

# Bioinformatics Web Services

**Boyu Yang, Tian Xue, Jing Zhao, Chaitanya Kommidi, Jeetendra Soneja, Jian Li,  
Rebecca Will, Bruce Sharp, Ron Kenyon,  
Oswald Crasta, Bruno W. Sobral**

Virginia Bioinformatics Institute  
Washington Street  
Blacksburg, VA 24061, U.S.A.

*Abstract - Web services is distributed computing technology that provides software services over the Web. The Cyberinfrastructure Group at the Virginia Bioinformatics Institute (VBI) has adopted the Web service architecture and wrapped a wide variety of bioinformatics applications and data resources using the technology. Users can browse various data resources and invoke analysis tools through the services deployed at VBI. We have also developed an integrated client environment, ToolBus, which is available from <http://pathport.vbi.vt.edu/download>. A use case is given to demonstrate the use of Web services through ToolBus. A complete list of Web services is included.*

**Key words:** Bioinformatics, ToolBus, Web services

## 1 Introduction

The amount of biological data is increasing rapidly due to the development of high throughput technologies. However, an increase in the amount of data does not automatically lead to an increase in the amount of biological knowledge unless it is accompanied with new or improved analytical tools. Bioinformatics data and analysis tools should be hosted centrally for performance, maintenance, management and access; however that has not largely been the way bioinformatics developments have occurred.

Web services are a distributed computing technology that provides software services over the web. They have the following unique features: transport over standard protocols like HTTP; messages are transmitted in XML format, so they are platform and language independent; a standard public interface described by WSDL; publishing and retrieving through a UDDI registry; and legacy software can easily be wrapped into Web services with limited effort. These features make Web services a good choice for deploying our bioinformatics applications.

The Cyberinfrastructure Group (CIG) at the Virginia Bioinformatics Institute (VBI) has adopted Web services technology and wrapped a wide variety of bioinformatics applications and data resources using the technology. An integrated client environment, ToolBus[1], has also been developed for accessing these Web services. Users can browse various data resources and invoke analysis tools deployed at VBI from anywhere in the world by using these services. In addition to accessing raw data in databases, 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. 1.

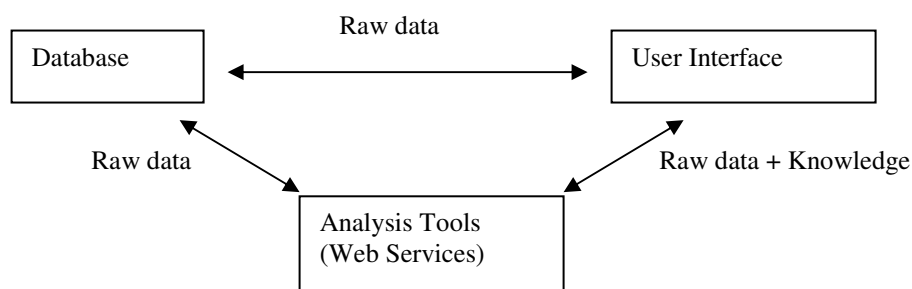


Fig. 1 Web services as a layer of data analysis

## 2 VBI Web Services Strategy

We have specifically addressed availability, compatibility, usability, security and performance issues in our Web services development.

*Availability:* VBI Web services are replicated on multiple, geographically distributed servers to help ensure the availability in case of a server outage. ToolBus, the interoperable client software developed by VBI, can automatically redirect service requests to another server when the current contacted server is not accessible. Details about ToolBus can be found on the PathPort web site (<http://pathport.vbi.vt.edu/>).

*Compatibility:* We follow industry standards as specified by W3C[2] in development of our web services, this means other parties can easily access our web services programmatically by following the W3C standards.

*Usability:* In order to make our Web services user-friendly, we provide a new function to compensate for a WSDL deficiency that default values, restrictions, and semantic descriptions of parameters are not specified, which makes it difficult for a user to enter an appropriate and correct value for a parameter. We developed a WSOPSL – Web Service OperationS Language, which provides default values, restrictions and semantic descriptions for all parameters of all operations. This allows a user to clearly understand the meaning of a parameter and select appropriate parameter values with a few mouse clicks. The WSOPSL document for each web service is independent of WSDL and can be obtained by a single method invoke, hence it does not break the industry standards. Another benefit of this extension is that it allows sequential execution of several operations in a Web service when used with ToolBus. This capability enables some basic pipeline functionality.

*Security:* Due to licensing restrictions or special computational resource requirements, we need to restrict some Web services to be accessible only by authorized users. AAA (authorization, authentication, and accounting) ticket system has been developed and deployed. Only those users with an AAA account can access the restricted Web services.

*Performance:* Many biological applications must process huge amount of data or need complex scientific computation, and these applications usually take a very long time. VBI Web services support a polling mechanism that allows the client software to submit a job with a unique job ID. When a Web service receives a job and finds that the job may take a long time to finish, it will inform the client by sending a polling message that contains an estimated completion time. The network connection will be closed and resources are freed to allow the client to process other tasks. The client can request the results at a later time when the service has finished by using the job ID. In addition, large datasets are compressed into zip format and transmitted as an attachment. Binary data, for example graphics, are also transmitted as an attachment.

## 3 Bioinformatics Web Services Available from VBI's CIG

After several years of development, a wide variety of Web services are now available, as listed in Table 1 by category (next page). The full list of services, WSDL URLs, and their descriptions can be found at <http://pathport.vbi.vt.edu/services>.

## 4 An Example Use Case - Comparing Genomes Using Mummer

This example demonstrates the use of ToolBus to access VBI Web services. Mummer[36] is a Web service we provide for genome wide sequence comparison to find Maximum Unique Matches(MUM) between two sequences. For this example we will use Mummer to compare one subtype of human flu virus (H1N1) and one subtype of avian flu virus (H5N1) to discover their sequences similarities.

First, launch ToolBus and select DNA: Compare – Mummer from the Tools panel as shown in Fig. 2. Then, double click on Mummer or select Mummer and click the *Use* button to invoke this Web service (See Fig. 3). The user interface allows the user to select the desired operation and parameters (Fig 3). There are five operations available: `mummer_Annotation_Virus` for selecting virus species from a database for comparison; `mummer_Annotation_Bacteria` for selecting bacteria species from a database for comparison; `MummerFileAnd-Annotation_Virus` and `mummerFileAndAnnotation_Bacteria` for selecting one sequence from the user's file system and another sequence from the database; and `MummerFiles` for selecting both sequences from the user's file system. In this example we select `mummer_Annotation_Virus` operation. The first parameter specifies the output format, either MUMs or Extended MUMs. Within the Reference Sequence and Query Sequence boxes, we select the species, their start and end positions and the type of annotation. Below the Query Sequence box, there is an *Advanced* button, which is not shown in the picture. Click the *Advanced* button and a group of advanced parameters is displayed; we set the minimum length of match to 10. Click the Run button to call the Web service. When the query result returns, a model chooser window pops up and the ToolBus discovered data model is displayed.

Table 1 VBI developed or wrapped tools

Data Browsing:	Sequence Alignment or Search:
Genome Browsing	BLAST(Cluster Server) [29]
Pathgen Background Information	BLAST(TimeLogic server)
GeneOntology	FASTA [30]
Protein Interaction Network	ClustalW[31]
Phylogeny:	Smith-Waterman [32]
Phylogenetic Tree Browsing	Stretcher[33]
Phylip[3]	Sean (find potential SNP)[34]
Gene Prediction:	Ssaha and SsahaSNP [35]
Genscan[4]	MUMmer[36]
Glimmer [5]	Hmmer[37]
GlimmerM[6]	Rfam/Infernal[38]
GlimmerHMM[7]	Cognitor[39]
TigrScan[7]	Blat[40]
Fgenesh[8]	Blocks[41]
GrailEXP[9]	Microarray analysis:
Orpheus[10]	Geneidmap(Gene alias search)
rRNA Scan[11][12]	Agnes (Cluster program in R [42])
tRNA Scan[13]	Hclust (Cluster program in R)
GeneMark[14]	Kmean (Cluster program in R)
SNAP[15]	Rpca (PCA classification in R)
Tfscan [16]	Rsom (Gene SOM in R)
Rankgene[17]	Anova (Anova analysis in R)
UNVEIL[18]	Cluster3.0[43]
TICO[19]	SAM(Significance Analysis in R)
Est2Genome[20]	Diana (Hierarchical clustering in R)
Probe Design:	Multtest (f/t tests, bioconductor [44])
PCR/Hybridization [21]	Rsvm (SVM classification in R)
YODA [22]	Rfda (Discriminant analysis in R)
Sequence Assembly:	Rknn (KNN classification in R)
Contigs from trace files [23]	Rlda (Supervised classification in R)
Sequence Cut:	Protein prediction:
Restriction enzymes [24]	InterProScan[45]
Structure Prediction:	Lipop[46]
Psipred[25]	ProteinPredict(P2SL)[47]
Memsat[26]	FindPept[48]
rbsFinder[27]	MSFit[49]
SignalP[28]	MSBlast[50]

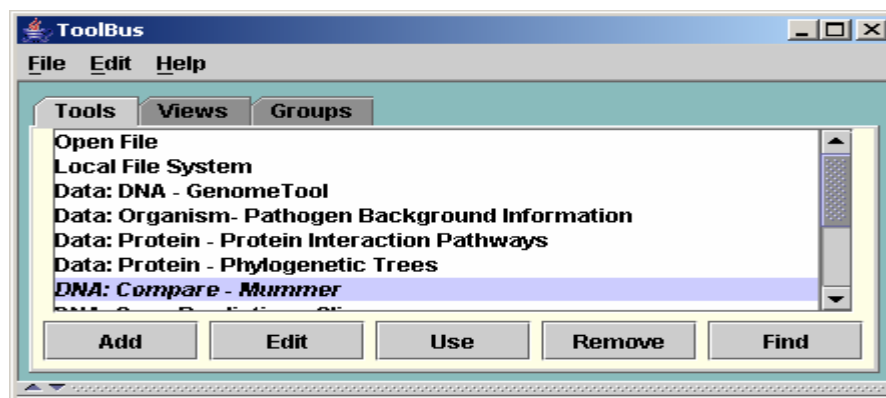


Fig. 2 ToolBus main window at startup

The data model for this result is called Sequence Comparison. Select this data model, then click OK, and the default view (perpendicular view) is displayed. A DAS [51] based parallel view is also available for selection under the View menu. Figures 4 and 5 show the two views separately.

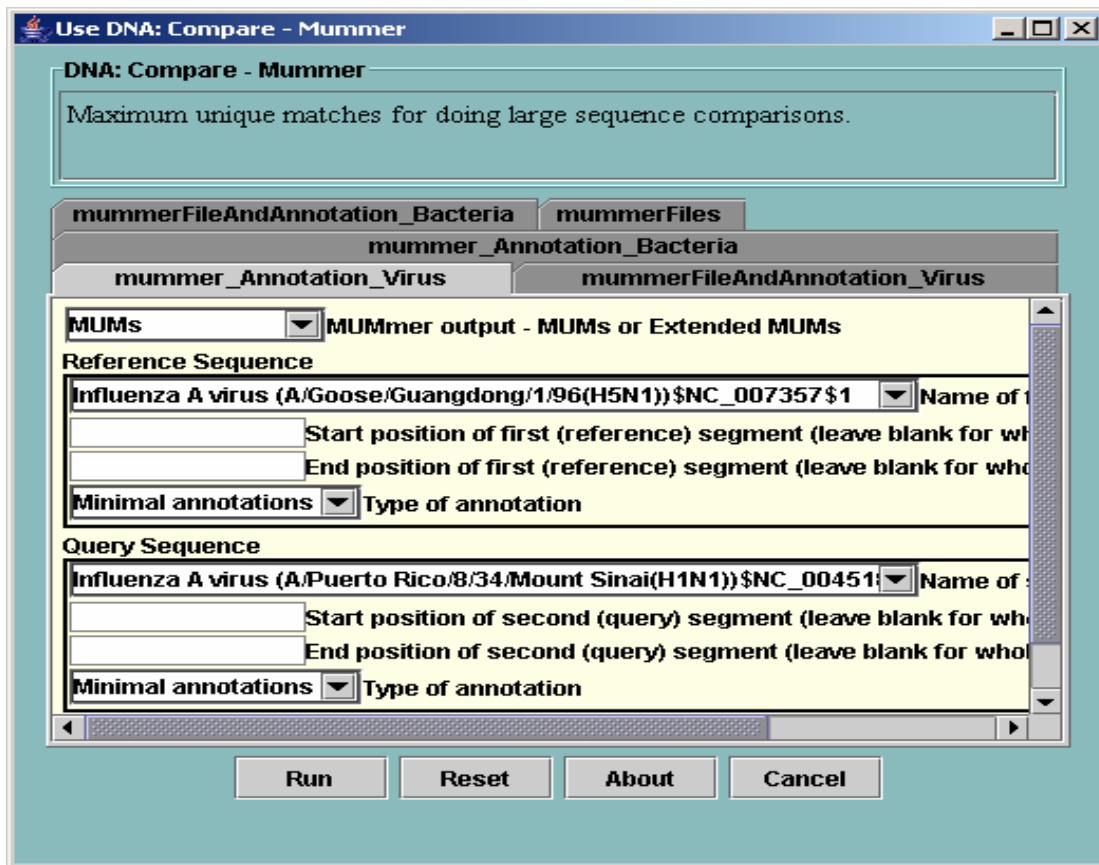


Fig. 3 ToolBus User interface for the Mummer Web service

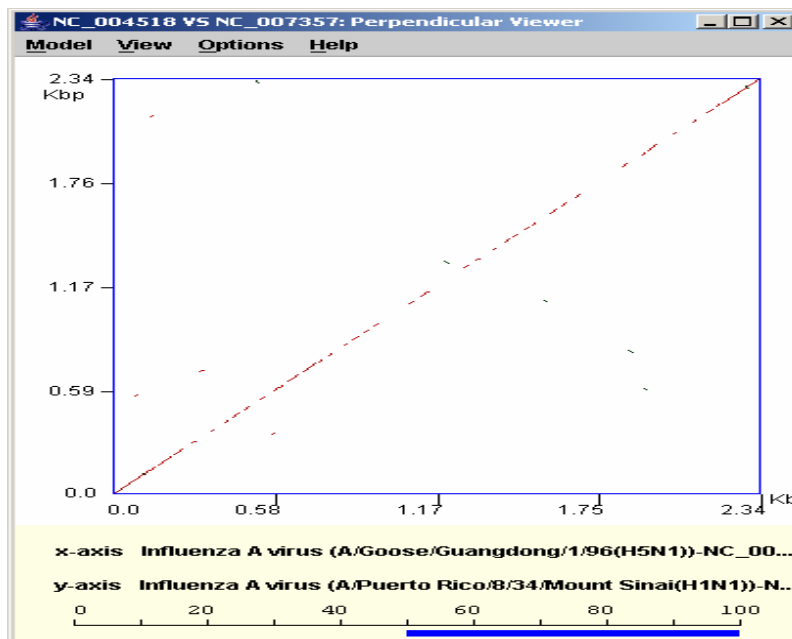


Fig. 4 Perpendicular view of comparing H5N1 and H1N1 sequences, the diagonal dashed line shows many MUMs between the two sequences

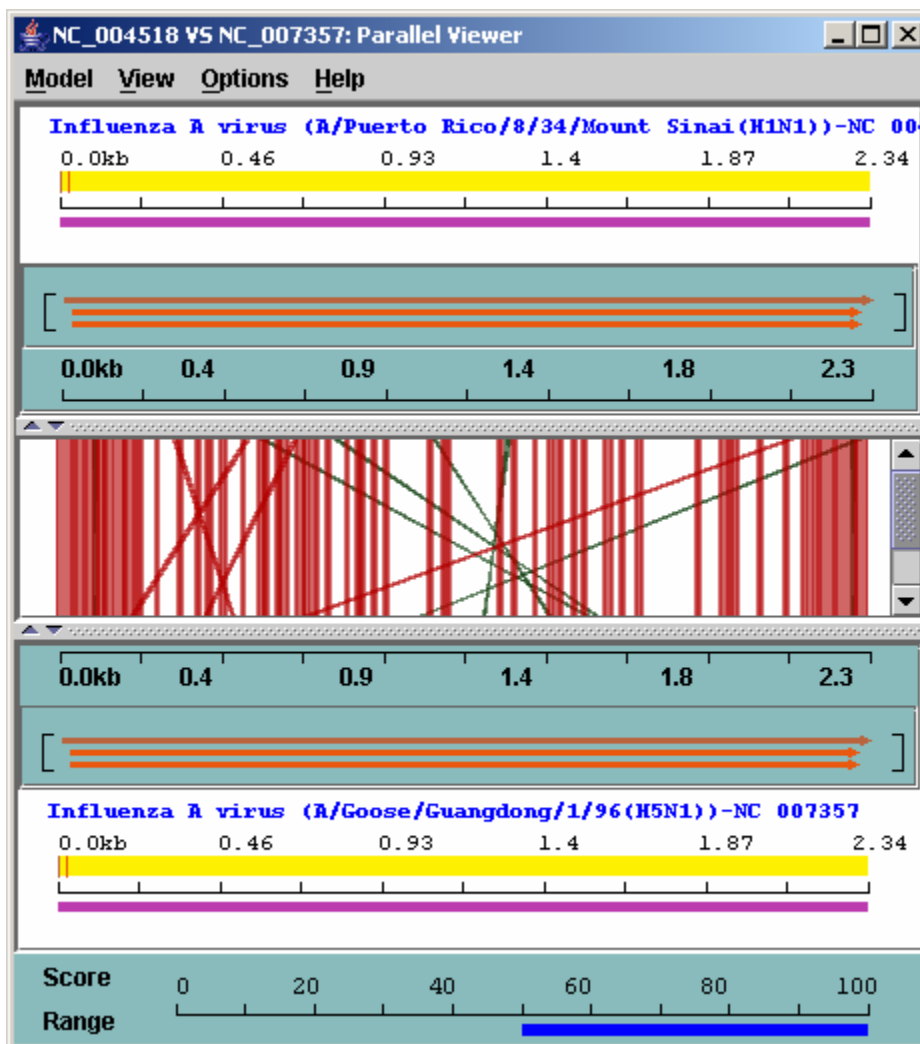


Fig. 5 Parallel view of comparing H5N1 and H1N1 sequences, the middle panel shows the MUMs, the top and bottom panel also shows all annotated features of the two sequences

Figures. 4 and 5 show high sequence similarities between the selected two replicons of the two viruses. This example illustrates that using VBI Web services through ToolBus allows applications that were not designed to work together to do so.

## 5 Conclusion & Future Directions

VBI's CIG has developed a wide variety of bioinformatics Web services. Users can browse various data resources and invoke analysis tools through the services deployed. Continued development of additional Web services is planned, with particular emphasis on microarray, proteomics, and biological pathway analysis support. Performance tuning and pipelining capability of Web services are in development. An open-source version of ToolBus for non-commercial use is available from <http://pathport.vbi.vt.edu/download>.

## 6 Acknowledgments

This work is supported by US Department of Defense, Grant number: W911SR-04-0045, awarded to Bruno W. S. Sobral. The development of our Web services has benefited considerably from the use of the following open source systems and tools: Tomcat and Axis from Apache, WSDL4J and UDDI4J from IBM, and Castor from Exolab. Various members of the CIG have contributed to the development of web services – we thank all of them for their invaluable contributions.

## 7 References

- [1] Boyu Yang, J Dana Eckart, Eric K. Nordberg and Bruno W. S. 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, June 2005, Las Vegas, NV.
- [2] <http://www.w3.org/TR/ws-arch/>
- [3] Felsenstein, J. PHYLIP (Phylogeny Inference Package). 3.6 edn (2004).
- [4] Burge, C. and Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78-94(1997).
- [5] 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).
- [6] Pertea, M. and Salzberg, S.L. *Using GlimmerM to find genes in eukaryotic genomes*. Current Protocols in Bioinformatics, 2002.
- [7] W.H. Majoros, M. Pertea, and S.L. Salzberg. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders, Bioinformatics @ Oxford University Press 2004.
- [8] Salamov A., Solovyev V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.*, 10,516-522(2000).
- [9] D. Hyatt, J. Snoddy, D. Schmoyer, G. Chen, K. Fischer, M. Parang, I. Vokler, S. Petrov, P. Locascio, V. Olman, Miriam Land, M. Shah, and E. Uberbacher, GRAIL-EXP and the Genome Analysis Toolkit, The 13th Annual Cold Spring Harbor Meeting on Genome Sequencing & Biology, May 2000.
- [10] Frishman, D., Mironov, A., Mewes, H.-W., and Gelfand, M., Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucl. Acids Res.*, 26, 2941-2947(1998).
- [11] X. Huang and W. Miller, *Adv. Appl. Math.* 12:337-357(1991).
- [12] Gribskov M, McLachlan AD, Eisenberg D. *Proc Natl Acad Sci U S A.* Jul; 84(13): 4355-8(1987).
- [13] Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25, 955-64 (1997).
- [14] Borodovsky, M. & J, M. GeneMark: Parallel Gene Recognition for both DNA Strands. *Computers & Chemistry* 17, 123-133 (1993).
- [15] Korf I. Gene finding in novel Genomes. *BMC Bioinformatics*, 5:59(2004).
- [16] Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31, 374-8 (2003).
- [17]. Yang Su, T. M. Murali, Vladimir Pavlovic, Mike Schaffer, and Simon Kasif, Rankgene: a program to rank genes from expression data, *Bioinformatics*, 19: 1578 – 1579(2003).
- [18] Henderson, J., Salzberg, S. and Fasman, K. H. Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol.*, 4, 127-141(1997).
- [19] M. Tech, N. Pfeifer, B. Morgenstern, P. Meinicke, TICO: a tool for improving predictions of prokaryotic translation initiation sites. *Bioinformatics* 21, 3568 – 3569(2005).
- [20] Mott R. EST\_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Applic.* 13:477-478(1997).
- [21] Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132, 365-86 (2000).
- [22] Nordberg, E. K. YODA: Selecting Signature Oligonucleotides. *Bioinformatics*, 21: 1365 - 1370(2004).
- [23] Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8, 186-94 (1998).
- [24] Roberts, R.J., Vincze, T., Posfai, J., Macelis, D. *Nucleic Acids Research* 33: D230-D232 (2005).
- [25] Bryson, K., McGuffin, L. J., Marsden, R. L., Ward, J. J., Sodhi, J. S. & Jones, D. T. Protein structure prediction servers at University College London. *Nucl. Acids Res.* 33(Web Server issue):W36-38(2005).
- [26] Jones D. T., Do transmembrane protein superfolds exist? *FEBS letters.* 423: 281- 285(1998).
- [27] Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL: A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, 17(12):1123-1130(2001).
- [28] Jannick Dyrlov Bendtsen, Henrik Nielsen, Gunnar von Heijne and Søren Brunak. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, 340:783-795(2004).

- [29] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* 215, 403-10 (1990).
- [30] W. R. Pearson and D. J. Lipman, Improved Tools for Biological Sequence Analysis, *PNAS* 85:2444-2448(1988).
- [31] 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).
- [32] Waterman, M.S. Efficient sequence alignment algorithms. *J Theor Biol* 108, 333-7 (1984).
- [33] E. Myers and W. Miller, "Optimal Alignments in Linear Space," *CABIOS* 4, 1 (1988), 11-17.
- [34] Derek Huntley, Angela Baldo, Saurabh Johri, and Marek Sergot, SEAN: SNP prediction and display program utilizing EST sequence clusters. *Bioinformatics Advance Access published online on December 15, 2005*
- [35] Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res* 11, 1725-9 (2001).
- [36] Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol* 5, R12 (2004).
- [37] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *The theory behind profile HMMs: Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.
- [38] Griffiths-Jones, S., Bateman, A., Marshal, M., Khanna, A., and Eddy, S. R. Rfam: an FNA family database. *Nucl. Acids Res.* 31:439-441(2003).
- [39] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucl. Acids Res.* 28: 33-6 (2000).
- [40] Kent, W.J. BLAT -- The BLAST-Like Alignment Tool. *Genome Research* 4: 656-664(2002).
- [41] J.G. Henikoff, E.A. Greene, S. Pietrokovski & S. Henikoff, "Increased coverage of protein families with the blocks database servers", *Nucl. Acids Res.* 28:228-230 (2000).
- [42] Dalgaard, P. *Introductory Statistics with R*, 288 (Springer Verlag, 2002).
- [43] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95, 14863–14868(1998).
- [44] Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5, R80 (2004).
- [45] Zdobnov, E.M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847-8 (2001).
- [46] A. S. Juncker, H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak and A. Krogh. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* 12(8):1652-62(2003).
- [47] Atalay V. and Cetin-Atalay R., Implicit Motif Distribution based Hybrid Computational Kernel for Sequence Classification, *Bioinformatics*, 21(8): 1429 – 1436(2005).
- [48] Gattiker A., Bienvenut W.V., Bairoch A., Gasteiger E.; FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification; *Proteomics* 2:1435-1444(2002).
- [49] C. R. Jimenez, L. Huang, Y. Qiu, and A.L. Burlingame, Searching Sequence Databases over the Internet: Protein Identification Using MS-Fit, *Current Protocols in Protein Science*, J. Wiley, Inc., 1998, 16.5.1-16.5.6
- [50] Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, Standing KG. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem.*; 73(9):1917-26(2001).
- [51] <http://www.biodas.org/documents/spec.html>