Bioinformatics Web Services

Mohamad Ibrahim Ladan
Computer Science Department, Haigazian University, *Beirut – LEBANON*

Abstract - *Bioinformatics is emerging as a new major or emphasis of study or work as a merge between biology and information technology majors or field of work. In addition, Web Services have emerged as a new Web-based technology paradigm for exchanging information on the Internet using platform-neutral standards, such as XML and adopting Internet-based protocols. This has helped in the birth of what is called Bioinformatics Web Services. In this paper, I will introduce bioinformatics web services, and survey the different existing tools and mechanisms available to develop such systems.*

Keywords: Bioinformatics, Web Services, Bioinformatics Web Services.

1 Introduction

These The recent advances in the field of molecular biology and genomic sequences technologies have resulted in flood of data and biological information from the research community. In order to utilize this huge volume of data in an efficient way, there was a need for the use of information technology and computerized tools to store, manage, view, index, and analyze this volume of data. This has led to the birth of what is called bioinformatics. It is a new science field in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of this field is to create a global perspective from which unifying principles in biology can be determined [1]. To accomplish this goal, there is a clear need for a technology environment or system that links together and make use of data and tools in different formats and shape found at different computers in different locations to create workflows that can be used by biologists from anywhere at anytime. Web Services technology is the right platform to be used to fulfill this requirement.

Web services represent a new programming approach based on a document-oriented model designed for interoperability at a document, typically XML, level. They are modular, self-describing, self-contained applications that are based on open standards and can be published, located, and invoked across the Internet/Web. Web services are a distributed computing technology that provides software services over the web and enable us to build Web-based applications using any platform, object model, and programming language that we may require [2]. Because of its features, Web Services is the perfect choice for bioinformatics applications developments.

The rest of this paper is organized as follows: Section 2 introduces and discusses the bioinformatics field. Section 3 introduces the Web Services technology and environment. Section 4 introduces and discusses the Bioinformatics Web Services in general, and surveys and discusses the different types of existing Bioinformatics Web Services in particular with their benefits and shortcomings. Finally, section 5 concludes the paper.Instructions for authors

2 Bionformatics

Please Bioinformatics is the analysis of biological information using computers and information technology. According to Oxford dictionary, bioinformatics is conceptualizing biology in terms of molecules and applying information technologies to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications. The National Center for Biotechnology Information defines bioinformatics as [3]: "Bioinformatics is the field of science in which biology, computer science, and information technology merges into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."

In the past, the main concerns of bioinformatics were storing, managing, analyzing volume of biological information, and development of complex interfaces to access this information and submit new and updated information by different researchers. In February 2001, the scientists have mapped the human genome, the complete set of genes. The process is called sequencing. It is an overwhelming process requiring complex analytical tools and techniques, and it was considered as the greatest success of bioinformatics tools [4]. With time, the bioinformatics field has evolved and is currently using different computational techniques which includes besides sequencing and structural alignment, database design and data mining, macromolecular geometry,

prediction of protein structure and function, gene finding, and expression data clustering.

In brief, the main objectives of bioinformatics can be stated as follows: The creation and maintenance of a database to store biological information, the development of complex interfaces for researchers to access and update existing data, and to develop tools and computational techniques for analyzing and interpreting the various types of data..

3 Web Services

Web Services are based on a collection of standards and protocols that allow us to make processing requests to remote systems by speaking a common, non-proprietary language and using common transport protocols such as HTTP and SMTP. Web services represent a new programming approach based on a document-oriented model designed for interoperability at a document, typically XML, level. They are modular, self-describing, self-contained applications that are based on open standards and can be published, located, and invoked across the Internet/Web. Web services enable us to build Web-based applications using any platform, object model, and programming language that we may require. In addition, they are implemented using a collection of several related, established and emerging technologies and communication protocols that include HTTP, XML, Simple Object Application Protocol (SOAP), Universal Description Discovery and Integration (UDDI), Web Services Description Language (WSDL), Common Object Request Broker Architecture (CORBA), Java Remote Method Invocation (RMI) , and .NET [2].

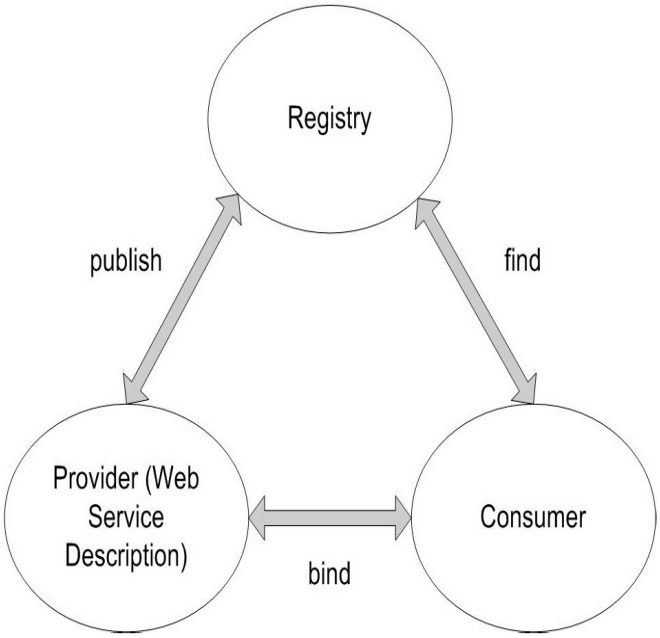

Figure 1. The web service model

The web service model consists of three entities, the service provider, the service registry and the service consumer. Figure 1 shows a graphical representation of the traditional web service model. The service provider creates or simply offers the web service. The service provider needs to describe the web service in a standard format, which in turn is XML and publish it in a central Service Registry. The service registry contains additional information about the service provider, such as address and contact of the providing company, and technical details about the service. The Service Consumer retrieves the information from the registry and uses the service description obtained to bind to and invoke the web service.

Web Services have several benefits and can offer solutions to several problems faced in bioinformatics. Web Services can make it possible for scientists to access biological data and analysis applications residing at different servers in different labs all over the world as if they were installed on their laboratory computers. In addition, Web Services can provide easier integration and interoperability between bioinformatics applications and the data they require from different locations. In the following section, I will be discussing and surveying some of the well known Bioinformatics Web Services.

4 Bioinformatics Web Services

This Web Services features and environment turned out to be the solution to some of the challenges faced in bioinformatics, in terms of integration and automation. Web Services can combine different types of bioinformatics tools available at different location on the Internet into one comprehensive set of bioinformatics services accessible from anywhere at any time. In addition, they provide easier integration and interoperability between bioinformatics applications and the data they require

Web Services technology enables scientists to access biological data and analysis applications as if they were installed on their local laboratory computers. Similarly, it enables programmers to build complex applications without the need to install and maintain the databases and analysis tools. Using Web Services users can browse various data resources and invoke analysis tools available on different computers/servers at different locations from anywhere in the world. In their simplest form, Web Services can provide a middle layer between a database and the user interface. This layer analyzes the user submitted data by intelligent computing or searching against certain databases, and finally provides user the domain knowledge as shown in Fig. 2 [5].

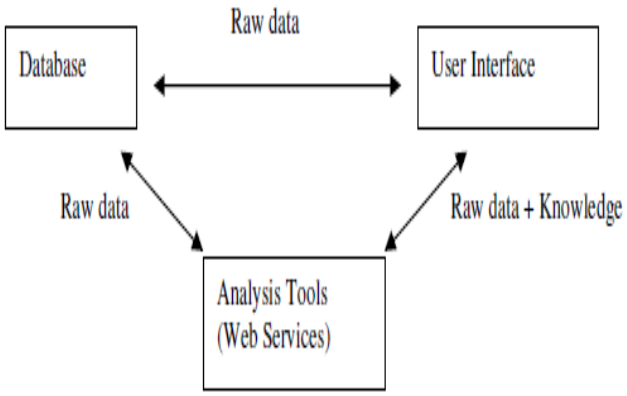

Figure 2. Web Services as a layer of data analysis.

Over the past decade many tools have been generated for the bioinformatics field; however most of these tools are web HTML forms-based tools. This is because it is easy for developers to develop an interface for their program that can be accessed using a web browser than to develop an interface for specific platform. In addition, the use of web browsers as the interface for bioinformatics services makes the development of simple graphical user interfaces relatively easy. Although these tools are very popular, they have a serious disadvantage which is the difficulty of integrating different tools and using different data from different sources to create workflows and data analysis. To overcome this difficulty, the bioinformatics community has generated several tools simplify the developing of workflows using Open Source libraries, such as BioPerl, BioJava and BioRuby [6]. These reusable procedures in different languages allow developers to develop systems for automatic generation of wrappers around web form-based tools to ease the integration of workflows and data from different sources [7, 8]. An example of a web form-based tool that does not need programming skills to use is the Sight project [9]. It is advertised as 'Automatic genomic data-mining without programming skills'. It is a web form analyzer that extracts data from a web form and presents it to the user. The user can then select the data of interest and create an agent from this selection. To create a workflow, it simply connects. However, the main disadvantage of this kind of tools is that each time a service provider updates its interface; the web form analyzer has to be used to reanalyze the interface and fix the corresponding agent.

More advanced tools are required to overcome the integration problems of web form-based tools. The introduction of Extensible Markup Language (XML) provided the solution for simplifying the application integration process. XML is a meta language that has a well-defined syntax and semantics [10]. It is used in the Web Services architecture as the format for transferring information/data between a Web Services provider application and a Web Services client application. It enables developers to separate the content of data exposed over the Web from its presentation. More importantly, XML has been widely accepted as the universal language of choice for exchanging information over the Web and is not the proprietary product of any company. As a result, researchers in bioinformatics can develop new standards for specific functions based on XML. They can define new tags like gene names and biology-specific names and tags.. This main property and others made XML very popular in Bioinformatics Web Services.

In addition to XML, SOAP (*Simple Object Access Protocol)* gained a lot of popularity in the bioinformatics web services community. It is an XML-based protocol for exchanging information in a decentralized, distributed environment [10]. It defines a mechanism to pass commands and parameters between clients and servers. The main reason for its popularity is its simplicity in using the Hyper Text Transfer Protocol (HTTP) for transporting data as messages instead of defining any new protocols. This use of HTTP ensures that Bioinformatics Web Services provider's applications and client applications can communicate using the Internet.

The numbers of Bioinformatics Web Services being developed are increasing every day. At the time of writing this paper, the number of such services listed in the BioCatalog (http://www.biocatalogue.org/) is 2278 services [11]. Most of these services provide programmatic access to data sources and/or algorithmic implementations to analyze biomedical data. These data and the corresponding analysis tools are mainly accessed using browser-based interfaces. They can efficiently answer specific data extraction and analysis needs. However, biomedical problems such as characterizing a gene in terms of a sequence, its translation, expression profile, function and structure requires accessing widely distributed services, exploring and globally evaluating the numerous available data, and the integration and linking of several database information retrieval and analysis services [12]. This tedious task can be achieved using Web Services technologies.

The European Bioinformatics Institute (EBI) has been using Web Services technology to enhance and ease the use of the bioinformatics resources it provides [13, 14] Currently, the European Bioinformatics Institute provides access to more than 200 databases and to about 150 bioinformatics applications.

Some of the well known Bioinformatics Web Services include the followings:

- *ToolBus* is an integrated environment in which bioinformatics data and tools can be interoperable and accessible in an open and flexible manner [5]. It is developed at The Cyber infrastrucure Group (CIG) at the Virginia Bioinformatics Institute.

- *Distributed Annotation System (DAS)* is open source software from biodas.org that provides access to complete genome annotations using a SOAP web interface [15, 16].

- *BLAST*, Basic Local Alignment Search Tool, is a Web Service family of applications that allow biologists and scientists to easily identify and find homologues of an input sequence in DNA and protein sequence libraries [17]. Many genomics laboratories provide a Web-based BLAST interface to their sequence databases for this purpose [18].

- *Pathway Database System* is an integrated system of a set of software tools for modeling, storing, analyzing, visualizing, and querying biological pathways data at different levels of genetic, molecular, and biochemical detail [19].

- *KEGG*, Kyoto Encyclopedia of Genes and Genomes, API was initiated by the Japanese human genome programme in 1995. It uses SOAP based interface to provide access to a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals [20].

- *PDBML*, Protein Data Bank Markup Language, is an XML-based schema for the data in the Protein Data Bank (PDB) [21, 22]. The PDB is a repository for the 3-D structural data of large biological molecules, such as proteins and nucleic acids. One of the members of the PDB organization, Protein Data Bank Japan (PDBj), has developed a tool called xPSSSS that provides a SOAPbased service to retrieve PDBML data [21].

- *MAGE-ML* Server is a tool to map proprietary database schemas for storage of microarray data into Microarray And Gene Expression Markup Language (MAGE-ML) and make them accessible using SOAP [23]. The main objective was to have a standardized Extensible Markup Language format for describing microarray experiments and their results.

- *AGML* Central provides access to databases containing proteomics information in Annotated Gel Markup Language (AGML) using a SOAP interface [24]. It is a web-based open-source public infrastructure for dissemination of two-dimensional Gel Electrophoresis (2-DE) proteomics data in AGML format. It includes a growing collection of converters from proprietary formats to AGML format. A JAVA applet visualizer was developed to visualize the AGML data with cross-reference links. In order to facilitate automated access a SOAP web service is also included in the AGML Central infrastructure.

- *EMBOSS*, European Molecular Biology Open Software Suite, is an Open Source analysis software suite that contains over 200 bioinformatics applications [25]. *Jemboss* is a graphical user interface for the *EMBOSS*, it consists of a client and server both written in Java [26]. The client communicates using SOAP with a Tomcat server that passes requests to the Jemboss server. The Jemboss server can then indirectly execute EMBOSS applications. This Jemboss server could easily be used to provide access via SOAP to other clients than the Jemboss GUI by describing and publishing the interface in WSDL.

- *BioMOBY* is an Open Source project that aims at providing a system for the discovery and processing of biological data using web services [27, 28]. It is emerging as the standard of fact for data exchange and web services inter-communication in bioinformatics. BioMOBY is actually two projects in one: there is Semantic MOBY (S-MOBY) and MOBY Services (MOBY-S). MOBY-S tries to solve the interoperability problem by specifying the syntax and messaging layer to link clients and service providers via information in a central registry. MOBY Services uses SOAP for communication between client, central registry and services. Semantic MOBY takes a little different approach. It tries to solve the interoperability problem by providing a way to clients and providers to describe their data and identify the data relevant to them.

- *MOWServ* is the bioinformatic platform offered by the Spanish National Institute of Bioinformatics to provide integrated access to databases and analytical tools [29]. It is a BioMoby-based web client that enables the secure and integrated analysis of data and straightforward access to databases, services and computational resources.

- *jORCA* is a desktop client aimed at facilitating seamless integration of Web Services [30]. It does so by making a uniform representation of the different web resources, supporting scalable service discovery, and automatic composition of workflows.

- *myGrid* is a project from the UK e-Science Programme funded by the Engineering and Physical Sciences Research Council (EPSRC). All myGrid components are developed in Java and its code base is available as Open Source [31]. It can access several types of services using Java and SOAP. The tool to create workflows for myGrid is called Taverna [32], which can be used to integrate several types of services including web services described by a WSDL document, SOAPlab services, and local applications. To describe a workflow, Taverna uses a custom XML-based language called simple conceptual unified flow language (Scufl).

- *caCORE* is a project developed by the National Cancer Institute Center for Bioinformatics and Information Technology (NCI CBIIT) to provide building blocks for development of interoperable information management systems and aimed at integrating bioinformatics services to support research in cancer biology and medicine [33]. It is an interconnected set of software and services. Enterprise Vocabulary Services (EVS) provide controlled vocabulary, dictionary and thesaurus services. The Cancer Data Standards Repository (caDSR) provides a metadata registry for common data elements. Cancer Bioinformatics Infrastructure Objects (caBIO) implements an object-oriented model of the biomedical domain and provides Java, Simple Object Access Protocol and HTTP–XML application programming interfaces. caCORE has been used to develop scientific applications that bring together data from distinct genomic and clinical science sources.

- *Mummer* is a Web service for genome wide sequence comparison to find Maximum Unique Matches between two sequences. MUMmer 3 is the latest version according to its web site: (http://mummer.sourceforge.net/). It is an open source project based on the mummer algorithm which is a suffix tree algorithm designed to find maximal exact matches of some minimum length between two input sequences [34]. The match lists produced by mummer can be used alone to generate alignment dot plots, or can be passed on to the clustering algorithms for the identification of longer non-exact regions of conservation. These match lists have great versatility because they contain huge amounts of information and can be passed forward to other interpretation programs for clustering, analysis and searching.

The above list is not a complete list of available Bioinformatics Web Services. I have only mentioned some of these services and tools to give an idea about the importance of this new field that combines computer science, information technology, biology, and the internet. In addition, it shows the continuous progress and advancement in this area starting from pure bioinformatics to HTML-based interfaces to bioinformatics arriving to more powerful and beneficial systems of what is called Bioinformatics Web services.

5 Conclusions

Researcher, in general, can easily publish their research results on the internet, compare their findings with others, and building on existing results to make new or more advanced progress. This is true in all fields of research, in general, and in the field of biology in particular. The value of accessing data from other institutions and the relative ease of disseminating this data has increased the opportunity for multi-institution collaborations, which produce dramatically larger data sets than were previously available and require advanced data management techniques for full utilization.

The rapidly emerging field of bioinformatics promises to lead to advances in understanding basic biological processes and, in turn, advances in the diagnosis, treatment, and prevention of many genetic diseases. Bioinformatics has transformed the discipline of biology from a purely lab-based science to an information science as well. Increasingly, biological studies begin with a scientist conducting vast numbers of database and Web site searches to formulate specific hypotheses or to design large-scale experiments. Users can access all data and applications as if they were installed in their local machines, providing seamless integration between disparate services and allowing the construction of workflows to perform complex tasks. However, these benefits come with some unresolved difficulties.

One of these difficulties is the service quality management. Several groups might offer same type of service for redundancy or load balancing, but they may be inconsistent or out of synchronization. In other words, some of these services may be out-of-date. It is currently not possible to discover the most up-to date service. Several servers may host different versions of the database and there might be changes in the available data. To deal with such issues, information about the quality of services needs to be implemented in the tools to handle and query the service directories. As an example, BioMOBY requires web service providers to register their services in a central repository. Service providers are expected to make sure that the information for their services is kept up to date.

Another difficulty has to do with standardization. Most, if not all Bioinformatics Web Services take advantages of the extensibility of XML to define their own tags to describe biology data. This extensibility feature of XML turns out to have a fire back effect in a sense that it enables scientists to describe every piece of data in the bioinformatics domain in XML by choosing different extensions for the same type of data. The problem surfaces when linking and integrating different services to form workflows to analyze collected data, and these set of data need to be converted from one XML schema into another. Therefore there is a need for standards in the area of bioinformatics or some kind of code of conduct for service providers such as the one proposed in [35] to prevent unnecessary and inefficient conversions between different data formats or tags. Although Web Services main feature is the ability to integrate and link data and tools with different formats. But this will only be efficient if the bioinformatics developers can reach consensus on one or couple of standards to describe bioinformatics data.

6 References

[1] A. Labarga, F. Valentin, M. Anderson and R. Lopez, Web Services at the European Bioinformatics Institute, Nucleic Acids Research, Web Server issue, Vol. 35, 2007.

[2] Mohamad Ladan, "Web Services: Technologies and Benefits" *Journal of Communication and Computer,* ISSN 1548-7709, Number 6, Volume 7, 2010.

[3] National Center for Biotechnology Information. http://www.ncbi.nih.gov/

[4] A Science Primer. Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources. http://www.ncbi.nlm.nih.gov, March 29, 2004

[5] B. Yang, J Eckart, E. Nordberg and B. Sobral. ToolBus: An Interoperable Environment for Biological Researchers,Proceedings of The 2005 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '05), p274, Las Vegas, NV, June 2005.

[6] Open Bioinformatics Foundation. http://www.open-bio.org/

[7] Rocco, D. and Critchlow, T. (2003), 'Automatic discovery and classification of bioinformatics web sources', Bioinformatics, Vol. 19, pp. 1927–1933.

[8] Kossenkov, A., Manion, F. J., Korotkov, E. et al. (2003), 'ASAP: Automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database', Bioinformatics, Vol. 19, pp. 675–676.

[9] Meskauskas, A., Lehmann-Horn, F. and Jurkat-Rott, K. 'Sight: Automating genomic data-mining without programming skills', Bioinformatics, Vol. 20, pp. 1718–1720, 2004.

[10] Mohamad Ladan ,"An Overview of XML and a Comparison with HTML and SGML", International Conference on Research Trends in Science and Technology, RTST 2002, Beirut and Byblos, Lebanon. March 4 - 6, 2002.

[11] Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., Goble, C.A.: BioCatalogue: a universal catalogue of web services for the life sciences, Nucl. Acids Res., 2010.

[12] Marco Masseroli, Giorgio Ghisalberti, and Stefano Ceri, Bio Search Computing: Bioinformatics web service integration for data-driven answering of complex Life Science questions. *Procedia Computer Science*, Volume 4, 2011, Pages 1082-1091.

[13] Pillai,S., Silventoinen,V., Kallio,K., Senger,M., Sobhany,S., Tate,J., Velankar,S., Golovin,A., Henrick,K. et al. (2005) SOAP-based services provided by the European Bioinformatics Institute. Nucleic Acids Res., 33, 25–28.

[14] Harte,N., Silventoinen,V., Quevillon,E., Robinson,S., Kallio,K., Fustero,X., Patel,P., Jokinen,P. and Lopez,R. Public web-based services from the European Bioinformatics Institute. Nucleic Acids Res., 32, 3–9. 2004.

[15] Dowell, R., Jokerst, R. M., Day, A. et al. 'The distributed annotation system', BMC Bioinformatics, Vol. 2, p. 7. 2001.

[16] BioDAS. http://biodas.org/

[17] Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology,* 215 (3), 403– 410. 1990

[18] Gish, W. (2002). BLAST. http://blast.wustl.edu/

[19] Krishnamurthy, L., Nadeau, J., Ozsoyoglu, G. et al. 'Pathways database system: An integrated system for biological pathways', Bioinformatics, Vol. 19, pp. 930–937. 2003.

[20] Kawashima, S., Katayama, T., Sato, Y. and Kanehisa, M., 'KEGG API: A web service using SOAP/WSDL to access the KEGG system', Genome Informatics, Vol. 14, pp. 673–674. 2003.

[21] Westbrook, J., Ito, N., Nakamura, H. et al., 'PDBML: The representation of archival macromolecular structure data in XML', Bioinformatics, Vol. 21, pp. 988–992. 2005.

[22] Bernstein, F. C., Koetzle, T. F., Williams, G. J. et al., 'The Protein Data Bank: A computer-based archival file for macromolecular structures', J. Mol. Biol., Vol. 112, pp. 535–542. 1977.

[23] Tjandra, D., Wong, S., Shen, W. et al., 'An XML message broker framework for exchange and integration of microarray data', Bioinformatics, Vol. 19, pp. 1844–1845. 2003.

[24] Stanislaus, R., Chen, C., Franklin, J. et al., 'AGML central: Web based gel proteomic infrastructure', Bioinformatics. 2005.

[25] Rice, P., Longden, I. and Bleasby, A., 'EMBOSS: The European Molecular Biology Open Software Suite', Trends Genet. Vol. 16, pp. 276–277. 2000.

[26] Carver, T. and Bleasby, A., 'The design of Jemboss: A graphical user interface to EMBOSS', Bioinformatics, Vol. 19, pp. 1837–1843. 2003.

[27] Wilkinson, M. D. and Links, M., 'BioMOBY: An open source biological web services proposal', Brief. Bioinform., Vol. 3, pp. 331–341. 2002.

[28] Wilkinson, D., Gessler, D., Farmer, A. and Stein, L.. 'The BioMOBY Project explores open-source, simple, extensible protocols for enabling biological database interoperability', in 'Proceedings of the Virtual Conference on Genomics and Bioinformatics', Vol. 3, pp. 17–27 , 2003.

[29] Sergio Ramírez, Antonio Muñoz-Mérida, Johan Karlsson, Maximiliano García, Antonio J. Pérez-Pulido, M. Gonzalo Claros, Oswaldo Trelles, MOWServ: a web client for integration of bioinformatic resources, Nucleic Acids Res., Web Server issue 38, July 1, 2010.

[30] Martín-Requena V, Ríos J, García M, Ramírez S, Trelles O. *jorca: easily integrating bioinformatics web services.* Bioinformatics; 26:553-559, 2010.

[31] Stevens, R. D., Robinson, A. J. and Goble, C. A., 'myGrid: Personalised bioinformatics on the information grid', Bioinformatics, Vol. 19. 2003.

[32] Oinn, T., Addis, M., Ferris, J. et al. , 'Taverna: A tool for the composition and enactment of bioinformatics workflows', Bioinformatics, Vol. 20, pp. 3045–3054. 2004.

[33] Covitz, P. A., Hartel, F., Schaefer, C. et al., 'caCORE: A common infrastructure for cancer informatics', Bioinformatics, Vol. 19, pp. 2404–2412. 2003.

[34] Bin Hu, Gary Xie, Chien-Chi Lo, Shawn R. Starkenburg, and Patrick S. G. Chain Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics *Briefings in Functional Genomics,* 10(6): 322-333, 2011.

[35] Stein, L., 'Creating a bioinformatics nation', Nature, Vol. 417, pp. k119–120. 2002.