AD________________

Award Number: DAMD17-03-1-0520

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

PRINCIPAL INVESTIGATOR: Subramanyam Swaminathan, Ph.D.

CONTRACTING ORGANIZATION: Brookhaven National Laboratory
Upton, NY 11973

REPORT DATE: August 2004

TYPE OF REPORT: Annual

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

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY **(Leave blank)**	**2. REPORT DATE** August 2004	**3. REPORT TYPE AND DATES COVERED** Annual (1 Aug 2003 - 31 Jul 2004)

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

5. FUNDING NUMBERS
DAMD17-03-1-0520

6. AUTHOR(S)

Subramanyam Swaminathan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Brookhaven National Laboratory
Upton, NY 11973

E-Mail: swami@bnl.gov

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. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Original contains color plates: ALL DTIC reproductions will be in black and white

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

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

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.

14. SUBJECT TERMS

Toxin data base, toxin homologs, bioengineered threat

15. NUMBER OF PAGES
29

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

Table of Contents

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

Annual Report for the Period ending July 2004

Introduction

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. TKB will be designed and developed to serve as a powerful analysis and decision system tool for the intelligence community. 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.

Body

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

In the first year we have modified, improved and expanded the previously existing data base for storing, managing and accessing molecular information on known as well as potential biological toxins. A detailed description of the new system follows. This section describes our work, progress and deliverables as given in Specific Tasks 1, 2 and 3.

1. Overview

Figure 1 is a representation of the Toxin Knowledge Base system and the interfaces it provides to the end user. The various interfaces and their uses are described below.

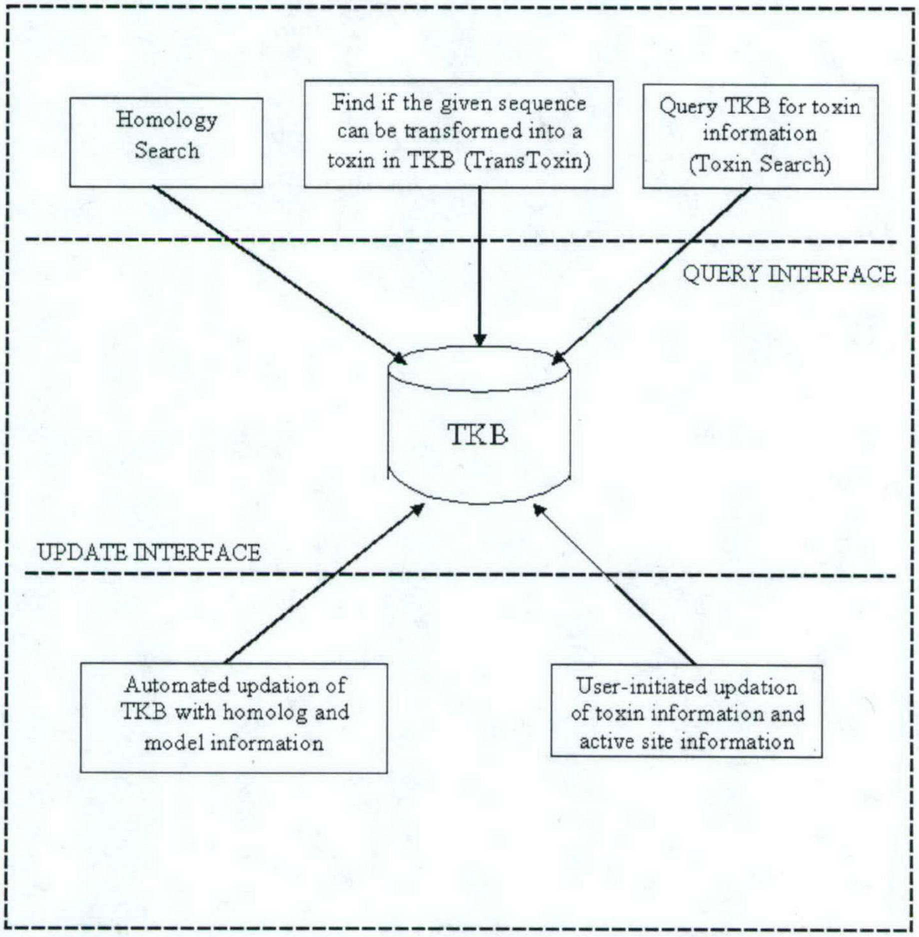

Homology Search
Find if the given sequence can be transformed into a toxin in TKB (TransToxin)
Query TKB for toxin information (Toxin Search)
QUERY INTERFACE
TKB
UPDATE INTERFACE
Automated updation of TKB with homolog and model information
User-initiated updation of toxin information and active site information

Figure 1: Project Architecture: The figure shows the architecture of the system.

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 as well as the models for the homologs. It provides two interfaces to the user namely:

1. **Query Interface**. 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. Trans Toxin - Determining whether a protein can be transformed into a toxin

2. **Update Interface**. 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.

2. Users of the System

The user interface for the Toxin Knowledgebase (TKB) is primarily user driven. The primary users for the system are scientists, biologists, and chemists. In addition, there is a database administrator and a system administrator who monitors and maintains the system as the need arises. A subcategory of biologists and chemists has special privileges. These users can perform additional tasks, such as scheduling regular updates to the TKB, running special programs (such as BLAST), and view the status of other users. The user interface for the system was designed with these requirements in mind. A hierarchy of users within the system is shown below in Figure 2.

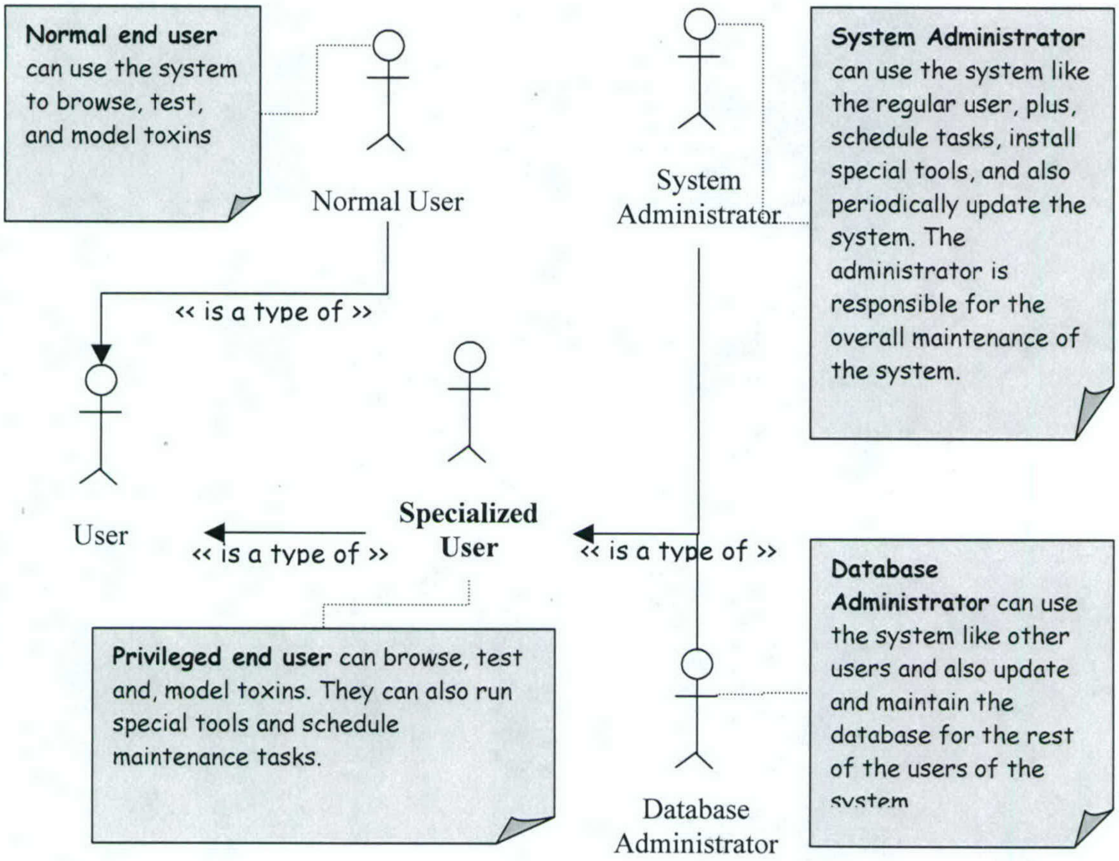

Normal end user can use the system to browse, test, and model toxins
Normal User
<< is a type of >>
User
<< is a type of >>
Specialized User
<< is a type of >>
System Administrator
System Administrator can use the system like the regular user, plus, schedule tasks, install special tools, and also periodically update the system. The administrator is responsible for the overall maintenance of the system.
Privileged end user can browse, test and, model toxins. They can also run special tools and schedule maintenance tasks.
Database Administrator
Database Administrator can use the system like other users and also update and maintain the database for the rest of the users of the system

Figure 2: User Hierarchy for Toxin Knowledgebase (TKB): A schematic view of various users and how they use the system.

TKB recognizes the following types of users:

- **Normal User:** Most of the users are of this type. They can browse TKB to find toxins, use the modeler to model toxins, and query the knowledge base for relevant information.
- **Privileged User:** A subset of the users has certain privileges within the system. Apart from using the system in the normal mode, they can also schedule tasks and use additional tools.
- **System Administrator:** A system administrator is responsible for the overall day-to-day maintenance of the system. Typically there would be one or two system administrators.
- **Database Administrator:** The database administrator is responsible for maintaining the database that forms the core of the system.

The user interface is designed to be:

- **Simple and Easy-to-use:** The user interface does not require much time to learn the basic functions of a system.
- **Uniform and Modular:** The user interface has a consistent and uniform look and feel on most available Web browsers. It is also easy to modify and reuse existing user interface elements, so additional features may be easily implemented
- **Platform Independent and Universally Accessible:** All users must be able to access the information in a format that is independent of any particular platform and must be globally accessible.

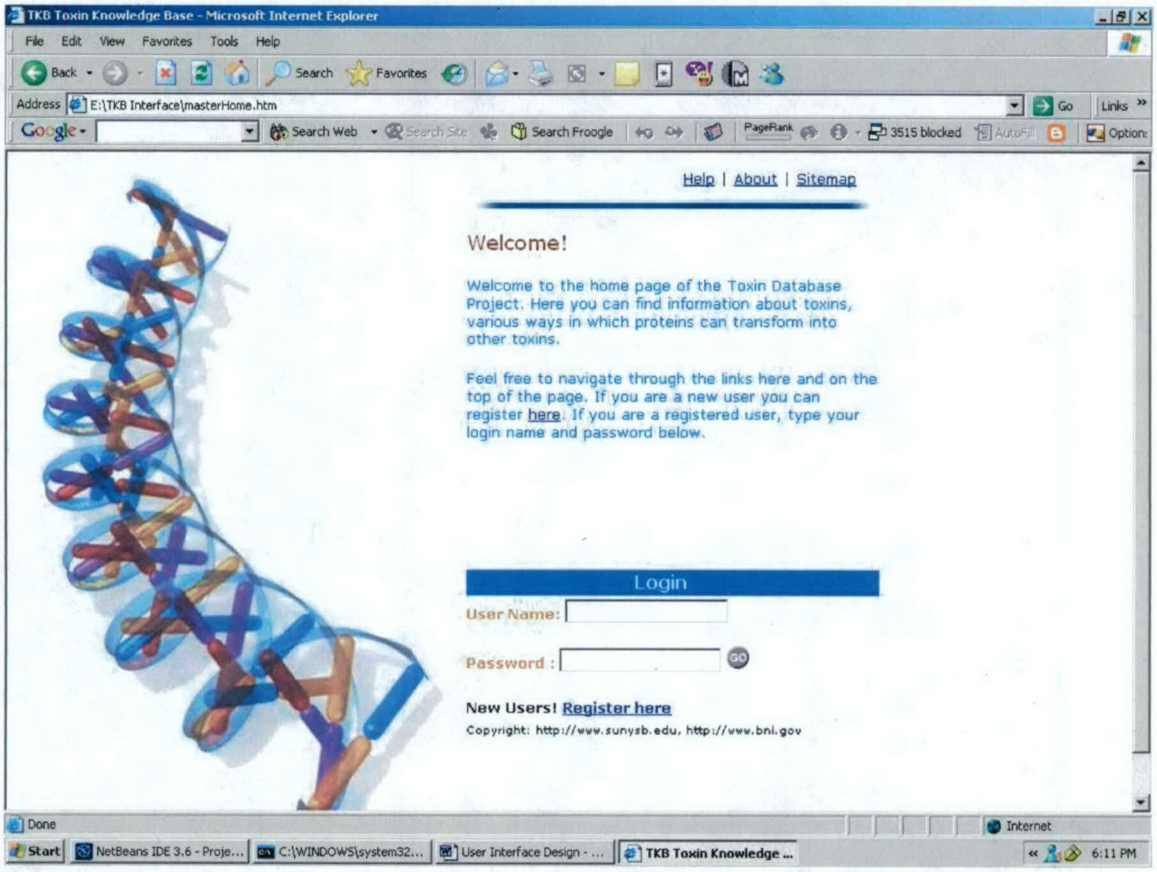
3. Project Details

This section describes the project components outlined in Figure 1. It also provides screen shots of the current state of the system, which should help better understanding of the features that have been incorporated into the system.

3.1. 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. All users need to log into the system. The login interface authenticates the user through a user-name/password combination and then determines the set of screens the user is authorized to access.

A screen shot of the login screen is shown in Figure 3.


TKB Toxin Knowledge Base - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address E:\TKB Interface\masterHome.htm
Go Links
Google Search Web Search Site Search Froogle PageRank 3515 blocked AutoFill Options
Help | About | Sitemap
Login
GO
Done
Internet
Start NetBeans IDE 3.6 - Proje... C:\WINDOWS\system32... User Interface Design - ... TKB Toxin Knowledge ...
6:11 PM

Welcome!

Welcome to the home page of the Toxin Database Project. Here you can find information about toxins, various ways in which proteins can transform into other toxins.

Feel free to navigate through the links here and on the top of the page. If you are a new user you can register here. If you are a registered user, type your login name and password below.

User Name:

Password :

New Users! Register here

Copyright: http://www.sunysb.edu, http://www.bnl.gov

Figure 3: Login Screen: The login screen is a simple welcome page that allows the user to log in. A separate form is used for new users who want to register. However, new users can log in only after they have been screened by the system administrator.

3.2. The Toxin Database

- The toxin database stores comprehensive information about the toxins. The information is collected from various online toxin databases such as Swiss-Prot (http://us.expasy.org/) and the Protein Data Bank (PDB) (http://www.rcsb.org/). These sites are worldwide repositories for the processing and distribution of 3-D biological macromolecular structure data. TKB stores all this data locally in order to make the access more efficient.
- The toxin features such as PDB ID, toxin name, synonym, cellular location, molecular weight, scientific name, inhibitors, biochemical information etc. are stored in the 'CONTENT' field of the database table. Many of the details are retrieved by TKB from various other sources also.

The detailed schema of the database table is as follows:

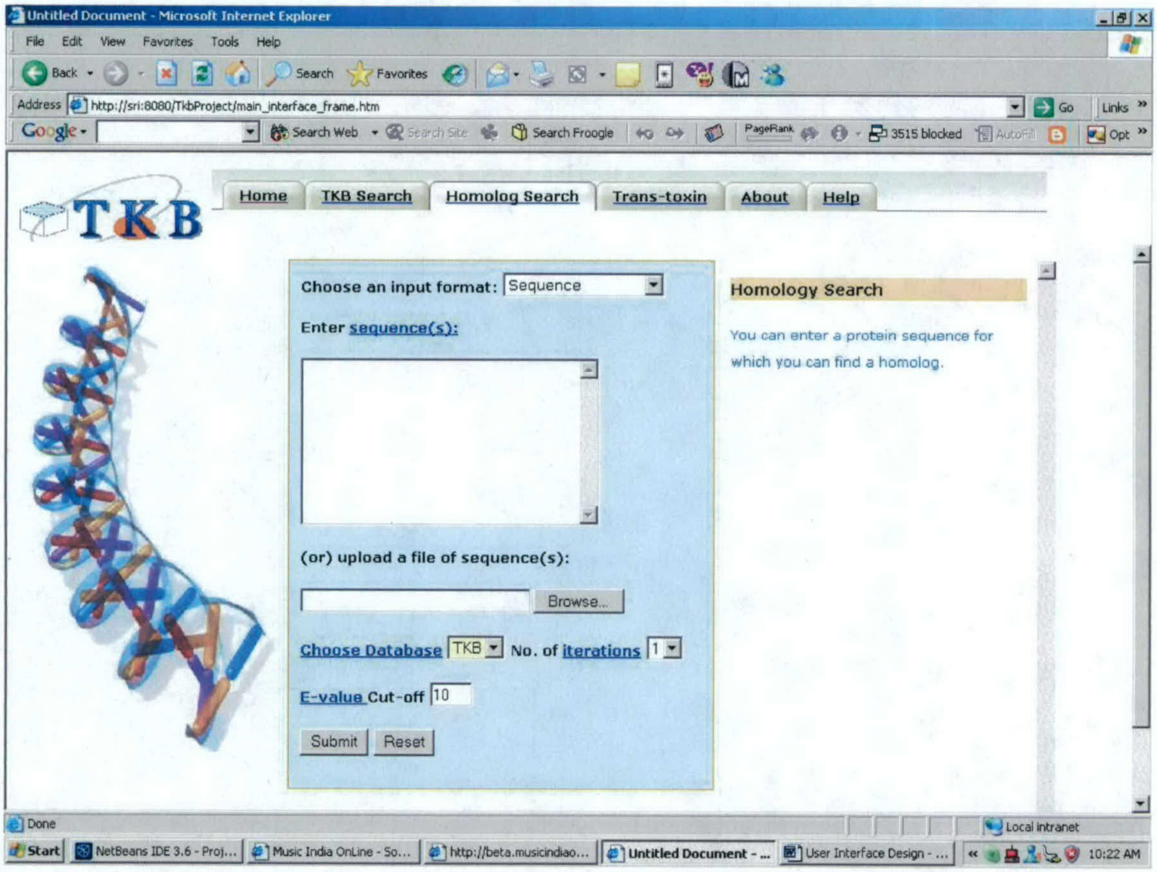
Field	**Purpose**
ID	Unique number to identify each toxin
CONTENT	Stores the toxin information
ACCESSION NUMBER	For information retrieval from Swiss-Prot
FASTA SEQUENCE	To find homologs
ACTIVE SITE ID	Used in trans-toxin queries
HOMOLOGS	Store homolog information
MODELS	Stores models of homologs

3.3. Query Interface

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

3.3.1. 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 (shown in Figure 4) 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)


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://sri:8080/TkbProject/main_interface_frame.htm
Go Links
Google Search Web Search Site Search Froogle PageRank 3515 blocked AutoFill Opt
TKB
Home
TKB Search
Homolog Search
Trans-toxin
About
Help
Browse...
Submit
Reset
Done
Local intranet
Start
NetBeans IDE 3.6 - Proj...
Music India OnLine - So...
http://beta.musicindiao...
Untitled Document - ...
User Interface Design - ...
10:22 AM

Choose an input format: Sequence

Enter sequence(s):

(or) upload a file of sequence(s):

Choose Database TKB No. of iterations 1

E-value Cut-off 10

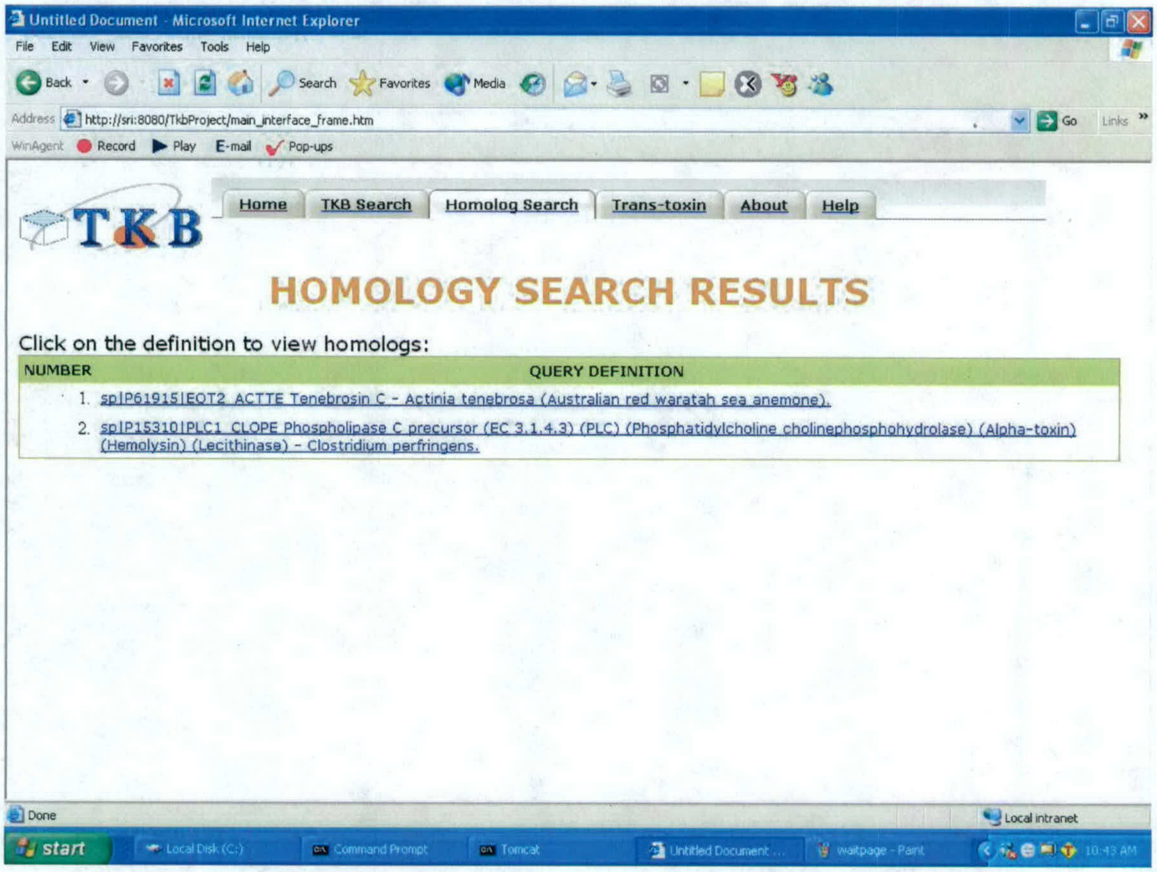
Homology Search

You can enter a protein sequence for which you can find a homolog.

Figure 4: Homology Search Interface: This interface helps the user to find a homolog based on a sequence or an accession number. The right hand panel shows how help can be dynamically provided depending on the links found on the form. This allows all users with minimal knowledge of the system to understand the terms and use the system with little 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 expected 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 shown in Figure 5.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://sri:8080/TkbProject/main_interface_frame.htm
Go Links
WinAgent Record Play E-mail Pop-ups
TKB
Home
TKB Search
Homolog Search
Trans-toxin
About
Help
Done
Local intranet
start
Local Disk (C:)
Command Prompt
Tomcat
Untitled Document ...
waitpage - Paint
10:43 AM

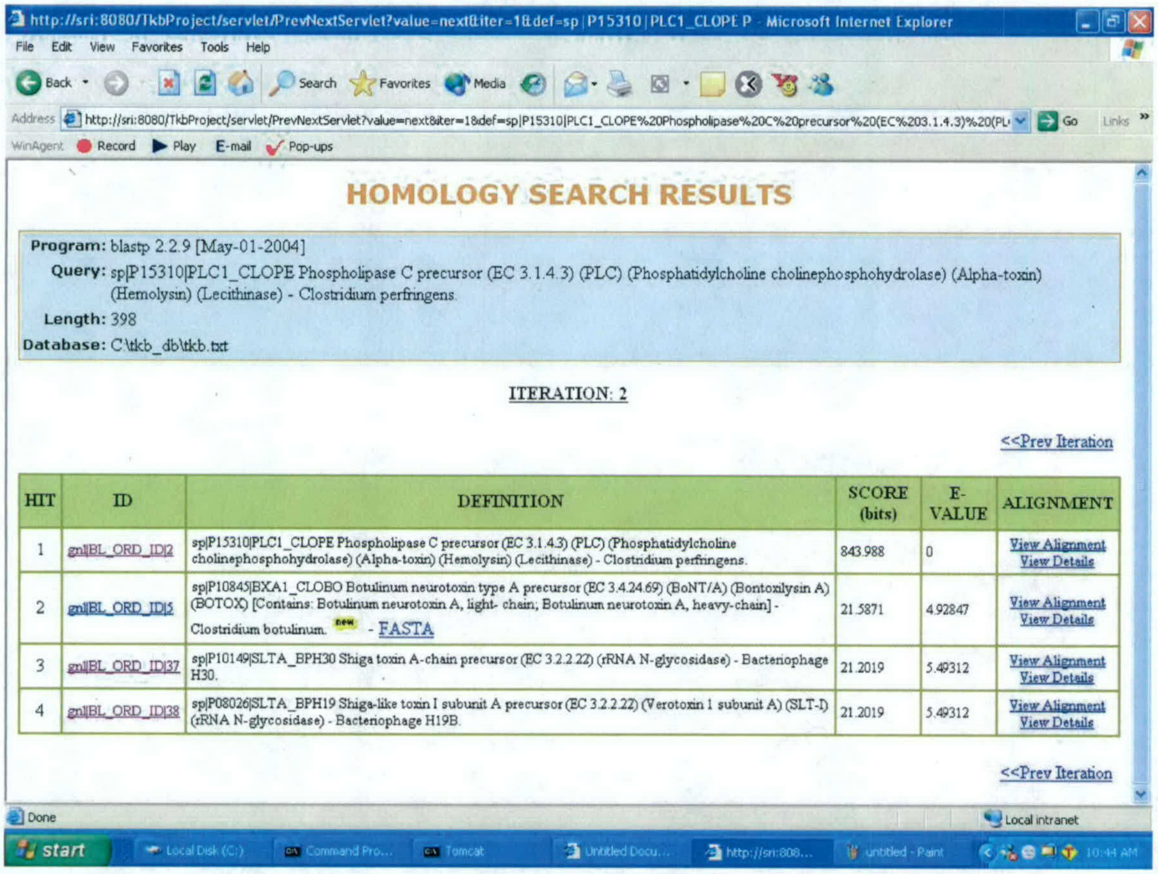
HOMOLOGY SEARCH RESULTS

Click on the definition to view homologs:

NUMBER	QUERY DEFINITION
1.	sp|P61915|EOT2_ACTTE Tenebrosin C - Actinia tenebrosa (Australian red waratah sea anemone).
2.	sp|P15310|PLC1_CLOPE Phospholipase C precursor (EC 3.1.4.3) (PLC) (Phosphatidylcholine cholinephosphohydrolase) (Alpha-toxin) (Hemolysin) (Lecithinase) - Clostridium perfringens.

Figure 5: Results for homology search: The initial results include the links, which take the user to the actual homologs that were found by the query.

- The list of homologs in a page-wise format, iteration by iteration, is shown in Figure 6.


http://sri:8080/TkbProject/servlet/PrevNextServlet?value=next&iter=1&def=sp|P15310|PLC1_CLOPE P - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://sri:8080/TkbProject/servlet/PrevNextServlet?value=next&iter=1&def=sp|P15310|PLC1_CLOPE%20Phospholipase%20C%20precursor%20(EC%203.1.4.3)%20(PL
Go Links
WinAgent Record Play E-mail Pop-ups
Done
Local intranet
start Local Disk (C:) Command Pro... Tomcat Untitled Docu... http://sri:808... untitled - Paint 10:44 AM

HOMOLOGY SEARCH RESULTS

Program: blastp 2.2.9 [May-01-2004]
Query: sp|P15310|PLC1_CLOPE Phospholipase C precursor (EC 3.1.4.3) (PLC) (Phosphatidylcholine cholinephosphohydrolase) (Alpha-toxin) (Hemolysin) (Lecithinase) - Clostridium perfringens.
Length: 398
Database: C:\tkb_db\tkb.txt

ITERATION: 2

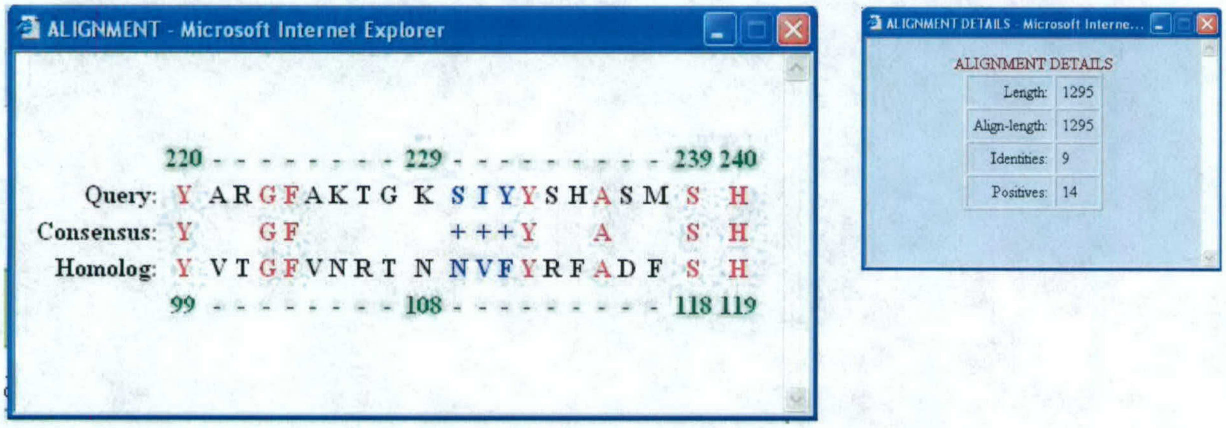
<<Prev Iteration

HIT	ID	DEFINITION	SCORE (bits)	E-VALUE	ALIGNMENT
1	gnl|BL_ORD_ID|2	sp|P15310|PLC1_CLOPE Phospholipase C precursor (EC 3.1.4.3) (PLC) (Phosphatidylcholine cholinephosphohydrolase) (Alpha-toxin) (Hemolysin) (Lecithinase) - Clostridium perfringens.	843.988	0	View Alignment View Details
2	gnl|BL_ORD_ID|5	sp|P10845|BXA1_CLOBO Botulinum neurotoxin type A precursor (EC 3.4.24.69) (BoNT/A) (Bontoxilysin A) (BOTOX) [Contains: Botulinum neurotoxin A, light- chain; Botulinum neurotoxin A, heavy-chain] - Clostridium botulinum. new - FASTA	21.5871	4.92847	View Alignment View Details
3	gnl|BL_ORD_ID|37	sp|P10149|SLTA_BPH30 Shiga toxin A-chain precursor (EC 3.2.2.22) (rRNA N-glycosidase) - Bacteriophage H30.	21.2019	5.49312	View Alignment View Details
4	gnl|BL_ORD_ID|38	sp|P08026|SLTA_BPH19 Shiga-like toxin I subunit A precursor (EC 3.2.2.22) (Verotoxin 1 subunit A) (SLT-I) (rRNA N-glycosidase) - Bacteriophage H19B.	21.2019	5.49312	View Alignment View Details

<<Prev Iteration

Figure 6: Results for Homology Search: The result of each query is shown on top. Homologs are shown in the table at the bottom.

- 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:
     - A score that represents the degree of similarity with the query toxin
     - The E-value
     - A link to pair-wise alignment of the query toxin and the homolog sequence. An example of such an alignment is shown in Figure 7
     - A link to alignment details (see Figure 7)



ALIGNMENT - Microsoft Internet Explorer

```
           220 - - - - - - - - 229 - - - - - - - - - 239 240
    Query: Y ARGFAKTG K SIYYSHASM S H
Consensus: Y   GF          +++Y   A   S H
  Homolog: Y VTGFVNRT N NVFYRFADF S H
           99 - - - - - - - - 108 - - - - - - - - - 118 119
```

ALIGNMENT DETAILS - Microsoft Interne...

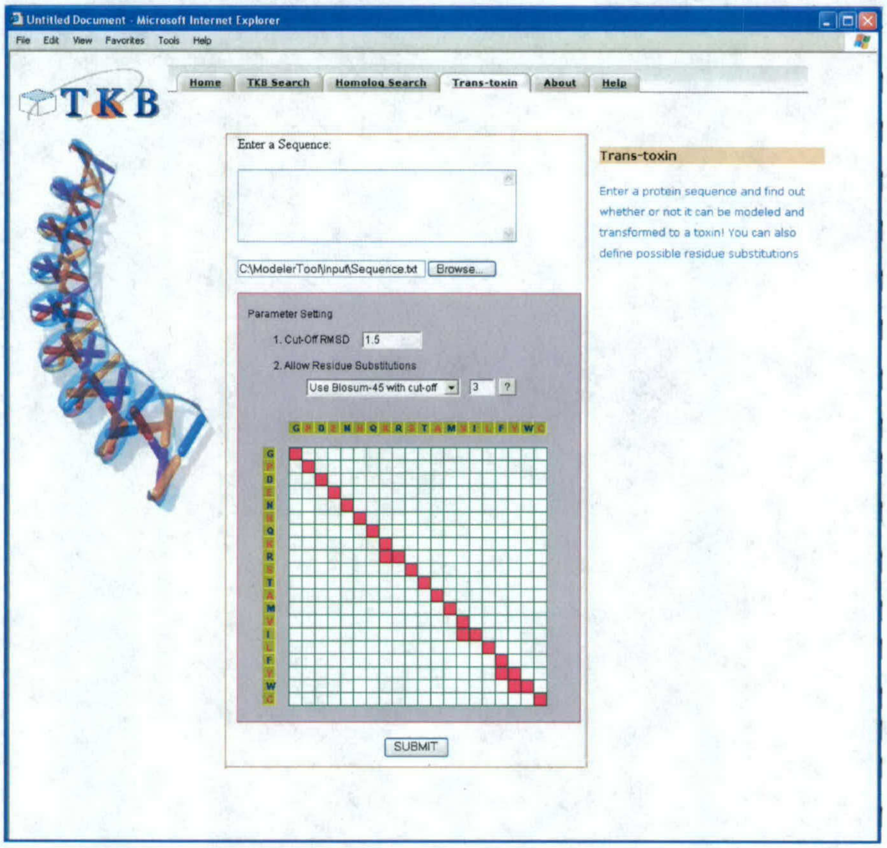
ALIGNMENT DETAILS

Length:	1295
Align-length:	1295
Identities:	9
Positives:	14

Figure 7: Pair-wise alignment and Details: Pair-wise alignment shows the identities (the residues marked red) and the positives (the residues marked blue). The alignment details are shown on the right.

3.3.2. Trans-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 as a text string. The user interface is shown in Figure 8.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
TKB
Home
TKB Search
Homolog Search
Trans-toxin
About
Help
Browse...
G P D E N Q K R S T A M V I L F Y W C
SUBMIT

Enter a Sequence:

C:\ModelerTool\input\Sequence.txt

Parameter Setting

1. Cut-Off RMSD 1.5

2. Allow Residue Substitutions

Use Blosum-45 with cut-off 3 ?

Trans-toxin

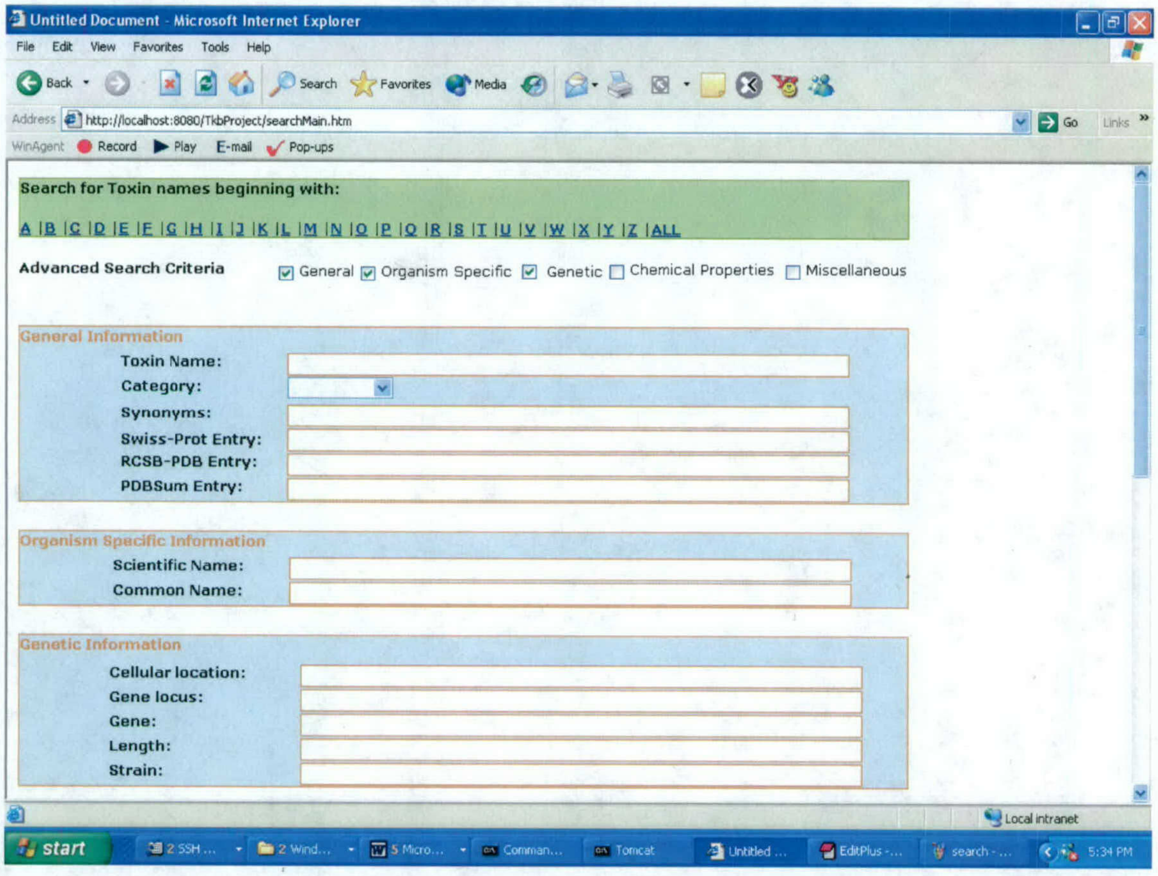
Enter a protein sequence and find out whether or not it can be modeled and transformed to a toxin! You can also define possible residue substitutions

Figure 8: TheTrans-toxin Interface: The user can provide the RMSD cut-off and select a residue substitution matrix. The user is also allowed to define possible residue substitutions through an interactive web page.

- PSI-BLAST is used to find sequences in TKB with more than 30% identity to the target protein, which might be toxins originally stored in TKB or homologs of toxins that are accumulated during the automated update process.
- MODELLER is then used to build models for the target protein.
- For each active site stored in TKB, the program SPASM (from Uppsala Software Factory) is run, to see whether there are sub-structures in the models of the target protein that match the active site. A sub-structure matches the active site in the sense that after super-positioning, the RMSD between the substructures is no more than the cut-off RMSD.
- Residue substitutions are allowed during super-positioning. Default RMSD cut-off and residue substitution matrix are provided. Users can specify their own RMSD cut-offs as well as the permitted residue substitutions.

3.3.3. TKB Search

- The TKB Search interface is useful for viewing 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 9 below.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://localhost:8080/TkbProject/searchMain.htm
Go Links
WinAgent Record Play E-mail Pop-ups
Local intranet
start 2 SSH ... 2 Wind... 5 Micro... Comman... Tomcat Untitled ... EditPlus -... search -...
5:34 PM

Search for Toxin names beginning with:

A |B |C |D |E |F |G |H |I |J |K |L |M |N |O |P |Q |R |S |T |U |V |W |X |Y |Z |ALL

Advanced Search Criteria ☑ General ☑ Organism Specific ☑ Genetic □ Chemical Properties □ Miscellaneous

General Information

Toxin Name:
Category:
Synonyms:
Swiss-Prot Entry:
RCSB-PDB Entry:
PDBSum Entry:

Organism Specific Information

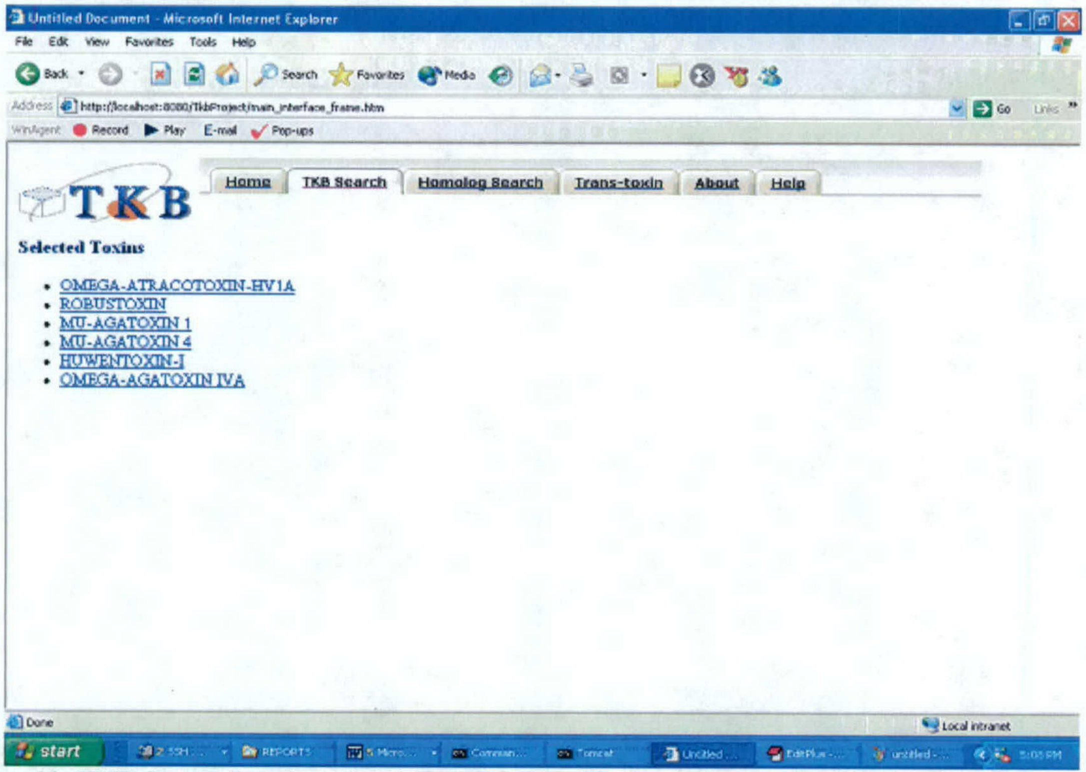
Scientific Name:
Common Name:

Genetic Information

Cellular location:
Gene locus:
Gene:
Length:
Strain:

Figure 9: TKB Search Interface: The figure shows the available search options

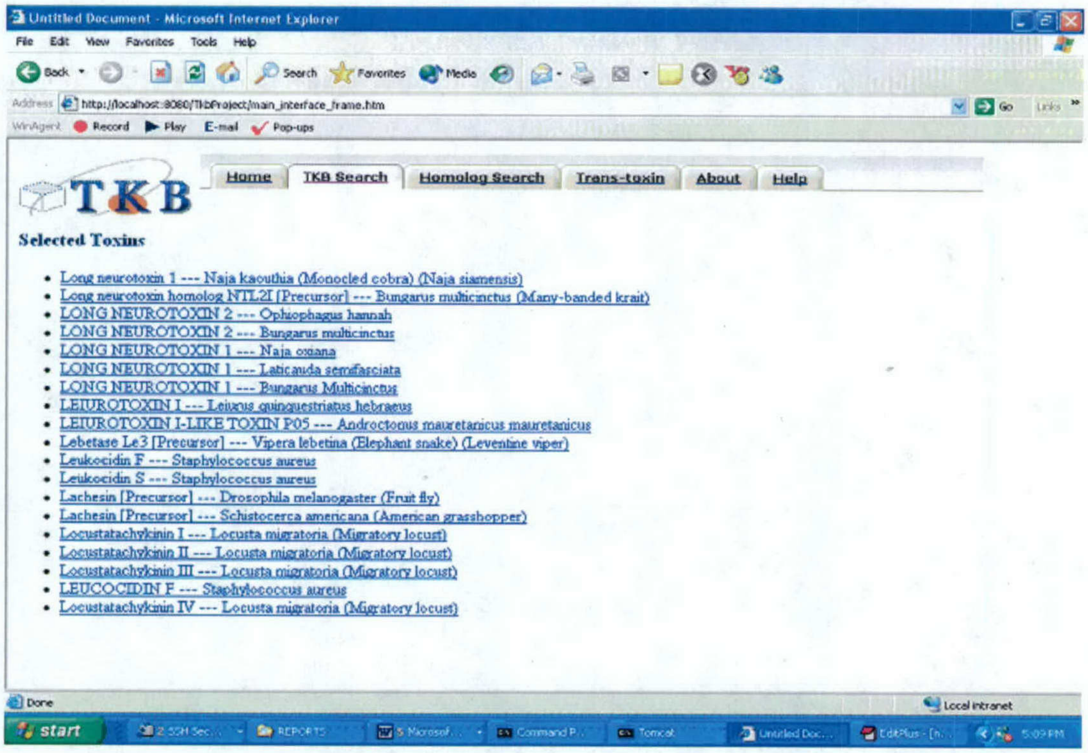
The search results containing the names of the toxins satisfying the chosen filter criteria are displayed as shown in the Figures 10 through 13.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://localhost:8080/TkbProject/main_interface_frame.htm
Go Links
WinAgent Record Play E-mail Pop-ups
TKB
Home
TKB Search
Homolog Search
Trans-toxin
About
Help
Done
Local intranet
start

Selected Toxins

- OMEGA-ATRACOTOXIN-HV1A
- ROBUSTOXIN
- MU-AGATOXIN 1
- MU-AGATOXIN 4
- HUWENTOXIN-I
- OMEGA-AGATOXIN IVA

Figure 10: **TKB Search Results:** The above results are displayed when the user searched for toxins that start with an 'L'

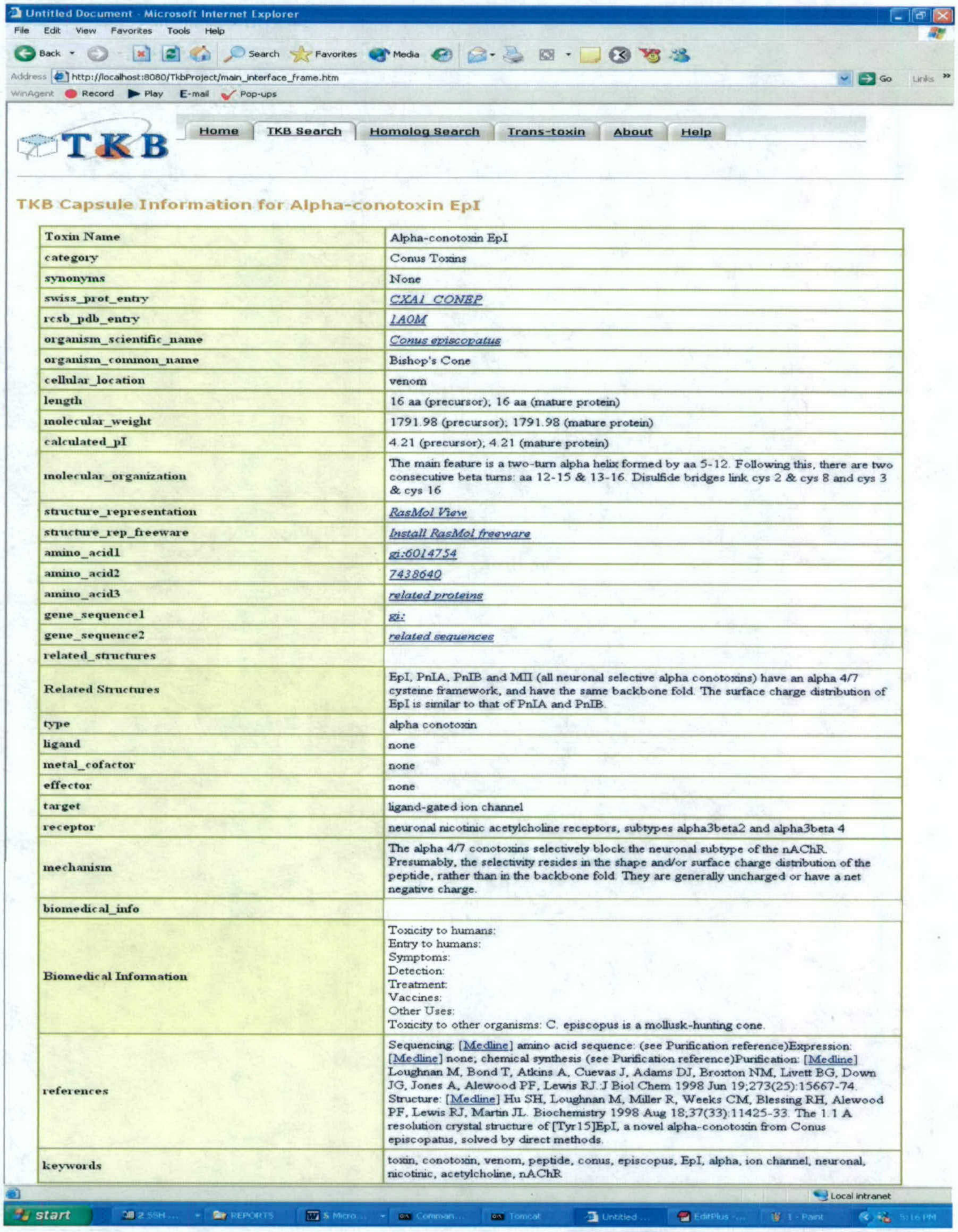

Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://localhost:8080/TkbProject/main_interface_frame.htm
Go Links
WinAgent Record Play E-mail Pop-ups
TKB
Home
TKB Search
Homolog Search
Trans-toxin
About
Help
Done
Local intranet
start

Selected Toxins

- Long neurotoxin 1 --- Naja kaouthia (Monocled cobra) (Naja siamensis)
- Long neurotoxin homolog NTL2I [Precursor] --- Bungarus multicinctus (Many-banded krait)
- LONG NEUROTOXIN 2 --- Ophiophagus hannah
- LONG NEUROTOXIN 2 --- Bungarus multicinctus
- LONG NEUROTOXIN 1 --- Naja oxiana
- LONG NEUROTOXIN 1 --- Laticauda semifasciata
- LONG NEUROTOXIN 1 --- Bungarus Multicinctus
- LEIUROTOXIN I --- Leiurus quinquestriatus hebraeus
- LEIUROTOXIN I-LIKE TOXIN P05 --- Androctonus mauretanicus mauretanicus
- Lebetase Le3 [Precursor] --- Vipera lebetina (Elephant snake) (Leventine viper)
- Leukocidin F --- Staphylococcus aureus
- Leukocidin S --- Staphylococcus aureus
- Lachesin [Precursor] --- Drosophila melanogaster (Fruit fly)
- Lachesin [Precursor] --- Schistocerca americana (American grasshopper)
- Locustatachykinin I --- Locusta migratoria (Migratory locust)
- Locustatachykinin II --- Locusta migratoria (Migratory locust)
- Locustatachykinin III --- Locusta migratoria (Migratory locust)
- LEUCOCIDIN F --- Staphylococcus aureus
- Locustatachykinin IV --- Locusta migratoria (Migratory locust)

Figure 11: TKB Search Results: The above results are displayed when the user searches for toxins in the category 'Spider'.

- The details of each toxin in the result can be seen by clicking on the toxin. A sample table containing toxin information is shown in Figure 12.



http://localhost:8080/TkbProject/main_interface_frame.htm

TKB Capsule Information for Alpha-conotoxin EpI

Toxin Name	Alpha-conotoxin EpI
category	Conus Toxins
synonyms	None
swiss_prot_entry	*CXA1_CONEP*
rcsb_pdb_entry	*1A0M*
organism_scientific_name	*Conus episcopatus*
organism_common_name	Bishop's Cone
cellular_location	venom
length	16 aa (precursor); 16 aa (mature protein)
molecular_weight	1791.98 (precursor); 1791.98 (mature protein)
calculated_pI	4.21 (precursor); 4.21 (mature protein)
molecular_organization	The main feature is a two-turn alpha helix formed by aa 5-12. Following this, there are two consecutive beta turns: aa 12-15 & 13-16. Disulfide bridges link cys 2 & cys 8 and cys 3 & cys 16
structure_representation	*RasMol View*
structure_rep_freeware	*Install RasMol freeware*
amino_acid1	*gi:6014754*
amino_acid2	*7438640*
amino_acid3	*related proteins*
gene_sequence1	*gi:*
gene_sequence2	*related sequences*
related_structures	
Related Structures	EpI, PnIA, PnIB and MII (all neuronal selective alpha conotoxins) have an alpha 4/7 cysteine framework, and have the same backbone fold. The surface charge distribution of EpI is similar to that of PnIA and PnIB.
type	alpha conotoxin
ligand	none
metal_cofactor	none
effector	none
target	ligand-gated ion channel
receptor	neuronal nicotinic acetylcholine receptors, subtypes alpha3beta2 and alpha3beta 4
mechanism	The alpha 4/7 conotoxins selectively block the neuronal subtype of the nAChR. Presumably, the selectivity resides in the shape and/or surface charge distribution of the peptide, rather than in the backbone fold. They are generally uncharged or have a net negative charge.
biomedical_info	
Biomedical Information	Toxicity to humans: Entry to humans: Symptoms: Detection: Treatment: Vaccines: Other Uses: Toxicity to other organisms: C. episcopus is a mollusk-hunting cone.
references	Sequencing: [Medline] amino acid sequence: (see Purification reference)Expression: [Medline] none; chemical synthesis (see Purification reference)Purification: [Medline] Loughnan M, Bond T, Atkins A, Cuevas J, Adams DJ, Broxton NM, Livett BG, Down JG, Jones A, Alewood PF, Lewis RJ. J Biol Chem 1998 Jun 19;273(25):15667-74. Structure: [Medline] Hu SH, Loughnan M, Miller R, Weeks CM, Blessing RH, Alewood PF, Lewis RJ, Martin JL. Biochemistry 1998 Aug 18;37(33):11425-33. The 1.1 A resolution crystal structure of [Tyr15]EpI, a novel alpha-conotoxin from Conus episcopatus, solved by direct methods.
keywords	toxin, conotoxin, venom, peptide, conus, episcopus, EpI, alpha, ion channel, neuronal, nicotinic, acetylcholine, nAChR

Figure 12: Toxin Information: Details of the toxin 'Alpha-conotoxin EpI --- Conus episcopatus'.

- The user can view a 3-D representation of the toxin structure by clicking on a link called ‘Structure’ in the detailed toxin information (uses a software in the public domain). A sample 3-D image of a toxin structure is shown below:

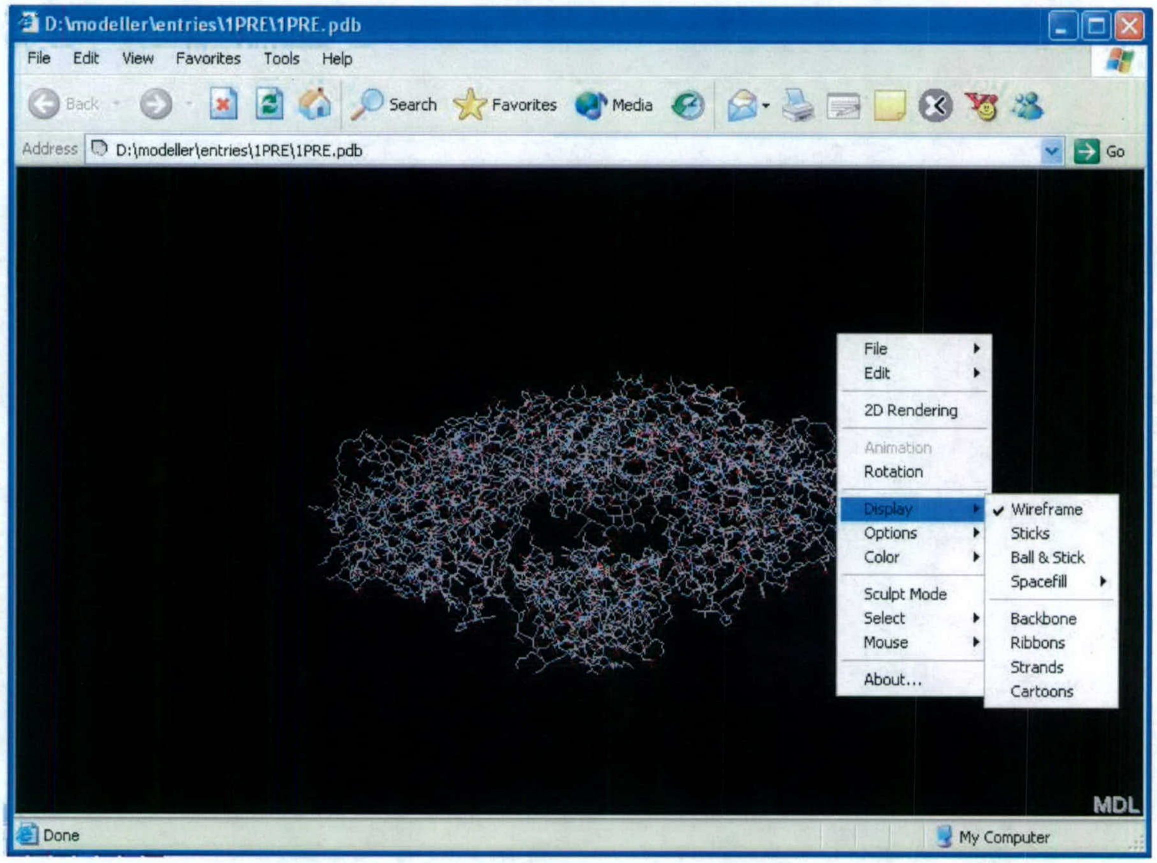


Figure 13: 3-D Visualization of the Protein Structure: User can choose different display modes such as Wireframe, Sticks etc. There are also options to rotate the toxin structure, change colors, and more.

3.4. Update interface

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 update*: This includes the following two tasks:
  - Update of TKB
  - Update of NR
- *User-initiated update*

These tasks are explained in detail in the following sections.

3.4.1. Update of TKB

The TKB has to be updated whenever new proteins are added to the NR database, since new entries in the NR database could bring 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 sequence against the most recently available updates to the NR database.

ii. Process the list of potential homologs to obtain the list of homologs for the input sequence. This involves filtering based on e-values and identity cut-offs.

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

3.4.2. Update of NR

A copy of the NR database is being maintained as part of the system. 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-redundancy of the NR database.

3.4.3. User initiated update of TKB

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

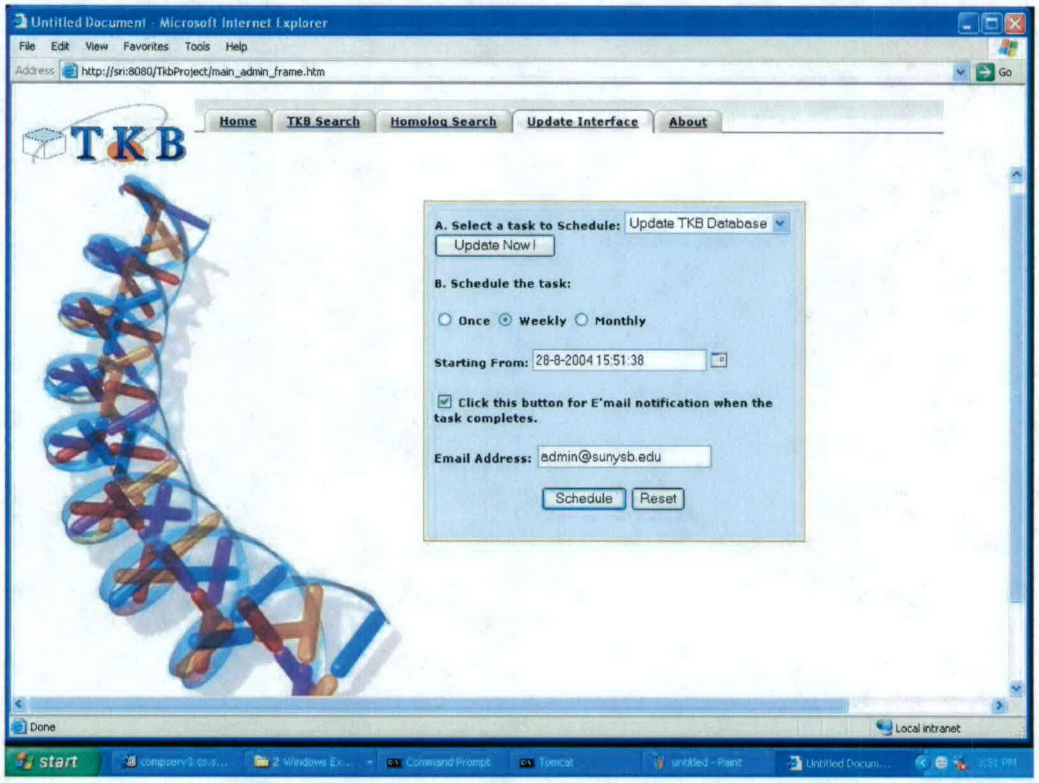
- *PDB ID*: Using WinAgent, Swiss-Prot is searched with the toxin's accession number (for example, P10844 for Botulinum neurotoxin type B), and the PDB IDs for the toxin are extracted.
- *Active site information*: Using WinAgent, PDB, PDBSUM, and LPC databases are searched using the PDB ID of a toxin, and the active site information, if any, is extracted.

- *Models*: If the structure information for the toxin is available, a model is built for the corresponding homologs. This is based on the alignment of the toxin and the homolog and uses MODELLER – a commonly used comparative modeling tool. MODELLER has a script language to control the modeling process. Our system generates appropriate scripts and in this way the use of MODELLER is completely automated.

3.4.4. Scheduler

- The above-mentioned update tasks are long-running transactions that are 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 with a given interval. The scheduler then performs the updates automatically.

The user interface for the scheduler is shown in Figure 14.



A. Select a task to Schedule: Update TKB Database

B. Schedule the task:

○ Once ● Weekly ○ Monthly

Starting From: 28-8-2004 15:51:38

☑ Click this button for E'mail notification when the task completes.

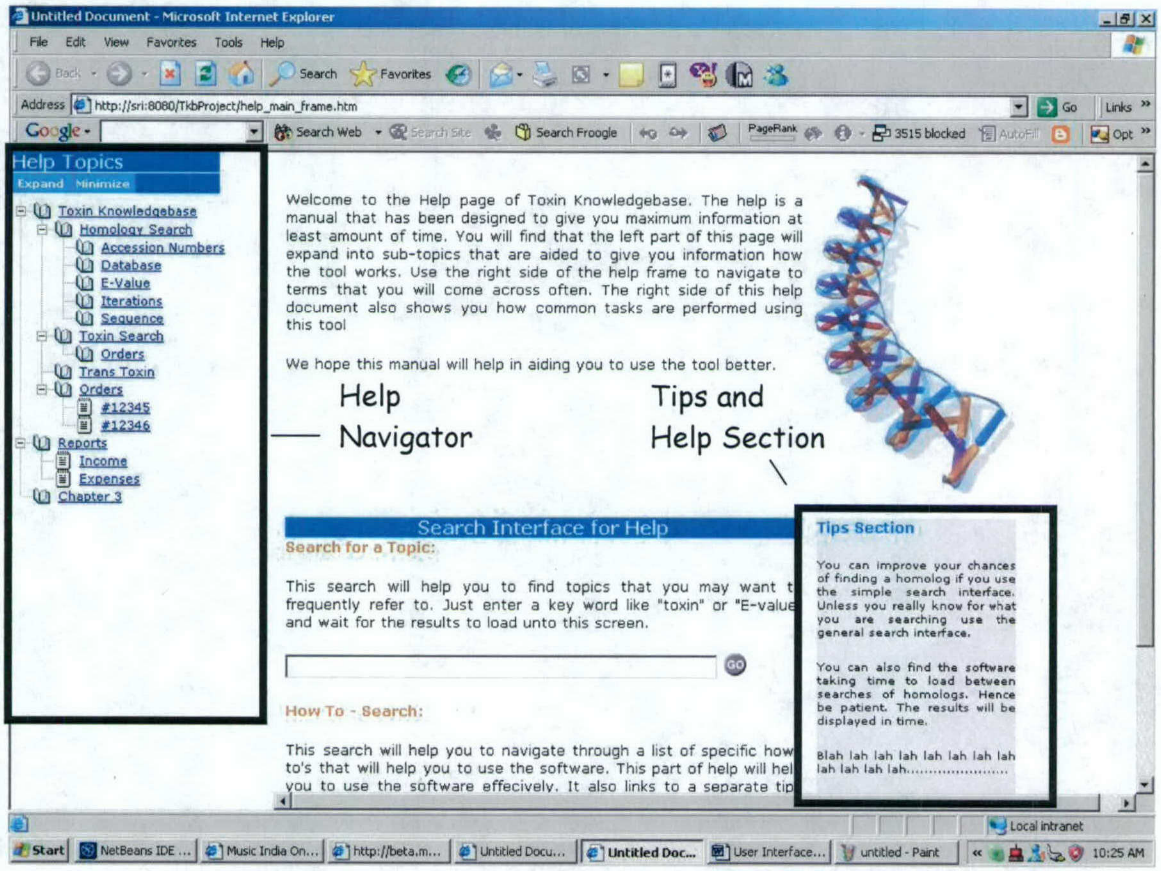
Email Address: admin@sunysb.edu

Figure 14: The Update Interface: The administrator can schedule Update Tasks by selecting the Task action from a menu. The administrator decides the frequency at which the scheduling has to take place and the start date. The administrator can choose to get an E-mail notification when the update completes.

The information about the status of the execution of the scheduled tasks is stored in the database. In the future the administrator will be provided with an interface to see the status of these scheduled tasks and will also be able to update the task information. The administrator will also be able to delete a pre-defined task.

3.5. Help interface

Help pages are provided for each module of the interface as shown in Figure 15. The help consists of hierarchical tree structure to help users navigate to the desired piece of information. A separate section provides information regarding the project and the site. Help pages can display information "inline." This is especially useful when forms are being used to obtain user input.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://sri:8080/TkbProject/help_main_frame.htm
Go Links
Google Search Web Search Site Search Froogle PageRank 3515 blocked AutoFill Opt
Help Topics
Expand Minimize
Toxin Knowledgebase
Homology Search
Accession Numbers
Database
E-Value
Iterations
Sequence
Toxin Search
Orders
Trans Toxin
Orders
#12345
#12346
Reports
Income
Expenses
Chapter 3
Help Navigator
Tips and Help Section
Search Interface for Help
Search for a Topic:
GO
How To - Search:
Tips Section
Local intranet
Start NetBeans IDE ... Music India On... http://beta.m... Untitled Docu... Untitled Doc... User Interface... untitled - Paint
10:25 AM

Welcome to the Help page of Toxin Knowledgebase. The help is a manual that has been designed to give you maximum information at least amount of time. You will find that the left part of this page will expand into sub-topics that are aided to give you information how the tool works. Use the right side of the help frame to navigate to terms that you will come across often. The right side of this help document also shows you how common tasks are performed using this tool

We hope this manual will help in aiding you to use the tool better.

This search will help you to find topics that you may want t frequently refer to. Just enter a key word like "toxin" or "E-value and wait for the results to load unto this screen.

This search will help you to navigate through a list of specific how to's that will help you to use the software. This part of help will hel you to use the software effecively. It also links to a separate tip

You can improve your chances of finding a homolog if you use the simple search interface. Unless you really know for what you are searching use the general search interface.

You can also find the software taking time to load between searches of homologs. Hence be patient. The results will be displayed in time.

Blah lah lah lah lah lah lah lah lah lah lah lah......................

Figure 15: Help Interface Design: The interface is designed to be user friendly and it allows the user to navigate easily through the help system. The help interface also provides tips and useful strategies for using the site as a whole.

4. System Requirements

- Oracle 10G
- Web Server/JSP Engine (Tomcat)
- Internet Explorer 5.5 onwards
- Stand-alone versions of PSI-BLAST, MODELLER, SPASM.
- MDL (Molecular Visualization Resources) CHIME plug-in.

B. Method development:

This section deals with methods that are being identified, analyzed, modified and developed to identify homologs and structural motifs and pertains to Specific tasks 3 and 4.

1.1. Experiments with Profile Hidden Markov Model (PHMM): We conducted three experiments with PHMMs to find homologous proteins. For each of the experiments, we begin using an initial set of pre-classified sequences known to be from that family, and build a PHMM from the multiple sequence alignment on the sequences. We then use PHMMs, to discover other protein sequences homologous to them.

Experiment 1 - Classification: Our first experiment was from an extensive dataset from iProClass, an integrated protein classification database. We built PHMM from a multiple sequence alignment of 61 sequences from the ras family. Then we scored 1912 other pre-classified sequences from this family and 3247 proteins not belonging to this family against the PHMM built. Using the scores of positive and negative examples, we determine an appropriate threshold score to optimize either precision, recall or both.

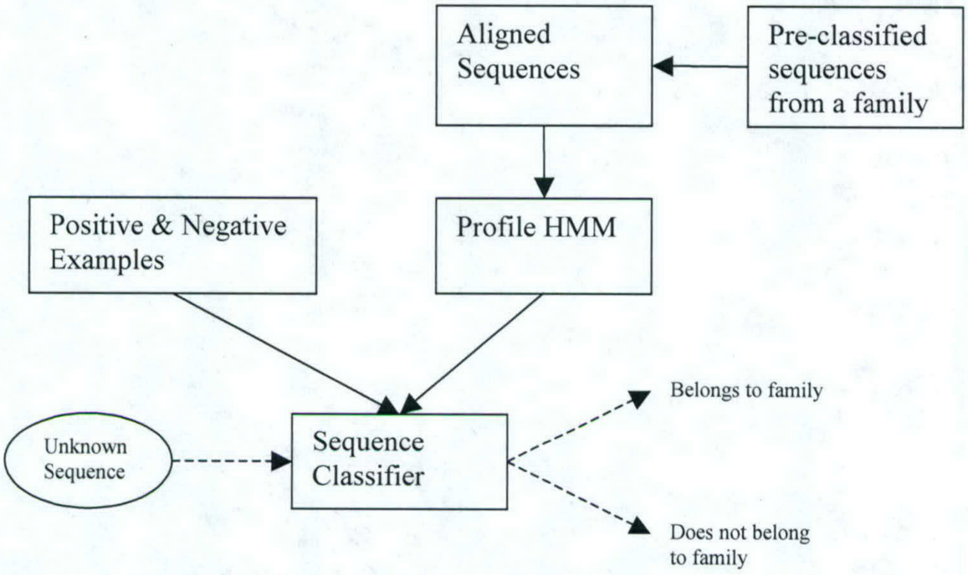

Aligned Sequences
Pre-classified sequences from a family
Positive & Negative Examples
Profile HMM
Unknown Sequence
Sequence Classifier
Belongs to family
Does not belong to family

Figure 16: Steps for Experiment 1 to classify proteins

Objective	Precision	Recall
Maximize Precision	100%	81.98%
Maximize Recall	44.23%	100%
Maximize Both	95.57%	95.18%

Table 1: Results

Experiment 2 – Discovering family members: Our second and third experiments were based on actual toxins from the existing TKB. We obtained 26 examples of Botulinum Neurotoxin family and constructed the PHMM. We also used the sequences to BLAST over the Protein database. We scored the 61 unique sequences obtained from the results and determined a safe threshold score since we did not have sufficient pre-classified examples. 47 other potential proteins were estimated from the family.

Experiment 3 – Protein to Protein Transformation Prediction: In this experiment, we had the case where a protein Staphylococcal Enterotoxin A (SEA) behaved like Staphylococcal Enterotoxin E (SEE). SEA and SEE have different V-beta specificity.

However, when residues 200-207 in SEA are substituted by corresponding residues in SEE, it begins to behave like SEE with regard to V-beta specificity. The task was to find other proteins that could potentially transform to SEE.

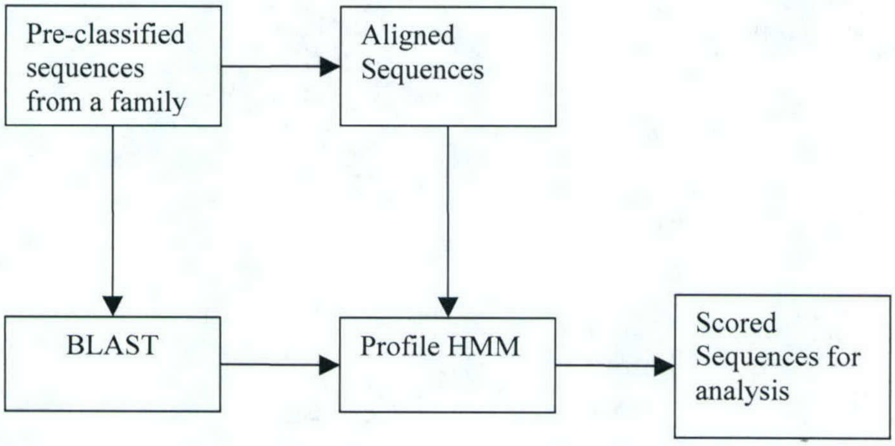

Pre-classified sequences from a family
Aligned Sequences
BLAST
Profile HMM
Scored Sequences for analysis

Figure 17: Steps for Experiment 2 and 3 to find related proteins

For this, a similar approach was used. We found 9 proteins from the Staphylococcal Enterotoxin family, and built a PHMM from it. The 141 unique BLAST results were sorted according to their scores against the PHMM. They are shown in the following table. SEA and SEE are in bold. Results are still being analyzed.

Protein Details	**PHMM Score**
optimized enterotoxin B [synthetic construct]	255.7884
enterotoxin sec variant [Staphylococcus aureus]	253.2815
Staphylococcal Enterotoxin B Complexed With Gm3 Trisaccharide	244.6821
A Chain A, Structural Basis For The Altered T-Cell Receptor Binding Specificity In A Superantigenic Staphylococcus Aureus Enterotoxin-B Mutant	243.0727
enterotoxin sec variant [Staphylococcus aureus]	232.1632
Staphylococcal Enterotoxin E	**224.2829**
enterotoxin P [Staphylococcus aureus subsp. aureus Mu50]	218.1125
staphylococcal enterotoxin A precursor [Staphylococcus aureus subsp. aureus MW2]	218.1125
enterotoxin seb variant [Staphylococcus aureus]	215.3183
A Chain A, Staphylococcal Enterotoxin A	208.3236
A Chain A, Crystal Structure Of Staphylococcal Enterotoxin A Mutant H187a With Reduced Zn2+ Affinity	207.225
Staphylococcal enterotoxin A	**207.0708**
D Chain D, Crystal Structure Of The D227a Variant Of Staphylococcal Enterotoxin A In Complex With Human Mhc Class Ii	206.9373

Table 2: Portion of the sorted list for Experiment 3 showing SEA and SEE

2. Motif Discovery from 3D Structure of Proteins

Structure motifs are as important as motifs found in the sequence because they reveal structure specific properties of the proteins that cannot be determined from sequence motifs because various proteins fold differently in their 3D structure.

This section provides a technique to discover motifs in protein structures by examining the neighborhood of the amino acids in the proteins. The resulting motifs reveal patterns in the neighborhood of the amino acids, thus revealing the structural patterns in the given protein family. It also attempts to classify whether a new protein belongs to a particular family based on the motifs discovered in that family. A viewer for the input protein 3D structures and the discovered motifs is also developed for convenience. The motifs discovered for the input protein structures considered are shown and the protein classification accuracies are included.

2.1. Approach: Given a set of protein families, a set of neighbor strings are calculated for each of the proteins in the family as illustrated in Figure 4. The neighbor string of an amino acid A represents the 3D neighborhood of A, i.e., the list of amino acids that are near to A within a given threshold in the same order as the backbone.

After the neighbor strings are computed, for each of the neighbor strings frequent prefixes and suffixes are found for it and packing patterns are computed using regular string matching techniques. These patterns contain sequences of amino acids in the 3D space and frequent patterns in a given family of proteins are categorized as 3D motifs. These set of 3D motifs found in a given family of proteins were then used to classify a new protein, i.e., to determine whether that new protein belongs to the given family of proteins.

2.2. Preliminary Results & Conclusions: The test bed for experiments is obtained from Protein Data Bank Web site. Three different families of proteins were used in the experiments: Scorpion neurotoxin family, Zinc Finger family and Immunoglobulin family.

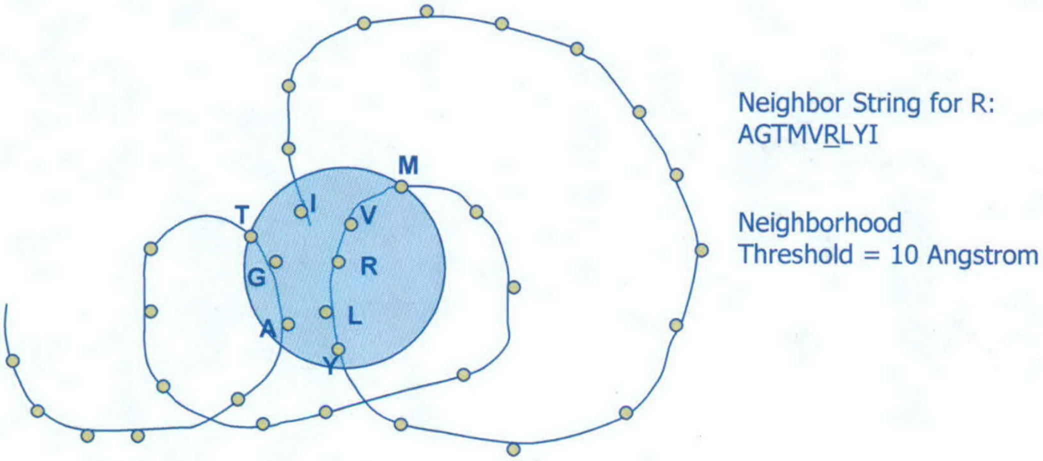


Figure 18: Illustration of how a Neighbor String is calculated

N-Fold methodology is used to test the efficacy of protein classification. The accuracies reported for each of three families are Scorpion neurotoxin family (100.00%), Zinc Finger family (78.57%), and Immunoglobin family (66.67%). The variance in accuracy is characterized by the number of proteins in each of the families and the complexity of their 3D structure. Also the 3D motifs obtained from this approach are short (the average length is 6.4 amino acids) and this affected the accuracy of protein classification. One of the future directions for the work is to enhance the motifs by using DBScan incremental clustering algorithm until it satisfies the support. Another future direction is along the lines of visualization that helps to better characterize the 3D motifs.

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.

Reportable outcomes

Four papers presented in a conference.

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. In Proceedings of Intelligent User Interfaces 2004.

2. A Gandhre, P Santhanagopalan, P Singh, D Ramavat, I Ramakrishnan, H. Davulcu. Creating and Managing Personal Information Assistants via a Web Browser: The WinAgent Experience. In Proceedings of VLDB Workshop on Information Integration on the Web (IIWeb'04), Toronto, Canada, August, 2004.

3. Hasan Davulcu, Zoe Lacroix, K Parekh, I Ramakrishnan, N Julasana. Exploiting Agent and Database Technologies for Biological Data Collection. In Proceedings of the International Workshop on Biological Data Management (BIDM'04), Zaragoza, Spain, September, 2004.

4. Saikat Mukherjee and I.V. Ramakrishnan. Taming the Unstructured: Creating Structured Content from Partially Labeled Schematic Text Sequences. Proceedings of International Conference on Ontologies, Databases and Application of Semantics (ODBASE), 2004 (Accepted for publication).

Conclusions

In our work so far, we have achieved establishing a powerful and useful data base which can used as a resource for identifying homologs of toxins which may become potential toxins.

Year 2 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. This approach will also follwed in the second year.
2. There are toxins which don't have active site information in PDB, PDBSUM, or LPC. In such cases, we need to find alternative information sources of active sites for such toxins. 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 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 may be eventually entered into the TKB.

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

One Ph.D. (full time) student and 10 M.S. students (all part time 20 to 50%)

Sub-contract to Arizona State University, Tempe, Arizona

1. H. Davulcu	Asst. Professor	10% effort

Two graduate students (part time)