BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database
Volume 10(12)

Intelligent Access to Sequence and Structure Databases (IASSD) – an interface for accessing information from major web databases

Sayak Ganguli*, Manoj Kumar Gupta, Protip Basu, Rahul Banik, Pankaj Kumar Singh, Vineet Vishal, Abhisek Ranjan Bera, Hirak Jyoti Chakraborty & Sasti Gopal Das

DBT Centre for Bioinformatics, Presidency University, Kolkata – 73; Sayak Ganguli - Email: sayak@bioinfo-presiuniv.edu.in; *Corresponding author

Received October 16, 2014; Accepted December 10, 2014; Published December 31, 2014

Abstract:
With the advent of age of big data and advances in high throughput technology accessing data has become one of the most important step in the entire knowledge discovery process. Most users are not able to decipher the query result that is obtained when non specific keywords or a combination of keywords are used. Intelligent access to sequence and structure databases (IASSD) is a desktop application for windows operating system. It is written in Java and utilizes the web service description language (wsdl) files and Jar files of E-utilities of various databases such as National Centre for Biotechnology Information (NCBI) and Protein Data Bank (PDB). Apart from that IASSD allows the user to view protein structure using a JMOL application which supports conditional editing.

Availability: The Jar file is freely available through e-mail from the corresponding author.

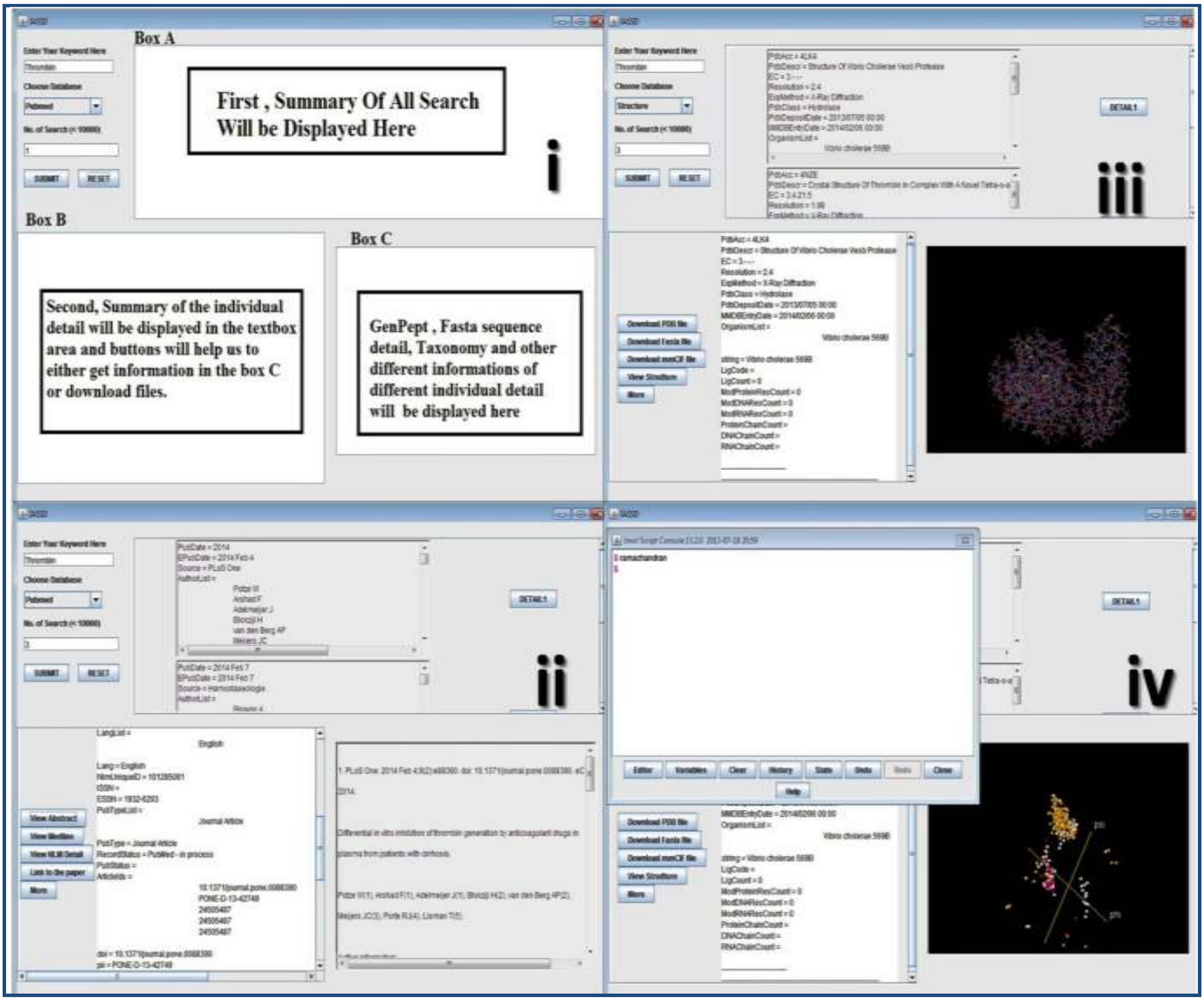
Keywords: IASSD, NCBI, PDB, WSDL, conditional editing

Background:

Biology today is a highly information rich science, thus necessitating the use of bioinformatics and computational techniques. The foremost challenge is the management of scientific data are which is important for the support of life science discovery. As computational models of essential cellular components are gradually produced, biology research would require computer intervention more often. Cross domain access to biological data and information followed by management, organization and integration is needed for the successful conversion to in silico biology. Biologists and other scientific research workers need to be provided with integrated platforms which would shorten the time for query processing as well as provide apt and intelligent answers to the queries. Quite a few interfaces have been designed for accessing and retrieving biological data in the past **[1]**. Few Of them being: The Sequence Retrieval System (SRS); TAMBIS: transparent access to multiple bioinformatics information sources; and The Kleisli Query System **[2, 3]**; however, the lack of continuous back end support and the ever changing nature of the biological ontologies have rendered them inconsequential.

Implementation:

IASSD integrates heterogeneous data sources from NCBI (PubMed, Nucleotide, Protein, NLM, MedLine and Structure Databases) and PDB. This interface is general purpose, written in Java, and provides a navigable interface for a variety of special-format bioinformatics repositories. The system has a universal internal data model that allows for the direct representation of relational, electronic data interchange (EDI) formats, including XML-based sources **[4].**



Box A

First , Summary Of All Search Will be Displayed Here

Box B

Second, Summary of the individual detail will be displayed in the textbox area and buttons will help us to either get information in the box C or download files.

Box C

GenPept , Fasta sequence detail, Taxonomy and other different informations of different individual detail will be displayed here

Figure 1: The features of IASSD; **(i)** the interface; **(ii)** Typical result for a keyword based query, **(iii)**PDB query and display in Jmol; **(iv)** Editing the structure using Jmol commands.

Key Features:

IASSD provides on demand access and retrieval of the most up-to-date biological data and is benchmarked with the ability to perform complex queries across multiple heterogeneous databases to find the most relevant information. Apart from this, it is also equipped with a robust information integration infrastructure that connects various computational steps involving database queries, computational algorithms, and application software with a strong base in the use of wsdl services across the web **[5]**. This is one of the unique interfaces which allow the user to navigate from a paper directly to the structure of the molecule described in the publication utilizing JMOL **(Figure 1).**

Handling of functional routines using Java:

IASSD uses log4j , NCBI eutils for sending and retrieving information from NCBI. Its uses Jmol Application for visualization of protein structure. Log4j can be configured by external configuration file at runtime. Log4j keep watch in terms of levels of priorities and provide mechanisms by which it direct logging information to a great variety of destinations like file, console, database etc. The eUtils(Entrez Programming Utilities),a group of eight server-side programs **[6]** that ensures availability, of a stable interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI), is always fixed and it translates a standard set of input parameters into the values necessary for searching and retrieving data from different NCBI software components. Thus the eUtils serve as organized interface to the Entrez system. For accessing data, software first sends an eUtils URL to NCBI, then get the results of this posting, after which data are being processes as required. The software can thus make use any computer language like Java, Perl, Python, for sending a URL to the eUtils server and interpret the XML response. Apart from the Jmol **[7]** java applet, a version of Jmol that works only when embedded in web browsers, the Jmol application is a stand-alone program (as are PyMol and Rasmol). The Jmol application runs independently and has nothing to do with web browsers. The Jmol program is implemented with help of Java and performs identically on Linux, Macs, Windows and Apple. The workflow is explained in **Figure 2.**

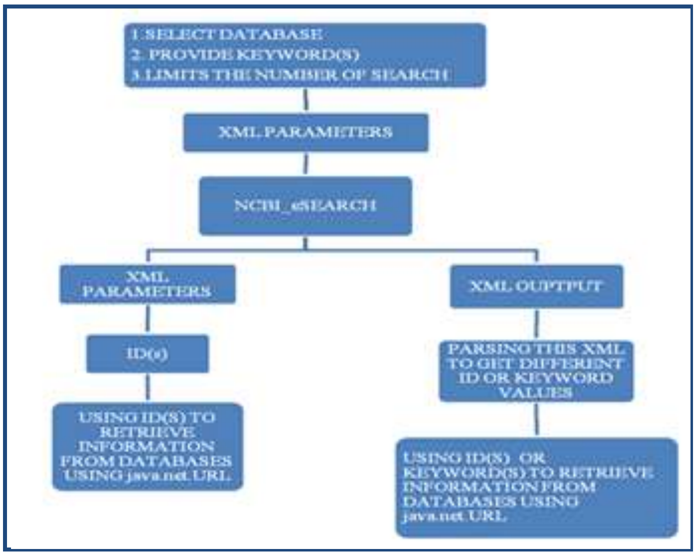

1.SELECT DATABASE
2. PROVIDE KEYWORD(S)
3.LIMITS THE NUMBER OF SEARCH
XML PARAMETERS
NCBI_eSEARCH
XML PARAMETERS
XML OUPTPUT
ID(s)
PARSING THIS XML TO GET DIFFERENT ID OR KEYWORD VALUES
USING ID(S) TO RETRIEVE INFORMATION FROM DATABASES USING java.net.URL
USING ID(S) OR KEYWORD(S) TO RETRIEVE INFORMATION FROM DATABASES USING java.net.URL

Figure 2: The scripting pipeline

Caveat:

IASSD requires a high-speed internet connection in the terminal where it is to be used since it does not come with a predefined database. Apart from that, the current version only supports key word based search strings and no database specific identifier can be used for the search.

Future Development:

Accessing and retrieving data from various databases; integrating more best-of-breed analytical tools and algorithms for extraction of useful information from the massive volume and diversity of biological data; Making the application available in LINUX and OS-X.

References:

[1] Etzold T *et al. Methods in Enzymology*. 1996 **266:** 114 [PMID: 8743681]
[2] Wong LJ, *Jour Func Prog.* 2000 **10**: 19
[3] Davidson S *et al. Int Journal of Digital Libraries.* 1997 **1**: 36
[4] www.w3.org/TR/xmlschema-{1,2}, May2001
[5] http://www.w3.org/TR/wsdl12, June 2003
[6] http://eutils.ncbi.nlm.nih.gov/entrez/eutils/
[7] Pavelin K *et al. PLoS Comput Biol*. 2012 **8:** e1002554 [PMID: 22807660]

Edited by P Kangueane
Citation: **Ganguli *et al.*** Bioinformation 10(12): 764-766 (2014)