active site of Thermolysin (PDB:1TLP), residues shown H142, E143 and H146

Similar Reactive Mechanism of Endoglucanase and Chitinase

We chose the sequence of Endoglucanase since it is an important protein that binds to cellulose, as well as having a multidomain enzymatic characteristic. A sequence comparison between Endoglucanase (Swissprot ID: P10477) and Chitinase (PDB ID: 1HJX) does not give a good hint about their similarity, as is shown in Figure 15-d. The length of the alignment, the number of identical matches and the number of positive matches are as follows.

Length	696
Number of identical matches	74
Number of positive matches	94

Moreover, because no structure is available for Endoglucanase, one may stop at the sequence level and conclude that Endoglucanase do not share functional similarity with Chitinase without further structural analysis.

Trans-toxin tries to build a model of Endoglucanase based on its sequence homology with proteins whose structures have been determined using Modeller. Then at the structure level, a putative active site in Endoglucanase is similar to that of Chitinase, as shown in Figure 15-e thus reveals that Endoglucanase is potentially a candidate to be transformed to Chitinase.

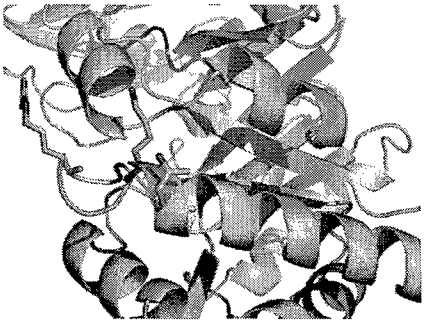


Figure 15-e (1): Active site of Chitinase (PDB: 1HJX), residues shown are R144, K147 and Q148

Figure 15-e (2): Putative active site of Endoglucanase E precursor (Swissprot: P10477), residues shown are R118, 20, and Q117.

Figure 15-d. Remote homology between Endoglucanase and Chitinase

Identifying Toxin Names and Interactions in Bio-Medical Abstracts

In addition developing the system we have embarked on text mining to identify new toxins from the available literature. A fully automated entity name extraction system, to identify toxins present in biomedical text has been developed. Our approach is based on identifying Sortal anaphors to extract proximal toxin names. We also extract protein-protein interactions related to the toxins talked about in the given abstract. Our extraction system handles complex sentences and extracts multiple and nested interactions specified in a sentence. Deliverables are toxin name list extracted from PubMed abstracts for the query "Toxin Survey" and protein interactions related to these toxins.

Key Research Accomplishments:

1. We have built a sophisticated Toxin Knowledge Base.
2. This can be used to identify and store homolog information.

3. A powerful tool WinAgent developed at Stony Brook University can be used to retrieve and collect data from various sources and incorporated into TKB.

4. Various links have been incorporated into TKB for easy use.

5. The TKB now contains molecular, structural and other information for over 1000 toxins.

6. TKB now stores homolog information for more than 500 toxins with an easy to use software to view the structural model.

7. Text mining has been developed to identify new toxins from web site literatures.

Reportable outcomes

One paper presented in a conference.

1. Arvind Ramanthan, Mike Kifer, I.V. ramakrishnan, Arvind Ramanathan, Chang Zhao, S. Jayaraman and S. Swaminathan . Toxin Knowledge Base: A system for discovering bioengineered threats. Presented as a poster in ISMB conference in Detroit, June 2005.

Conclusion

TKB has thus provided an engineering solution to a widely acknowledged problem of analyzing information from various resources by combining several off-the-shelf software tools and developing an integrated work-flow that offers biologists with the ability to analyze the nature of toxins. It also provides information to users if a non-toxin protein can be potentially transformed into a toxin using simple substitution of amino-acid residues at their active sites. It is also the single largest resource on information regarding toxins, where in biologists can easily synthesize and disseminate knowledge about toxins.

Apart from our engineering processes, our current research effort focuses on the development of methods to classify toxins into families based on profiles (using profile based Hidden Markov Models [8]). These models take into account information about variations even across distant homologs and can thus identify remotely related proteins and toxins. We are also investigating methods to build profiles of structures to compare active site information of proteins.

It is also equally important to enrich the knowledgebase with more knowledge about toxins. But, this process is generally not very simple and often unintuitive, because, the identification of even a single new toxin can mean that a biologist has to potentially

go through hundreds, perhaps even thousands of abstracts and scientific articles. Our text mining will solve this problem.

Future Plans

1. Some toxins do not have any structure information available. In this case, we can instead get a model for the toxin from MODBASE. For example, Tityustoxin ts3 (accession number P01496) does not have a PDB entry in Swiss-Prot, but it has a highly reliable model in MODBASE.

2. There are toxins for which we can not find the active site information in PDB, PDBSUM or LPC. We need to find alternative information sources of toxin active sites. Because active sites are often associated with structural pockets and cavities, one possible approach is to locate the active site with the help of CastP, which can provide identification and measurements of surface accessible pockets as well as interior inaccessible cavities.

3. Instead of building a model based on a pair-wise alignment and a toxin structure, we can build a model based on a multiple alignment of the homolog and several toxins to which it is homologous. This can help us to remove redundant models and improve the reliability of the models.

4. Hidden Markov Models have been used to build profiles of protein sequences of the same family. We want to extend the Profile Hidden Markov Model to the three-dimensional space and use it to model the active sites of toxins from the same family. Instead of comparing each active site with the target protein, we just need to compare each active site profile with the target protein.

5. The user will be allowed to submit a new toxin. The new toxins will be studied and it may become an entry in the TKB.

References

1. Using a Library of Structural Templates to Recognize Catalytic Sites and Explore their Evolution in Homologous Families. J. Mol. Biol. Vol 347 2005, pages 565-581
2. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics Vol 19 no. 13 2003, pages 1644-1649

3. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research Vol 32, Database issue D217-D222, 2004.
4. Nikeeta Julasana, Akshat Khandelwal, Anupama Lolage, Prabhdeep Singh, Priyanka Vasudevan, Hasan Davulcu, I. V. Ramakrishnan: WinAgent: a system for creating and executing personal information assistants using a web browser. Intelligent User Interfaces 2004: 356-357
5. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
6. G.J. Kleywegt (1999). Recognition of spatial motifs in protein structures. J Mol Biol 285, 1887-1897.
7. Sali & T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.
8. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Cambridge Publications, 1998.

Personnel in the Project

1. S. Swaminathan (PI)	Scientist	20% effort
2. S. Jayaraman	Sr. Research Associate	40% effort

Sub-contract to State University of New York at Stony Brook

1. Mike Kifer	Professor	10% effort
2. I.V. Ramakrishnan	Professor	10% effort

Two Ph.D. (50%) student and 12 M.S. students (all part time 20 to 50%)

Sub-contract to Arizona State University, Tempe, Arizona

1. H. Davulcu	Asst. Professor	10% effort

Seven M.S. students (part time)

A copy of the poster presentation is attached. Since it was prepared by Powerpoint, the two pages actually should come side by side.

Toxin Knowledge Base: A System

Michael Kifer, I. V. Ramakrishnan, Arvind Ramanathan, Chang Zhao

Department of Computer Science, Computer Science Building,
Stony Brook University, Stony Brook, NY 11794.

Abstract

Recent developments in recombinant DNA technology has given rise to the possibility of producing bioengineered pathogens like toxins and other products on scales that could make them into formidable weapons of bioterrorism. Yet another kind of threat is through chimeric molecules, where in the virulent domain of a toxin is hidden in a non-pathogenic protein. The Toxin Knowledge Base (TKB) has been established as a bioinformatics resource to tackle the problem of identifying potential bio-warfare agents as well as chimeric proteins. It is a tool that can be used to assimilate, synthesize, analyze and disseminate genomic and structural information on biological and potential biological warfare agents, identify and develop counter measures such as vaccines, antitoxins and inhibitors and also understand the mode of actions of these toxins at the cellular, sub-cellular, and molecular levels. The system has been designed using a novel workflow mechanism and is seamlessly integrated using an easy-to-use web based portal.

System Description

The system has been developed with an emphasis on the end user's perspective in mind. Figure 1 shows the system architecture of TKB from that perspective. The Query interfaces comprise of a sophisticated inference engine, based on derivable knowledge in terms of the structures of toxins that exist within the database. This knowledge is used as a means to design the Trans-toxin workflow, whose schematic is shown in figure 2. Once a user inputs a sequence, a search against the sequences in the RCSB yields the answer to the question whether the input sequence has a structure. If the sequence does not have a structure, the input sequence is first 'BLAST'ed against the TKB sequence information to determine whether it is possible to model the sequence using the structural information in the knowledge base. If so, it is passed on to the Modeller program, which provides a three dimensional model, which is then compared against the active site templates using SPASM and the results are tabulated to the end user.

The administrator's interface is based on the fact that he spends most of his time in keeping the knowledge base up-to-date. The data acquisition and curation workflow is shown in figure 3.

The data for toxins lies embedded within several sources. Principal sources for the current work include public domain databases like SWISSPROT, EMBL and RCSB. Using a specially built tool called WinAgent, the data from these various sites are mined and assembled as XML data. This XML data is then processed using an XSLT to merge into the Toxin Knowledgebase. This provides a simple and extensible mechanism for mining toxin data from the internet. The tool created for this purpose (a modified version of WinAgent) is very easy to use, such that even a novice user can create powerful agents to mine for toxin data. This tool has been embedded within the design of the system such that it is transparent to the user and hence, the user can simply create agents by specifying a specific website and click on a few instances that he wants the agent to pick up (during the training phase), and then automatically execute the code to fetch similar sets of data from the site and then integrate it into the toxin repository.

This has allowed us to develop a simple, yet powerfully extensible system, that can be potentially applied to not only toxins, but also different classes of recognized proteins.

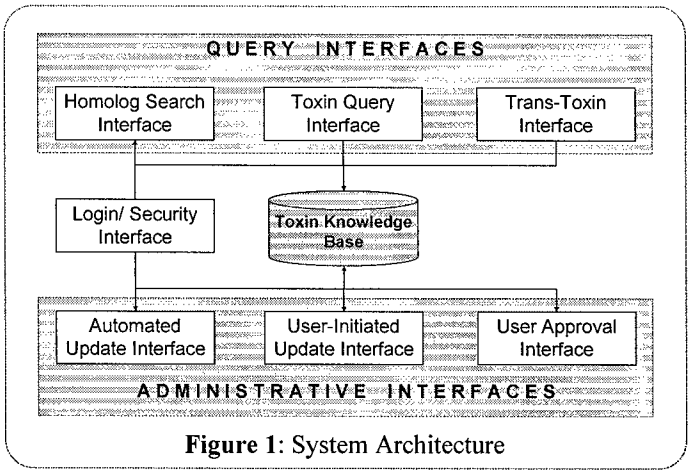


Figure 1: System Architecture

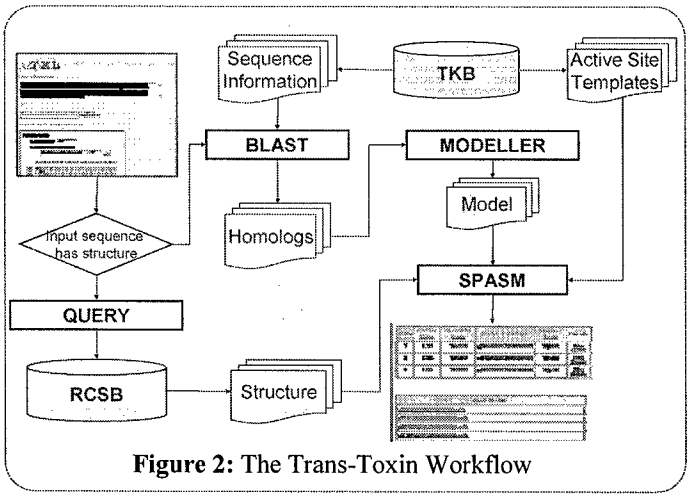


Figure 2: The Trans-Toxin Workflow

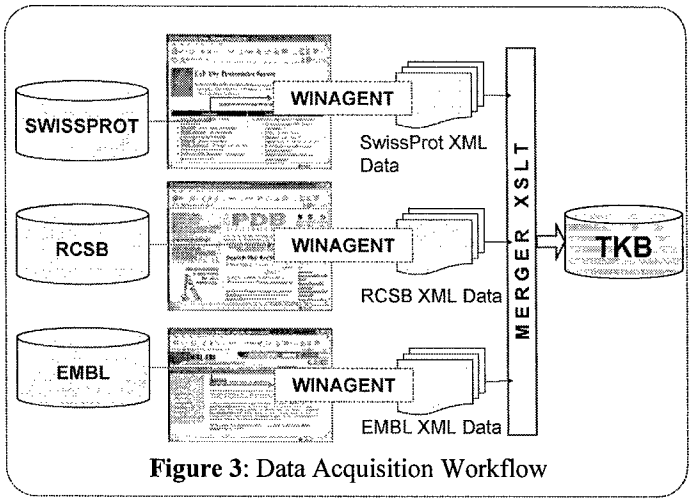


Figure 3: Data Acquisition Workflow

for Discovering Bio-Engineered Threats

Seetharaman Jayaraman, Subramanyam Swaminathan

Biology Department, Brookhaven National Laboratory, Upton, NY 11973

BROOKHAVEN NATIONAL LABORATORY

Results

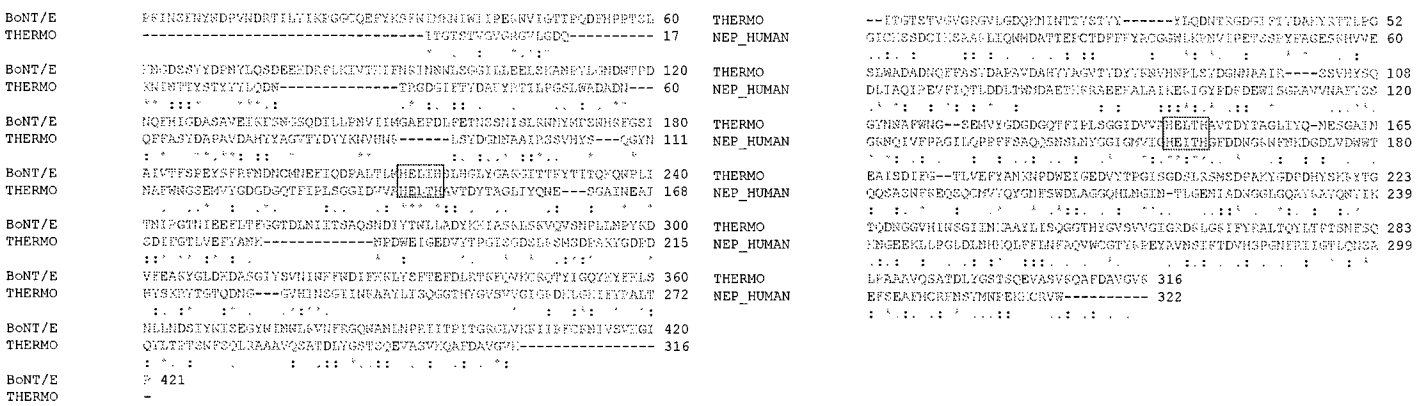


Figure 4: Sequence comparison between botulinum neurotoxin serotype E Light Chain (BoNT/E), thermolysin (THERMO), and neprilysin (NEP_HUMAN), similar to a report presented in Toxin Knowledgebase. One can observe that, BoNT/E and THERMO are homologs; THERMO and NEP_HUMAN are homologs, and share the same active site motif (Zinc binding motif: HEXXH).

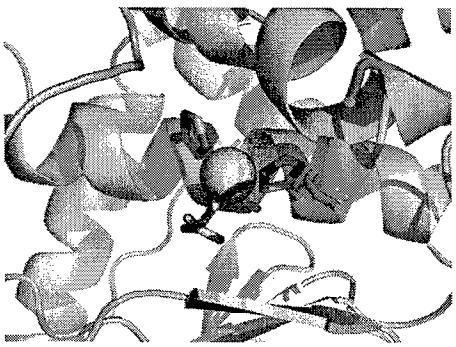


Figure 5 † (a): Active Site of botulinum neurotoxin E Light Chain (PDB: 1T3A), residues shown are H211, E212 and H215.

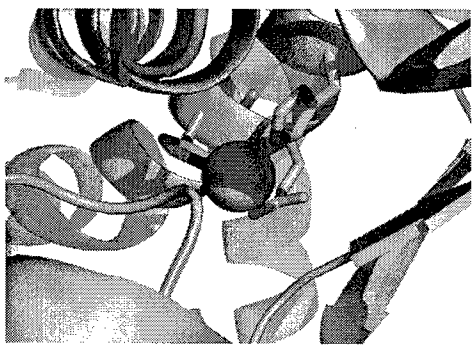


Figure 5 † (b): Active Site of human neprilysin (PDB: 1DMT), with residues H583, E584 and H587.

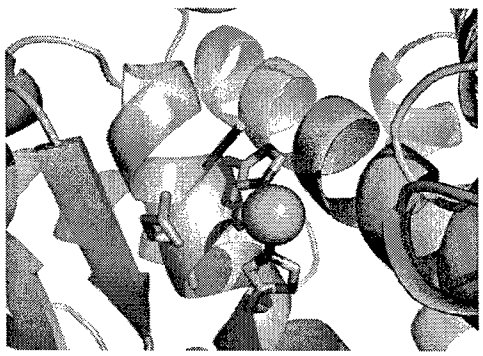


Figure 5 † (c): Active Site of thermolysin (PDB: 1TLP), residues H142, E143 and H146.

† Figures generated using PyMOL Molecular Graphics software, (http://www.pymol.org).

The initial results from using TKB are shown in figures 4 and 5. Figure 4 presents a sequence comparison of botulinum neurotoxin, thermolysin and neprilysin. As shown in the figure, all the three proteins share the same motif – HEXXH. However, one must note that botulinum neurotoxin serotype E and neprilysin are rather distantly related homologs. They do not show a high sequence similarity and hence one may conclude that the two may not share any significant structural or functional similarity.

If one follows the workflow illustrated in figure 2, the output from TKB will show that the active sites of the three proteins to be very similar. The arrangement of the residues at the active site and the coordination of the Zinc ion shows that all the three proteins not only share the same motif, but also that these proteins could be functionally similar.

Conclusions and Future Work

The current work represents a successful prototype for relating sequence, structure and functional aspects of proteins. Sequence based comparison methods although valuable, can allow detection of related families of proteins, but in order to relate the functions of proteins better, better structure based methods are necessary. In the future, the authors plan to develop three dimensional structural profiles, that would be used to identify remote homologs, that may be functionally related to various proteins. It is also planned to expand the knowledgebase and allowing the bioinformatics community to use this web-based service.

References

1. Nikeeta Julasana, Akshat Khandelwal, Anupama Lolage, Prabhdeep Singh, Priyanka Vasudevan, Hasan Davulcu, I. V. Ramakrishnan. WinAgent: a system for creating and executing personal information assistants using a web browser. Intelligent User Interfaces 2004: 356-357.

2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-410.

3. G.J. Kleywegt (1999). Recognition of spatial motifs in protein structures. J Mol Biol 285, 1887-1897.

4. A. Sali & T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.

Acknowledgements

The authors would like to thank Sridhama Vempaty, Padmavathy Malligarjunan, Jigar Mehta, Darshan Ramavat, Venkat Reddy Nukala, Sheetal Shah, Mayur Shetty, Debashish Rawal and Pooja Virkud for their help in developing the system. The grant from US Army Medical Research Acquisition Activity (Contract No.: **DAMD17-03-1-0520**) is also gratefully acknowledged.