Towards a Knowledgebase for Caspases Substrates

Hani Qudsi[1], Abed Al-Rahman Haddad[1], Islam Itkaidek[1], Hasan Al-Taradeh[2], Hashem Tamimi[12], Mahmoud Alsaheb[1], Yaquob Ashhab[2]
[1]Information Technology Department, Palestine Polytechnic University, Hebron, Palestine.
[2]Biotechnology Research Center, Palestine Polytechnic University, Hebron, Palestine

Abstract—

The idea of the project is to develop a user friendly knowledgebase that contains all proteins that are cleaved by caspase enzymes.

The proteins are graphically represented to facilitate the analysis and studying of caspases substrates. Each protein file contains rich information such as the cleaving caspase(s), the exact cleavage position, functional domains, biological implications of the cleavage process, and links for the pathway and interaction for each protein.

The database offers easy to use searching tools that allow simple as well as advanced search. The database will be freely available via a web site through Palestine Polytechnic University's server.

The unique feature of this database is its ability to evolve through the contribution of browsers (biologists) from all around the world. The browsers can contribute to grow the database through adding more caspase substrates or editing already available data. The contributions of browsers were designed in a way to ensure high security data entry as well as to minimize effort needed by the administrators to follow-up and validate the contributions.

In addition, our database is capable to automatically and regularly update all its content based on recent data that is available through various central databases such as Uniprot and PubMed.

.

Keywords; Caspases; Knowledgebase; Substrate.

I. INTRODUCTION

Information technology has affected every aspect of our life. Besides being a stand-alone area, IT has become a part of each field of science in terms of reducing time and efforts in accomplishing certain tasks. Biological science is one scientific field that benefited from IT in an excellent way, since IT offered effective remarkable solutions for utilizing a lot of processes done in biomedical science. This paper is presenting a solution for some problems that were detected in biotechnology, specifically the knowledge basis for proteins.

Caspases are cysteine proteases that play an essential role in programmed cell death and inflammation [1]. Caspases mediate their functions through the cleavage of wide range of protein-substrates. The cleavage of all caspases substrates is characterized by the presence of Asparatic acid at the P1 site [2]. The expanding list of caspases' substrates motivated the idea of a database that stores the substrates and the related information to it. However, the few available caspases databases are suffering from a lack of periodic update of the proteins information, poor and static presentation of the data [3, 4]. Caspasome was developed to be a dynamic database that allows the users to participate in the development and update of the data. A search engine with a graphical integration were implemented to simplify the use of caspasome and make the data easier to understand by biologists.

II. DATABASE DESCRIPTION

A. Database Content

The presented knowledgebase is concerned with the proteins that are cleaved by caspase. This type of information is not completely stable since biologists continue to discover new cleavage positions for proteins with known and unknown positions. Therefore, the data should be updated periodically. This information is collected in two ways. First, the information that is concerned with the cleavage positions is collected from the biologists themselves since these kind of information is relatively new. This information took the form of Excel files or by the biologists' contribution. Second, other information concerned with proteins such as proteins' name, domains ..etc is collected from the uniprot website in form of XML files.

B. Database Construction

The knowledgebase is implemented using SQL Server 2008 which provides a complete and easy platform for databases including database creation, controlling, and management. It supports three type of authentication: SQL Server authentication, Windows Authentication, and Mixed mode, which uses the both types. This knowledgebase is publically available at *http://195.3.189.246/BrowsersPages/Default.aspx*.

III. UTILITY

This knowledgebase contains information for all known substrates of all caspases. The data of the substrates is easily accessible through an online website that represents a unified reference for all the information related to that kind of proteins with an efficient search methods. It also provides a graphical and user-friendly interface for representing the proteins information. The interface was designed with ASP.Net. It represents a safe environment for biologists around the world to add contributions about proteins' information and an easy way to provide a feedback to the system administrator. In

addition, the website provides a tool called CAT3 (Caspase 3 Tool), which is a powerful bioinformatics tool that can predict the cleavage site of caspase-3 substrates with accuracy higher than any other similar software [5].

The system functions, interactions and the data flow between the system components and the actors are illustrated in figure 1.

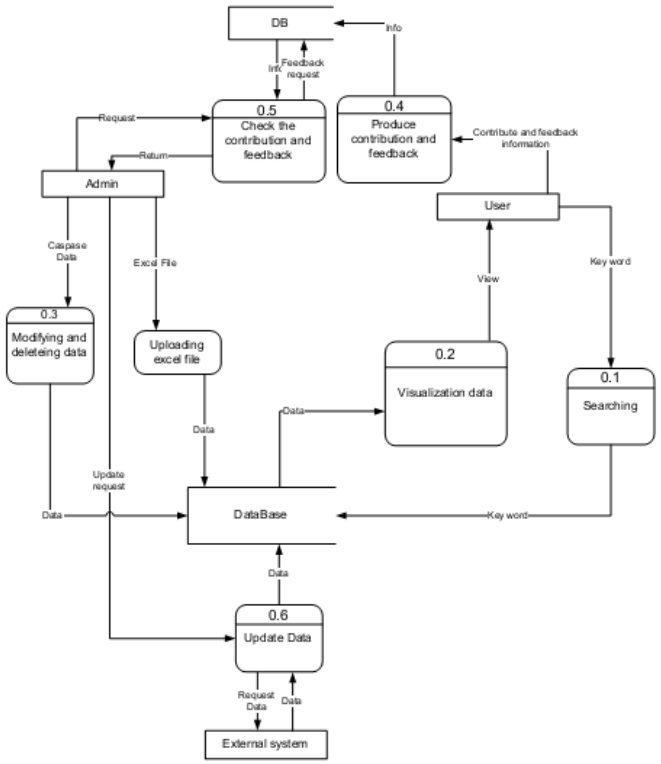

Figure 1 Data Flow Diagram. The interactions and data flow between the system components.

The flow of data, as represented in figure 1, will starts when the user search for a protein (process #1) then the data fetch from database and return selected protein information and the graphical view for protein sequence (process #2). The user can modify or add new protein information through the contribution interface. These contributions will by checked by the admin. Confirmed proteins will automatically get basic information retrieved from parsed XML files of UNIPROT and PUBMED databases sources.

The knowledgebase was configured and uploaded on a web server. The website is browsed using any browser-controlled environment and can be accessed over the Internet. It was deployed using the three-tier architecture. The following figure illustrates the system architecture.

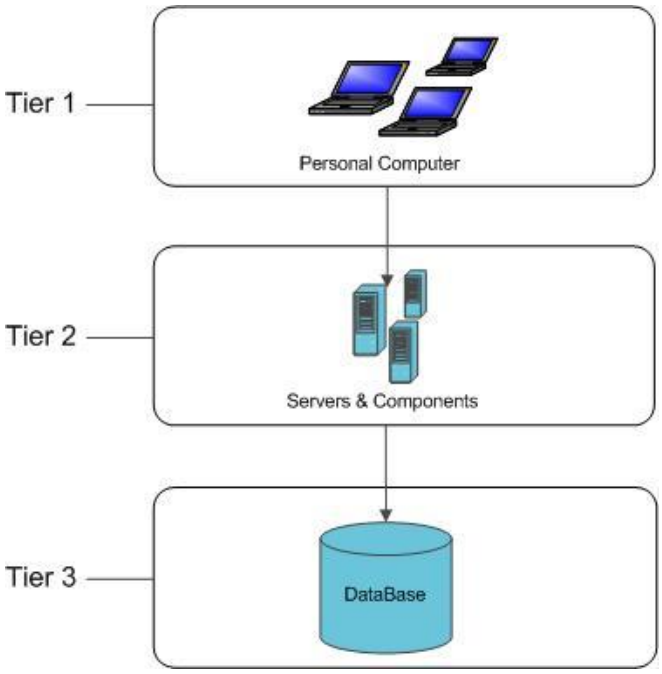

Figure 2 The three-tier system architecture. Three-Tier Architecture has three layers (1) User interface layer, it is representing web pages (2) Business logic, it is use classes and builds functions in it, these classes receive data as parameter and pass the parameters to the procedure in the database, so it is an intermediary between user interface and database procedure (3) DBMS it includes stored procedures that execute SQL statements in databases, which use to insert, update and delete data.

The website consists of four main tools: a search engine (basic and advanced), data view, biologist's contribution and the CAT3 tool.

A. Search

The website supports two search mechanisms: basic and advanced.

1) Basic Search

By entering a keyword in the provided space, the system will search the database for any related information. The results will be shown in a way that will easily provide the basic information for the retrieved protein. The protein information includes: the accession number, entry name and the recommended name. The figure below shows the view of the result.

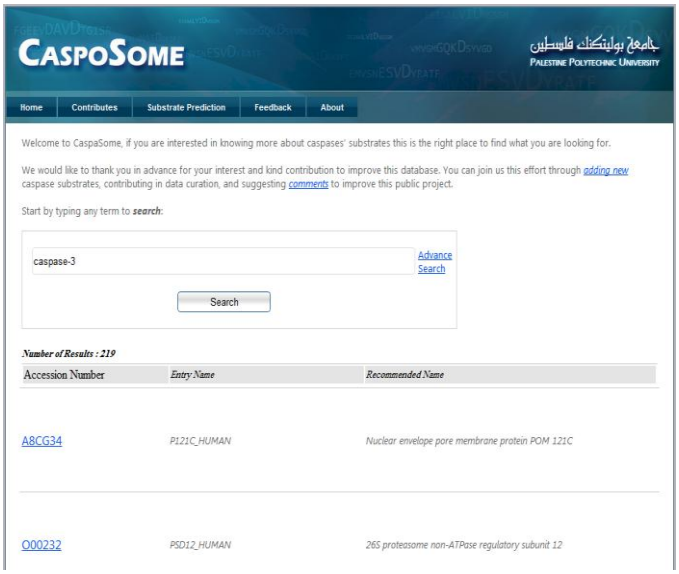

Figure 3 Basic Search Snapshot. A snapshot of a basic search and the way results are represented.

2) Advanced Search

This technique enables the browser to search in more details and specific fields. Browsers can select field to search about it, and then enter the keyword that is related with the selected field. If the browser selected more than one field, the user can connect between these fields by one of the logical relationship (And, Or, and Not).

The system offers auto complete for each field, when the browser enters a keyword, the system will suggest a keyword completion to facilitate the search process.

The results will be shown to deliver the basic information for the retrieved protein. It will show the accession number, entry name and the recommended name. the figure below shows the view of the result.

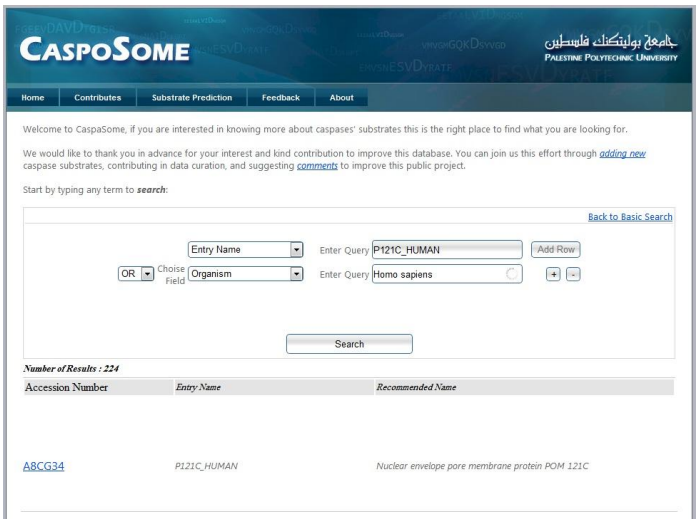

Figure 4 Advance Search Snapshot. This figure shows the page of the advance search and result.

3) Contribution

Browsers can contribute to add new protein information or add new cleavage site position of protein to the knowledgebase.

A contribution has two types of information, first type is personal information to identify the contributor and this information are Name, Scientific Degree, email, and Country, and the second information type is related to the new protein information and this information are Accession-Number, Cleavage-Position, Evidence, Cleavage assay, Tetra peptide Motif, and Cleavage outcome.

The figure below shows the view of the contribution page.

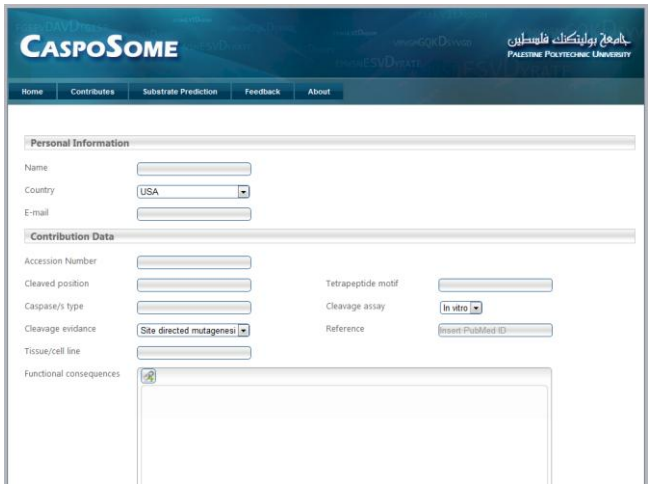

Figure 5 Advance Search Snapshot. This figure shows the page of add new contribution.

4) Data View

After searching and selecting specific protein, the browser will see the page of protein information.

The protein page divided to seven main blocks. Identity data block which have the main information for protein as Accession number, entry name …etc. Cleavage data block which have information about the cleavage site or sites for this protein. Sequence data block have the graphically present of the protein sequence and domains, and this block present the sequence in Fasta format. General annotation block, gene ontology block and external links block will be shown too. Finally the references block will shows the protein cleavage information references using for proof the cleavage position and using in cleavage outcome papers

The follow figure shows the Sequence block which has the graphically representation of the protein domains and cleavage positions.

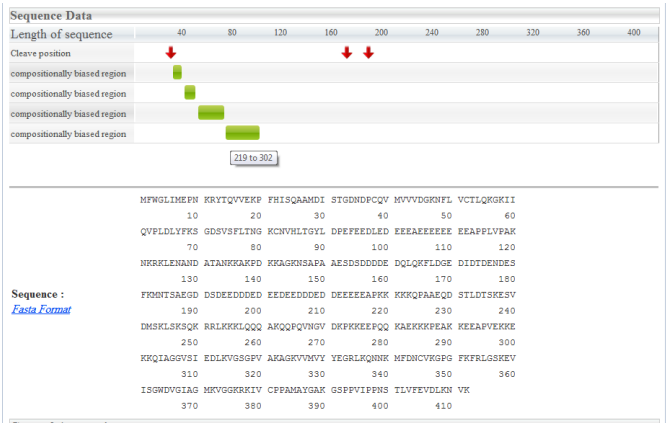

Figure 6 Graphically Representation of Protein. This figures show the graphically representation of the protein cleavage, domain and sequence information and fasta format for protein sequence.

5) Substrate Prediction Tool

substrate prediction tool published at [3] 2012 and was added to the website as a rich and important tool which provide a prediction about cleavage position in a protein sequence and the cleavage score for each position, this prediction reduce the time and the cost for researchers for being test the cleavage in laboratories. This tool work when the browsers adding the cleavage name, protein sequence and by which caspase type you want to test then the results are cleavage name, tetrapeptide motif and the score for each cleavage position, the fisrt version of this tool will support caspase-3 only.

Figure 7 shows the prediction tool page.

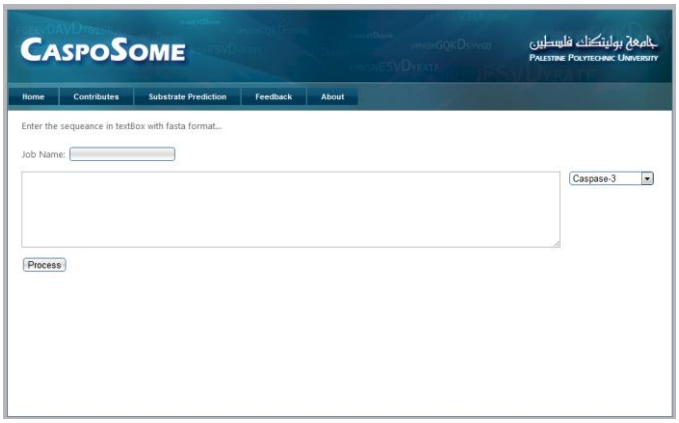

Figure 7 Prediction Tool Page Snapshot. This figure is shows the page of the prediction tool, where has the text box for the project name and the text box for the sequence.

CONCLUSION

This web application has been build to achieve the objective of the project, Caspases Knowledgebase is a unique biological resource for researchers from various biomedical disciplines, which represents a reliable, sustainable and self evolving bioinformatics system.

REFERENCES

[1] Pop C, Salvesen GS. Human caspases: activation, specificity, and regulation. J Biol Chem. 2009. 14;284(33):21777-81.

[2] Timmer JC, Salvesen GS. Caspase substrates. Cell Death Differ. 2007. 14(1):66-72.

[3] A U Lüthi1 and S J Martin1. The CASBAH: a searchable database of caspase substrates Cell Death Differ (2007)14, 641-650.

[4] Lawrence JK Wee, Tin W Tan and Shoba Ranganathan.SVM-based prediction of caspase substrate cleavage sites. BMC Bioinformatics 2006, 7(Suppl 5):S1.

[5] Ayyash M, Tamimi H, Ashhab Y. Developing a powerful in silico tool for the discovery of novel caspase-3 substrates: a preliminary screening of the human proteome. BMC Bioinformatics. 2012. Jan 23;13:14.