

Genome Annotation and Comparison System

*Jing Zhao, *Tian Xue, Boyu Yang, Kelly Williams, Alice R. Wattam,
Rebecca Will, Bruce Sharp, Ron Kenyon,
Oswald Crasta, Bruno W. Sobral

Virginia Bioinformatics Institute
Washington Street
Blacksburg, VA 24061

Abstract – The VBI GenomeACS (VBI Genome Annotation and Comparison System) is a genome annotation and comparison system for prokaryotes and eukaryotes. It has been developed by the Virginia Bioinformatics Institute's (VBI) Cyberinfrastructure Group (CIG) as part of the Pathogen Portal (PathPort) project. Backed by a scalable genome relational database VBI GenomeDB and a Web service business layer, the system provides an extensible and user friendly framework for VBI's genome curation efforts, with a genome browser interface that allows the user to view, search and compare complete genomes, including plasmids and eukaryote organelles.

Key words: Genome, Web Services, PathPort, Cyberinfrastructure, Bioinformatics, Computational Biology

*** Equal Contribution**

1 Introduction

Genomics is a major focus in bioinformatics. Challenge involved in genome annotation and analysis include the ever increasing amount of genome annotation data, and the complexity and diversity of genome analysis tools. VBI's Cyberinfrastructure Group has designed and developed VBI GenomeACS (VBI Genome Annotation and Comparison System) as a comprehensive and extensible software infrastructure aiming to provide data scalability, capability of running various genome analysis software tools within a single application, and the flexibility of integrating new tools without any change of the software infrastructure.

VBI GenomeACS was built within the ToolBus/PathPort [1,2] systems, which provide a generic Web Service framework to integrate backend data sources through a business logic layer, and present data to a front end visualization suite (refer to Section 5, Architecture Design, and Figure 1 diagram for details). The Web Service technology achieves service layer platform independence and easy distribution over the Web.

The VBI GenomeACS is designed to serve scientific users. For example, there has been a requirement from genomic researchers for the capability to view the entire genome. This requirement includes not only the core nuclear chromosomes of a specific organism, but also plasmids or organelles. An organism's complete genome may contain multiple chromosomes, plasmid components for certain bacteria, and in the case of eukaryotes, extranuclear chromosome fragments such as mitochondria and plastid DNAs. Automatic grouping of those individual chromosome pieces poses a technical challenge since our major data source, RefSeq at the National Center for Biological Information (NCBI) [3], only provides GenBank data files at the single-replicon level. A manual grouping or any hard-coded approach would be too costly, especially for prokaryotes and viruses. To meet this challenge, VBI GenomeACS implements a taxonomy-based checking mechanism, the details of which will be provided in Section 3.

The Web service layer of VBI GenomeACS aims for a broader utility in building a flexible bioinformatics infrastructure. Bioinformatics applications are critical for biological research, and many of those tools are available and continue to be developed. With rapid increase of biological data and more tools become available, there needs a bioinformatics tools framework for tools integration. Sometimes, different tools are used together to provide better analysis into biological data, this requires not only data transformation between tools, but also requires those tools to be seamlessly integrated with output of one tool to be input of another.

At VBI, Web services and XML have been chosen as the core technology for an extensible, scalable and open standard based bioinformatics tools framework. First, all tools functionalities are exposed as Web services, and new tools are easily added into the framework. Second, the executions of those Web services are coordinated under one application framework. Finally, compatible XML documents are used to communicate among the different Web services, thus facilitating the automated orchestration of these sometimes disparate services. VBIGenomeACS is part of this tools framework.

2 Genome Database

With the scalability goal in mind, VBIGenomeDB was designed to use a distributed architecture. The system consists of a central registration database and multiple satellite category databases. At present, the following eight categories are available, and these include genomes from bacteria/archaea, viruses, *Plasmodium*, *Drosophila melanogaster*, *Rattus norvegicus*, *Mus musculus*, as well as the organelles, which include mitochondria and plastid sequences. The central registration database provides the URL of each of these eight databases, as well as an entry point for each segment accession identifier. The eight database instances share an identical database schema, allowing a generic implementation for the Web service to fetch a detailed annotation from any of the eight databases on-the-fly. Databases from additional categories can be easily incorporated into the existing system with a minimal addition at the central registration database, requiring no change at the Web service layer.

Currently, the main data source for VBIGenomeDB is NCBI's RefSeq database [3], an integrated non-redundant set of sequences that includes genomic DNA, RNA, and proteins. A Perl script was developed to extract data from GenBank flat files that support a number of predefined feature types such as CDS, rRNA, tRNA, repeat, etc. Each feature type may have a rich set of qualifiers and values that include gene, function, product, db_xref, and translation, as examples [4]. The database schema uses a generic mechanism to store all feature types, qualifier types and values, therefore there is no need to hard code predefined types. Thus the database and Web service developers do not need to be concerned about missing data or exceptions that might result from an update of the predefined feature/qualifier type by NCBI. Another benefit of the generic schema is that other data sources, such as EMBL, FlyBase, Human Genome Project, or even unpublished customer datasets, can be plugged into the VBIGenomeDB without difficulty.

VBIGenomeDB hosts both prokaryote genomes and eukaryote genomes. For ongoing VBI curation projects, contig assembly information was retrieved by a Perl script that processes the seq_contig.md file for each available organism and stores it in the central registration database. Our GenomeTool Web service obtains these contig coordinates from the Genome Annotation Database then passes them on to the client side for chromosome assembly [5, 6].

3 Genome Grouping

The ability to view an entire genome for a specific organism, including plasmids and organelle DNAs together with the core nuclear chromosomes is desirable for biologists and bioinformaticians in whole-genome study. Automatic grouping of those individual chromosome pieces was achieved by a taxonomy-based approach designed by VBI's Cyberinfrastructure Group. The idea is that any sequences that have the same NCBI taxonomy id would be from the same organism and therefore should be grouped together.

The approach works very well for grouping multiple core chromosomes within one organism. For example, NC_003317 (*Brucella melitensis* chromosome I) and NC_003318 (*Brucella melitensis* chromosome II) share the taxonomy id 224914. It also works well for grouping bacterial chromosomes with relevant plasmids, such as NC_000918 (*Aquifex aeolicus*) and NC_001880 (*Aquifex aeolicus* plasmid eel1), which share the taxonomy id 224324.

The more complicated eukaryotic organisms require a different approach. For the *Plasmodium* category, which is composed of the complete genome of *Plasmodium falciparum*, grouping the mitochondria (NC_002375) and plastid (X95275, X95276) genomes together with the nuclear genomes required an extension of the search that included not only the taxonomy id, but also the direct parent and child taxonomy ids within NCBI's taxonomy hierarchy [7]. In this case, NC_002375, X95275, X95276 have the same taxonomy id 5833 while the other fourteen regular chromosomes of *Plasmodium falciparum* 3D7 use a more specific taxonomy id 36329, a direct child of taxonomy id 5833.

Although using the taxonomy id has worked well in the majority of cases, an exception arises in the few cases where independent projects have sequenced nearly identical genomes that have been assigned the same taxonomy id, as for *Agrobacterium tumefaciens* str. C58. The sequences from Cereon Genomics included the RefSeq numbers NC_003062,

NC_003063, NC_003064, NC_003065 , and NC_003304, NC_003305, NC_003306, NC_003308 are from the University of Washington sequencing project As shown in Table 1, eight components would have been grouped together if grouping was made solely on taxonomy id. Our solution to this problem was to apply a stronger checking mechanism that combines the taxonomy id and sequence definition as well as the literature reference.

Table 1. *Agrobacterium tumefaciens* str. C58, a sample case of taxonomy based grouping

Accession	NCBI Taxonomy Id	NCBI Sequence Definition	Sequenced by
NC_003062	176299	<i>Agrobacterium tumefaciens</i> str. C58 chromosome circular, complete sequence.	Cereon
NC_003063	176299	<i>Agrobacterium tumefaciens</i> str. C58 chromosome linear, complete sequence.	Cereon
NC_003064	176299	<i>Agrobacterium tumefaciens</i> str. C58 plasmid AT, complete sequence.	Cereon
NC_003065	176299	<i>Agrobacterium tumefaciens</i> str. C58 plasmid Ti, complete sequence.	Cereon
NC_003304	176299	<i>Agrobacterium tumefaciens</i> str. C58 chromosome circular, complete sequence	U. Washington
NC_003305	176299	<i>Agrobacterium tumefaciens</i> str. C58 chromosome linear, complete sequence.	U. Washington
NC_003306	176299	<i>Agrobacterium tumefaciens</i> str. C58 plasmid AT, complete sequence.	U. Washington
NC_003308	176299	<i>Agrobacterium tumefaciens</i> str. C58 plasmid Ti, complete sequence.	U. Washington

The grouping algorithm is summarized as:

- 1) Loop through the accessions and group together segments with the same taxonomy ids.
- 2) Within each taxonomy id group, (in the sample case, 176299, consisting of 8 segments), compare the sequence definitions.
- 3) If any multiple segments have identical definition but different accession IDs, (in the sample case, see NC_003062 and NC_003304 as an example), go check their respective literature references.
- 4) If references are not the same, this suggests that they may be the same chromosomes from different sequencing projects and therefore should not be grouped together.

4 Web Service Realization

The VBIgenomeACS server-side design is based upon the ubiquitous use of Web services [8] and XML messaging. All business functionalities such as database query or sequence similarity search are presented as a Web service; this includes native programs, publicly available tools and services, and VBI implemented tools and backend databases. This design allows for the addition of new tools, or new implementations for existing tools without affecting existing business interfaces. It also provides easy distributions of Web services to different machines. Web services also provide greater flexibility for the client (ToolBus, for example) since users can choose any Web service that they wish to access, including those of their own development and those they have discovered via Universal Description, Discovery and Integration (UDDI) [9] or similar repository directories.

GenomeTool is also part of this package. It is a genome annotation tool that provides a genome sequence with feature annotations. Used in ToolBus/PathPort, the genome information returned from this tool are made available in a genome viewer that allows visualization of sequence features and the results of analyses within the context of the sequence, as well as its six-frame translation. The VBIgenomeACS provides other sequence analysis tools that aid researchers in developing further annotations of the genomic information. These include gene prediction programs like GlimmerTool [10] and GeneMarkTool [11], sequence alignment programs (WaterTool [12], StretcherTool [13], MSATool [14]), and whole genome comparison tools (MummerTool [15]).

5 Architectural Design

The VBIGenomeACS server-side architecture is designed around Web service technology and all business functionalities are exposed as Web services. It has a multi-tier architecture with client/presentation, web service, business logic, and backend data layers. XML messages are used for communication between the client layer and Web services layer. The following is a brief introduction to each tier and how they work together.

The client/presentation layer is implemented via ToolBus, a rich Java client that analyzes and visualizes data returned from the VBIGenomeACS server side. It is not tied to any particular server, but instead can easily contact any servers from which compatible Web services are deployed. This makes ToolBus highly adaptable. When one server environment is down, ToolBus can switch to other active servers. The communication between the client and server is via Simple Object Access Protocol (SOAP) [16] over HTTPS.

The Web service layer is the interface between the client and server business functionalities. This interface abstracts away the actual implementations and locations of server functionalities so that changes in implementation details or changes in location do not affect clients. Apache Axis [17] is used as the Web service application framework and is essentially the Web services engine.

The business logic layer is where the data are retrieved and processed from backend databases, or from processing results of native programs, or from publicly available services.

The data layer is where data are stored or processed. Other than native programs or public databases, VBI also has some specialized databases that may provide data that has been more recently updated than the NCBI RefSeq data present in the VBIGenomeDB. Separating the server architecture into different layers improves flexibility, availability, manageability, and performance.

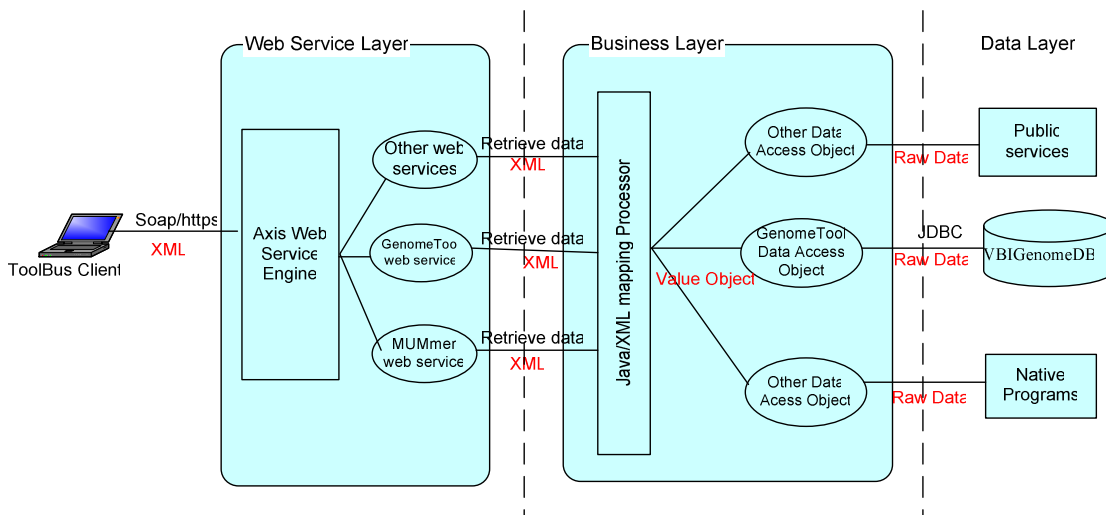


Fig. 1 VBIGenomeACS server-side architectural design

6 Use Case Example 1

The VBI Genome Annotation System offers an interface that allows the user to search for an organism of interest. During Web service initialization, it performs a dynamic all-ID search based on the organism type (See Fig. 2).

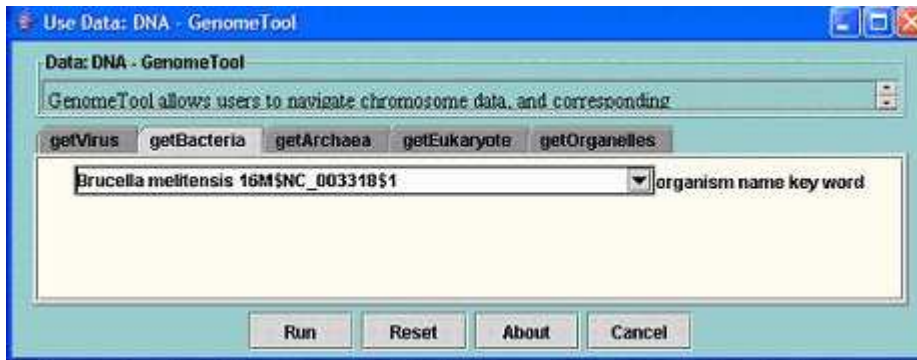


Fig. 2 GenomeTool user interface. User select organism: *Brucella melitensis* 16M

The GenomeTool Web service returns two *Brucella melitensis* 16M genomes by grouping taxonomy_ids. The visualization results start with a summary of chromosomes for a specific organism and provide the ability to zoom into specific regions to explore the genome at the sequence level, and also levels where genes or CDS features are available (See Fig. 3).

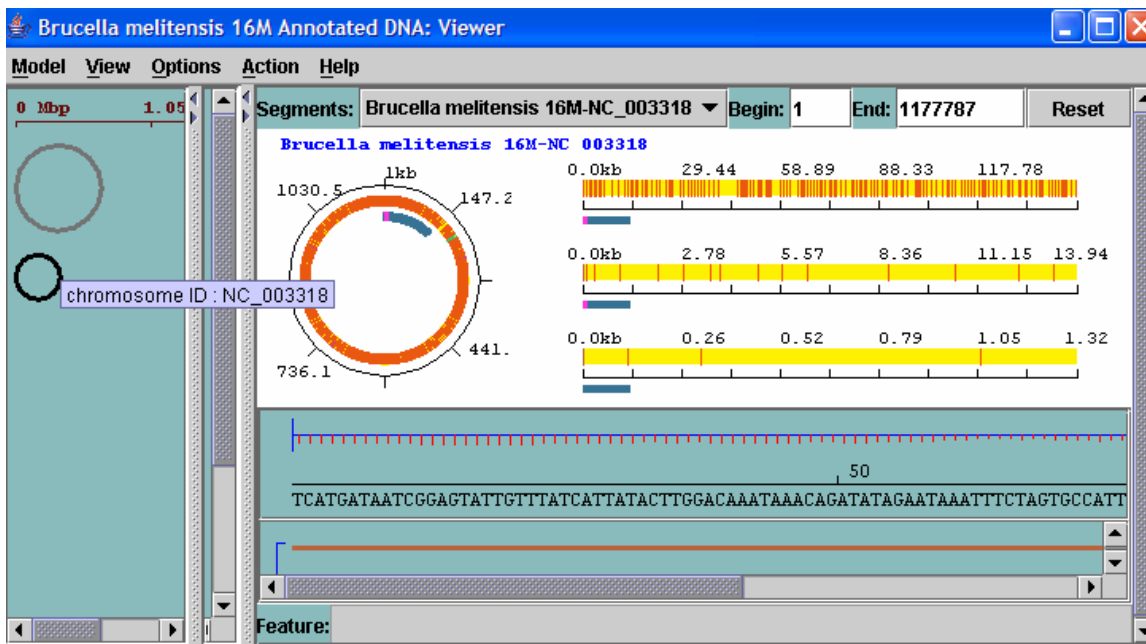


Fig. 3 *Brucella melitensis* 16M-genome information (left panel) and feature information (right panel)

7 Use Case Example 2

The Mummer Web service, which wraps the MUMmer native program, was developed to provide genome-wide sequence comparison. When the Mummer Web service is initialized, it invokes the GenomeTool Web service to dynamically retrieve all sequence ids of bacteria or virus from the VBIGenomeDB. An example of this comparison can be seen when MUMmer is used to compare sequence similarities of two *Brucella* strains (*Brucella melitensis* 16M-NC_003317 and *Brucella suis* 1330-NC_004310).in Fig. 4.

8 Conclusion and Discussions

The overall design including the Web services architecture, the design of theVBIGenomeDB, and the use of XML messaging to communicate among different bioinformatics tools makes the VBI Genome Anotation and Comparison

System a powerful and versatile tool for the biological researcher. This system offers a generic Web services architecture and bioinformatics applications that are presented as easy-to-use services for client applications. Furthermore, independently developed but related applications may be linked as a workflow process under this framework and offer functionalities that are not possible individually, a feature that is very useful to biologists that are not adept programmers. Since Web services can be distributed and discovered, they provide high availability to the biological research communities.

9 Future Work

Using taxonomy ids at NCBI GenBank is not guaranteed to meet every need of the bioinformatics researcher. Therefore even the most sophisticated grouping algorithm may still result in exceptions on some occasions. The taxonomy-based automatic grouping algorithm may need further refinement when new exceptions are detected.

Support for annotation edition and version control was built into the database schema and the implementation of this feature in the Web service layer is in progress.

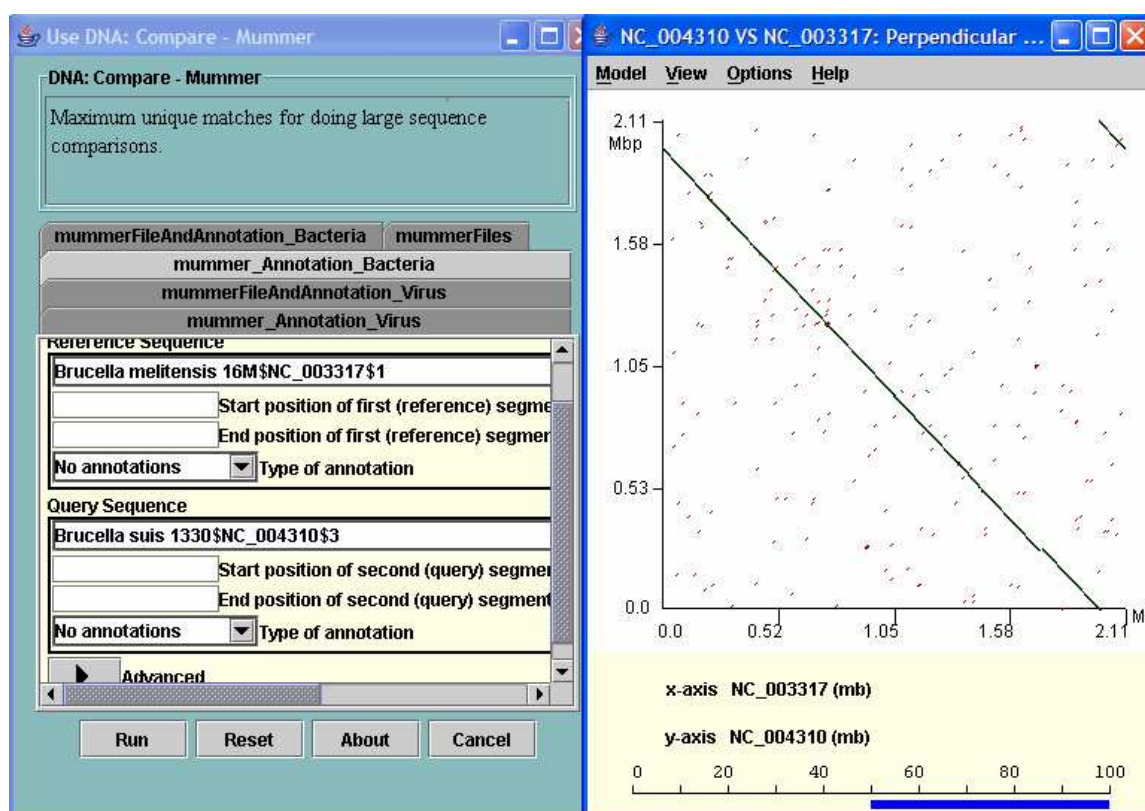


Fig. 4 Mummer Web service user interface (left panel) with the user selecting *Brucella melitensis* 16M-NC_003317 against *Brucella suis* 1330-NC_004310. Information on both sequences is retrieved by the Mummer Web service by invoking the GenomeTool Web service. Via ToolBus, the sequence comparison model identifies with Mummer Web service results. This is a perpendicular view (right panel) showed for comparing similarities of NC_003317 and NC_004310. For this dot plot, the reference sequence (NC_003317) is mapped across the x-axis, while the query sequence (NC_004310) is on the y-axis. Wherever the two sequences agree, a colored line or dot is plotted. The forward matches are displayed in red, while the reverse matches are displayed in black.

10 Acknowledgments

This work is supported by US Department of Defense, Grant number: W911SR-04-0045 to Bruno W. S. Sobral. The development of VBIGenomeACS has benefited from the following open source code: MySQL, BioPerl, Tomcat and Axis from Apache, WSDL4J and UDDI4J from IBM, and Castor from Exolab.

11 Reference

- [1] J Dana Eckart, and Bruno W. S. Sobral, A Life Scientists Gateway to Distributed Data Management and Computing: The PathPort/ToolBus Framework. *OMICS: A Journal of Integrative Biology*, Volume 7, Number 1, pp 79-88 (2003).
- [2] Boyu Yang, Eric K. Nordberg, J Dana Eckart, and Bruno W. S. Sobral, ToolBus - An Interoperable Environment for Biological Researchers. 2005 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '05).
- [3] Pruitt KD, Tatusova, T, Maglott DR, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, Jan 1;33(1):D501-D504 (2005).
- [4] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL., GenBank. *Nucleic Acids Res.*, Jan 1;28(1):15-18 (2000).
- [5] Assembling Genomic Sequences. <http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html>.
- [6] NCBI Genomic Sequence Assembly and Annotation Process. <http://www.ncbi.nlm.nih.gov/genome/guide/build.html>
- [7] Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, Jan 1;28(1):10-4 (2000).
- [8] Web Service: <http://java.sun.com/webservices/>.
- [9] Universal Description, Discovery and Integration (UDDI): <http://www.uddi.org>.
- [10] Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L., Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27, 4636-41 (1999).
- [11] Borodovsky, M. & McIninch, J., GeneMark: Parallel Gene Recognition for both DNA Strands. *Computers & Chemistry* 17, 123-133 (1993).
- [12] Smith TF, Waterman MS, Identification of common molecular subsequences. *J. Mol. Biol* 147(1);195-7 (1981).
- [13] E. Myers and W. Miller, Optimal Alignments in Linear Space. *CABIOS* 4, 1, 11-17 (1988).
- [14] Thompson, J.D., Higgins, D.G. & Gibson, T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-80 (1994).
- [15] Arthur L. Delcher, Simon Kasif, Robert D. Fleischmann, Jeremy Peterson, Owen White and Steven L. Salzberg, Alignment of whole genomes. *Nucleic Acids Research*, Vol. 27, No.11 2369-2376 (1999).
- [16] Simple Object Access Protocol (SOAP): <http://www.w3.org/TR/soap/>.
- [17] Apache Axis: <http://xml.apache.org/axis/>.