RECIIS
Electronic Journal of Communication Information & Innovation in Health
[www.reciis.cict.fiocruz.br]
ISSN 1981-6286

SUPPLEMENT – BIOINFORMATICS AND HEALTH

Original Articles

The PDTIS bioinformatics platform: from sequence to function

DOI: 10.3395/reciis.v1i2.Sup.98en

Thomas Dan Otto
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil
otto@fiocruz.br

Marcos Catanho
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil
mcatanho@fiocruz.br

Wim Degrave
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil
wdegrave@fiocruz.br

Antonio Basílio de Miranda
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil
antonio@fiocruz.br

Abstract

Since the eighties, Bioinformatics has acquired a growing importance in the biological sciences. Due to the complexity of hardware and software tools, and the need for a specialized management of network infrastructure, data processing and storage, many research institutions organized a bioinformatics core facility. In this work, we describe typical bioinformatics activities and the organization of a bioinformatics core with institutional support. Since the end of the 1980s, Fiocruz has contributed for the development and establishment of Bioinformatics in Rio de Janeiro, through different initiatives. Starting with sequence analysis courses, today our institution has a permanent Bioinformatics service, integrated with other strategic activities through a thematic network of technological platforms imbedded in the Program for Technological Development of Health Products (PDTIS), where the main attributions of such platforms are concentrated in support, establishment of new methodologies and active participation in research projects and technological development.

In this paper we present the PDTIS Bioinformatics Platform, its scope, resources and its support, teaching and research activities. Bioinformatics has emerged as a new scientific area and a short history of its development at Fiocruz is also summarized.

Keywords

Bioinformatics, computational biology, databases, genome, sequence analysis

Bioinformatics

In this section we present a summary of aspects of the history of Bioinformatics, emphasizing the first sequence analysis programs and databases, its interactions with automatic sequencing methods, the genome projects, comparative genomics and metagenomics, citing hallmarks of this process.

The origins of Bioinformatics and Computational Biology can be situated in the sixties, when technological advances allowed the emergence of computers as important tools in the field of Molecular Biology (as well as in all other areas). We can site three main factors responsible for this: (i) the growing number of available protein sequences, at the same time an important source of new data and questions; (ii) the idea that macromolecules carry information becomes a fundamental part of the conceptual model of Molecular Biology and (iii) the availability of more powerful computers in the main universities and research centers. Amongst the pioneers of Bioinformatics we can cite Dr. Margaret Dayhoff, who collected, organized and made available the first atlas of protein sequences in 1965 (DAYHOFF et al. 1965). Another important contributor is Dr. Temple Smith, co-author of the algorithm of dynamic programming known as Smith-Waterman (SMITH & WATERMAN 1981), which is employed in tools for similarity searches.

Until the end of the sixties, several computational techniques (algorithms and computer programs) for the analysis of structure, function and evolution of nucleotide and protein sequences were developed, as well as simple protein databases (HAGEN 2000; OUZOUNIS & VALENCIA, 2003). New techniques and approaches came about in the following decades, mainly (i) algorithms for sequence alignment; (ii) creation of databases with public access; (iii) implementation of fast methods for sequence searches and retrieval in databases; (iv) development of more sophisticated systems for protein structure prediction and (v) tools for the annotation and comparison of genomes and systems for the functional analysis of genomes (OUZOUNIS, 2002).

But it was only in the decade of the 80's that Bioinformatics and Computational Biology took the shape of independent disciplines, with their own problems and advances, and efficient algorithms, able to deal with the growing volume of information, were conceived and implementations of these algorithms became available to the scientific community (OUZOUNIS & VALENCIA, 2003). It was during this decade that several program packages for sequence analysis were created, like Staden (STADEN, 1977), Pustell (PUSTELL & KAFATOS, 1982) and GCG (DEVEREUX et al., 1984), as well as public databases serving as repositories for sequences and the results of their analysis, like (GENBANK, 2007), EMBL (EMBL, 2007), DDBJ (DDBJ, 2007) and Swiss-Prot (SWISS-PROT, 2007) (today part of UniProt (UNIPROT, 2007)). The definitive affirmation of these new disciplines came about in the nineties, with the appearance of genome, transcriptome and proteome projects (supported by important advances in DNA sequencing methods, by the development of microarrays and biochips and mass spectrometry), of computer networks at a global scale (internet), of huge biological databases, supercomputers and powerful personal computers.

In this context, besides the support to the analysis of data generated by high throughput technologies (cited above), Bioinformatics finds its usefulness in the solution of problems related mainly to the analysis of nucleotide and protein sequences, through similarity searches, annotation of genes and genomes, sequencing and genome assembly, as well as in the areas of comparative genomics, molecular phylogeny and molecular modeling.

Further important advances in DNA sequencing technology made genome projects, initiatives aiming the determination of the complete genomic sequence of a specific organism, more accessible to the scientific community, which sees them as an excellent opportunity for obtaining a general view of the metabolism, biochemistry and genetics of the organism under study (more information about genome projects can be found in the GOLD database (GOLD, 2007)). Initially aiming to study important organisms from a medical, industrial or experimental point of view (model research organisms), such as the human genome (VENTER et al., 2001; LANDER et al., 2001), mouse (Mouse Genome Sequencing Consortium et al. 2002) and several bacteria like *Escherichia coli* (BLATTNER et al., 1997) among others, today the number of genome projects has grown rapidly with the inclusion of new organisms, important for researchers with different viewpoints, as in the case of the mammal known as ornitorynch (*Ornithorhynchus anatitus)*, important for evolutionary reasons. As a matter of fact, a compilation of all genome projects from all over the world shows that there are 660 completed genomes, with more 2,258 projects under way, of which 60 of archaeabacteria, 1,344 of eubacteria and 854 of eucariotes (GOLD, 2007). In addition, use of new sequencing technologies in environmental samples gave birth to the so-called metagenomics, where all the genetic content present in such samples is studied. Contrary to usual genome projects, in which the DNA of a single organism is studied, in metagenomics DNA of a whole community of organisms is assessed. Nowadays, 114 metagenome projects are under way (GOLD, 2007).

The Bioinformatics PDTIS Platform

In this section a short history of the Bioinformatics Platform, beginning with the first sequence analysis courses at FIOCRUZ, will be presented. Following this, the Platform's main tasks and characteristics will be addressed in four sections: (i) ***computational resources*** *– equipment available to the users; (ii)* ***support activities*** *– services offered by the platform (databases and programs for sequence analysis, genomics, modeling, etc.); (iii)* ***teaching activities*** *– how the platform can help the user to obtain more information and improve his work (courses, tutorials, manuals, links to other resources, etc.) and (iv)* ***research activities*** *– results of research performed by the bioinformatics team.*

In Rio de Janeiro, the first Bioinformatics courses were organized by Wim Degrave at the Oswaldo Cruz Institute in 1989, as disciplines in the pos-graduate courses, from the Biochemistry and Molecular Biology

Department (DBBM), with emphasis in the computational analysis of nucleotide and protein sequences using a VAX MX850, based in the old model PDP 11. Other courses followed, like the course *Automatic DNA Sequencing* (1992), *International Training Course on Computer Analysis of Nucleic Acid and Protein Sequences and the Use of Networks* (1992, with support from OMS/TDR). Later, other national and international courses were organized, supported by Fiocruz, OMS/TDR e FAO. In 1994, the first high capacity server was acquired: *gene.dbbm.fiocruz.br*, a Silicon Graphics Challenge server with 4 CPUs, and a graphical station *SGI Indigo2*, substituted in 1999 by the server SGI Origin2000, called *amoeba*. In 2003, the Bioinformatics PDTIS Platform was officially created as part of the Technological Development Program for Health Products (Plataforma de PDTIS Bioinformática 2007).

PDTIS has, as a main goal, to foster the technological development at Fiocruz, through the articulation of multidisciplinary and cooperative networks, aiming at the development of products, processes and services with impact in public health and the economic and social development of Brazil. At the same time, it promotes and stimulates cultural changes in the institution, through the implementation of bridges between applied research, the production of inputs for health, and institutional technology management. The adopted model for this structuring, Cooperative Networks, aims at motivating researchers to work in an organized and collaborative fashion around common goals and similar technologies, thus obtaining significant results with the optimization of human and financial resources. The program is managed through a Management Committee composed by the program coordinators, the cooperative networks coordinators and by the quality, technology and financial supervisors.

Equipped with modern instruments and operated by specialized staff, the technological platforms offer not only support and service, but actively participate in research projects, technological development projects and platform networking. They are characterized by their high strategic value, which makes them indispensable for public and private sectors. Their main activities concentrate on support, development of methodology and active participation in research projects.

The PDTIS Bioinformatics Platform (Figure 1) makes available the main bioinformatics tools and sequence databases for the Fiocruz community. This has been achieved in two ways: (i) via internet, offering free access to the majority of programs and databases, and (ii) through the bioinfo server, where all programs and databases are installed, and which is dedicated mainly to the execution of computationally intensive tasks. The staff offers advice to the casual user interested in the analysis of one or a few sequences, using web interfaces, and giving access to several utilities over the Internet. For users with a larger data volume or with special needs, the opening of an account for access to the services on a server is suggested. Rules and procedures for requesting accounts and platform utilization can be found on the home page (PDTIS, 2007). The installation of new databases and programs can be done upon demand, in these cases users should schedule he service directly with the platform administrators.

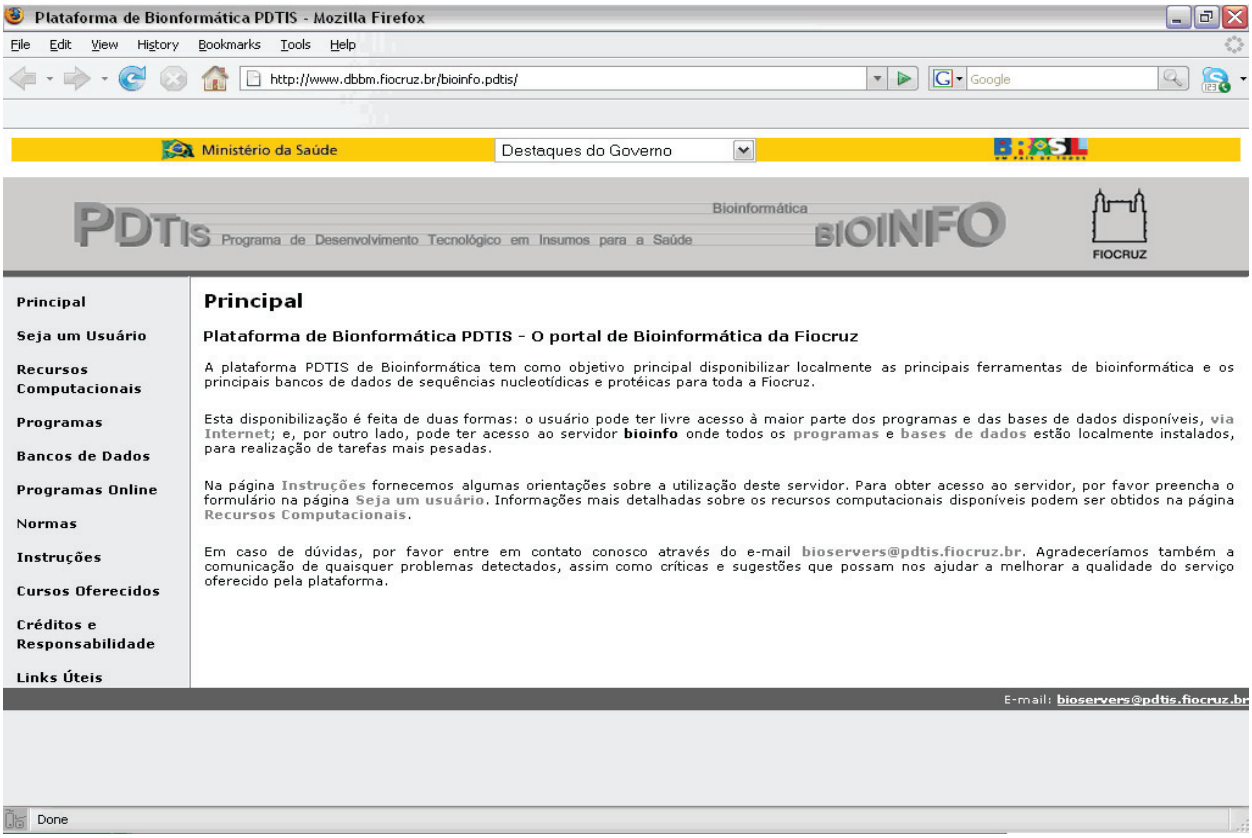

Figure 1 - Home page of the PDTIS Bioinformatics Platform.

Today, the Bioinformatics Platform counts with six servers for use by the FIOCRUZ community and several personal computers for students and visiting researchers (Table 1). There is the *bioinfo* server, hosting databases and computational tools, open for all interested users; the *genoma* server, dedicated to genome assembly; the *tremendao* server, dedicated to the storage and processing of sequences originating from the DNA Sequencing PDTIS Platform, as well as some databases; the *magneto* server, dedicated to the execution of web services and online tools like PISE (LETONDAL, 2001); and the *neptune* and *pluto* servers, reserved for data processing and web services.

Table 1 - Computational resources available at the PDTIS Bioinformatics platform

Computer	Name	TCP/IP	Services	Model
Bioinfo	bioinfo.pdtis.fiocruz.br	157.86.152.23	Databases and programs, open for all users	A
Genoma	genoma.pdtis.fiocruz.br	157.86.152.20	Genome assembly	A
Tremendao	-	157.86.44.10	Databases	D
Magneto	-	157.86.44.10:808	Web server, online programs	B
SeqAnalist	-	157.86.44.10	Sequence analysis	C
Proteoma	proteoma.pdtis.fiocruz.br	157.86.44.32	Proteome analysis programs	B
Neptune	-	157.86.176.101	Data processing	B
Pluto	-	157.86.176.102	Data processing	B

Server configurations

A: SunFire v65-X with 2 Intel Xeon processors of 3.2 gigahertz (GHz) and 5 gigabytes (GB) RAM with 1 36 GB HD and 3 72 GB HD's.

B: Intel Pentium-4.3 GHz, 1 GB RAM, 80 GB HD.

C: Intel Pentium-4.3 GHz, 256 MB RAM, 30 GB HD.

D: Athlon 3 GHz with 64 bits processor, 1 GB RAM, 200 GB HD.

The main nucleotide and protein sequence databases, like GenBank (BENSON et al., 2007), InterPro (MULDER et al., 2007), Swiss-Prot (BOECKMANN et al., 2003), Blocks (HENIKOFF et al., 1999; HENIKOFF et al., 2000), Prosite (HULO et al., 2006), TrEMBL (BOECKMANN et al., 2003), are available for querying by the Fiocruz community (Table 2). Similarly, the main programs for sequence analysis and other tasks like genome assembly are also available, *e.g.* EMBOSS (*European Molecular Biology Open Software Suíte*), an open source program package developed to meet the need of bioinformatics and molecular biologists (RICE et al., 2000); BLAST (*Basic Local Alignment and Search Tool*), an algorithm for the local alignment of nucleic or protein sequences (ALTSCHUL et al., 1997), HMMER (*Hidden Markov Modeler*), an open source implementation of an algorithm for the construction of PHMMs, *Profile Hidden Markov Models*, that can be used in similarity searches (EDDY, 1998); ClustalW, a multiple sequence alignment programs (THOMPSON et al., 1994; CHENNA et al., 2003), amongst others (Table 3).

Table 2 - Installed databases available for querying, at the Bioinformatics platform

Name	Description	Size*
AAindex	Aminoacid indexes and similarity matrices.	1.1 MB
BIOCATAL	Software catalogue for molecular biology.	2.1 MB
BLAST	Databases for BLAST execution.	32 GB
Blocks	Database of multiple alignments, no gaps, corresponding to the most conserved positions in a group of related proteins.	1.1 GB

cont.

Table 2 - Installed databases available for querying, at the Bioinformatics platform (cont.)

ChEBI	Chemical Entities of Biological Interest (ChEBI): dictionary of small molecular entities like atoms, molecules, ions, radicals, etc., identifiable as different entities. May be natural or synthetic products.	201 MB
CPGISLE	Database of CpG islands.	21 MB
CUTG	Organism-specific codon usage obtained from sequences available at GenBank.	689 MB
GO	Controlled vocabulary of terms for the description of attributes of gene products.	2.4 GB
InterPro	Database of protein families, domains and functional sites where identifiable characteristics found in known proteins may be applied to unknown protein sequences.	4.2 GB
KEGG	Database of enzymes and metabolic pathways.	29 GB
MassSpecDB	Database of non-identical protein sequences, compiled from several primary resources.	2.5 GB
PDB	Repository for data processing and distribution about the tri-dimensional structure of proteins and nucleic acids.	148 MB
PeptideSearch	Mass spectrometry data.	53 MB
Pfam	Collection of multiple alignments and Hidden Markov Models covering several protein families and protein domains.	22 GB
Prints	Database of protein families and domains.	59 MB
PROSITE	Database of protein families and domains.	192 MB
RefSeq	Reference database from NCBI.	12 GB
RESID	Database of protein modifications. It is a collection of annotations and structures for post-translational protein modifications.	3.5 MB
Swiss-Prot	Protein database.	6.8 GB
TRANSFAC	Database of eukaryotic transcription factors and their DNA binding sites.	9.5 MB
TrEMBL	Supplement of Swiss-Prot, automatic annotation, containing the conceptual translations of all nucleotide sequences from EMBL not integrated into Swiss-Prot.	1.8 GB
UniGene	Experimental system for the automatic partition of GenBank in a non-redundant set of gene oriented clusters.	8.6 GB
UTR	Database of untranslated regions of eukaryotic mRNA.	872 MB

* Space used by each database in 08/10/2007. Size of databases varies according to their versions, increasing at every release. KB: kilobytes; MB: megabytes; GB: gigabytes.

Table 3 - Available programs at the Bioinformatics PDTIS platform

Name	Description	Basic commands
Similarity Searches		
BLAST	BLAST executes local similarity searches of one or more sequences in nucleic acid or protein databases.	blastall
WU-BLAST	WU-BLAST executes local similarity searches of one or more sequences in nucleic acid or protein databases.	blastn blastp blastx tblastn tblastx
FASTA	FASTA executes searches based on Pearson e Lipman's algorithm, aiming local similarity detection between a given sequence and a group of sequences from the same type (nucleic acids or proteins).	fasta34

cont.

Table 3 - Available programs at the Bioinformatics PDTIS platform (cont.)

SSEARCH	SSEARCH executes a rigorous search, based of the algorithm of Smith-Waterman, aiming local similarity detection between a given sequence and a group of sequences from the same type (nucleic acids or proteins).	ssearch34
HMMER	Program for similarity searches using statistical descriptions of the consensus obtained for a family of sequences (Profile Hidden Markov Models).	hmmalign hmmcalibrate hmmsearch hmmbuild
Multiple Alignments		
ClustalW	Program for multiple alignment (global) of nucleic or proteic sequences.	clustalw
T-Coffee	Program for multiple alignment (global) of nucleic or proteic sequences.	t_coffee
MUMmer	System for multiple alignment and visualization of whole genomes, or big DNA or protein segments, in a fast and efficient way.	mummer nucmer promer run-mummer1 run-mummer3
SeaView	Graphical editor of multiple sequence alignments.	seaview
Sequence Assembly		
AMOS	Open source program package offering support for the development of tools for genome assembly	runAmos
PCAP	Program for nucleic sequences assembly based on the detection of overlaps between the fragments.	pcap
CAP3	Program for nucleic sequences assembly based on the detection of overlaps between the fragments.	cap3
Phred/Phrap/ Consed	Package for chromatogram reading, base-calling, quality attribution for each base (Phred), sequence assembly (Phrap), visualization, edition and closing of the assemblies (Consed).	phredPhrap phredPhrap.poly consed
TIGR Assembler	Tool for the assembly of large sequence datasets like ESTs, BACs or small genomes.	TIGR_Assembler
Sequence Annotation		
Glimmer	System developed for gene prediction in microbial genomes, particularly bacteria, archaeabacteria and viruses.	glimmer3
Artemis	Free tool for visualization and annotation of genomes.	art
ACT	Artemis Comparison Tool. Tool for interactive visualization of comparisons between genomic sequences and their annotations.	act
Other programs		
EMBOSS	Free package for sequence analysis	wossname
REPuter	Famíly of programs used for detection and visualization of repetitive sequences in whole genomes or chromosomes.	repfind repvis repselect
PerlPrimer	Open source graphical interface, written in Perl, developed for helping primer design for standard PCR, real-time PCR and sequencing.	perlprimer.pl
CodonW	Program developed to simplify multivariate analysis (correspondence analysis) of codons and aminoacid usage.	codonw

Besides all these databases and programs, the Bioinformatics Platform has generated other data repositories and computational tools, free for academic use. We can cite BioParser (CATANHO et al., 2006a), a program that analyses the results of sequence similarity searches using BLAST or FASTA (PEARSON & LIPMAN, 1988; Pearson 1990), generating a local database that can be queried through a graphical web interface; GenoMycDB (CATANHO et al., 2006b), a database for comparative analysis of genes and genomes from mycobacteria; LocalCOG (LocalCOG, 2005), relational implementation of COG database (TATUSOV et al., 1997, 2003), with local installation, with an interface developed for complex queries; Squid (CARVALHO et al., 2005), a program for sharing computational resources, allowing the execution of BLAST in a distributed way using personal computers, maximizing utilization of available computational power; among others in different stages of development (PDTIS, 2007; THE BIOINFORMATICS TEAM, 2007).

The Bioinformatics Platform has also contributed in student formation, offering courses and opportunities to develop Master and PhD thesis. Regularly, we offer the courses *Nucleotide and Protein Sequence Analysis*, *Introduction to PERL Programming*, *Origin, Structure and Evolution of Genomes*. We also offer basic and introductory bioinformatics winter and summer courses (CURSOS DE FÉRIAS DO IOC, 2007), extension courses promoted by the Posgraduation Program in Cellular and Molecular Biology (PGBCM) of the Oswaldo Cruz Institute/Fiocruz.

In addition, present research lines of the bioinformatics team include (i) functional classification and annotation of proteins, metabolic pathways, analogous enzymes, pseudogenes and repetitive sequences; (ii) modeling of biological data; (iii) developing of new algorithms for genome assembly and (iv) annotation of trypanosomatid and mycobacterial genes. Recently, the calculation stage of the *Genome Comparison Project* (GENOME COMPARISON PROJECT, 2007), in which all protein coding genes from all completed genomes so far were compared in a pairwise manner, using the Smith-Waterman algorithm on the World Community Grid (WORLD COMMUNITY GRID, 2007), was finished. Results of these comparisons allow for the development of several studies in phylogeny and molecular evolution, and will soon be available for the whole scientific community.

Perspectives

In this section we present the perspectives for the Bioinformatics Platform.

As described above, a Bioinformatics Platform must act supporting research projects, technological development and activities of other platforms in the network. We have identified several points that are critical to improve the quality of the services.

Continuous interaction between the platform and the users is essential. Beginners in the field request adequate information, on an appropriate level, to help them using the resources of the platform. Although the platform has a collection of tutorials and links to other sites as well as reference lists, in their majority these documents do not attend the needs of the occasional user or beginner, who needs a step-by-step guide. Therefore, gradually, simplified protocols describing the utilization of the main programs and databases are being provided. These will cover simple sequence assembly, SNP identification, execution of similarity searches, etc.

It is also necessary to strengthen the interaction with the other platforms, implementing, for example, a pipeline for sequence clustering aiming SNP identification from the sequences produced by the DNA Sequencing Platform. The availability of a server with a high quality graphics card is necessary to meet the demands of the users interested in molecular modeling.

Also, a cluster of machines with high performance is desirable for projects with heavy workload. In fact, we are planning the implementation of a grid at Fiocruz, which can use the idle capacity of the hundreds of personal computers spread throughout the institution (similar to the *World Community Grid*, WORLD COMMUNITY GRID, 2007]), without affecting the daily work of those participating.

The Bioinformatics Platform must maintain quality services, offering to the Fiocruz community the main programs and databases used in Molecular Biology and Bioinformatics, improving its relationship with the users and supporting the development of research in the Institute.

Bibliographic references

ALTSCHUL, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res**, v.25, n.17, p.3389-402, 1997.

BENSON, D.A. et al. GenBank. **Nucleic Acids Research**; v.35:D21-D25, 2007.

BOECKMANN B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. **Nucleic Acids Research**, v.31, n.1, p.365-70, 2003.

CATANHO, M.et al. BioParser: a tool for processing of sequence similarity analysis reports. **Applied Bioinformatics**, v.5, n.1, p.49-53, 2006a.

CATANHO, M. et al. GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. **Genetic Molecular Research**, v.5, n.1, p.115-26, 2006b.

CHENNA, R. et al.. Multiple sequence alignment with the Clustal series of programs. **Nucleic Acids Research**, v.31, n.13, p.3497-3500, 2003.

CURSOS DE FÉRIAS DO IOC. Available at: <http://bioinfo.pdtis.fiocruz.br/cursosdeferias_ioc/>. Accessed: 14 Nov. 2007.

DAYHOFF, M.O. (Eds). **Atlas of Protein Sequence and Structure**. Maryland, USA: National Biomedical Research Foundation, Silver Spring, 1965.

DDBJ. Available at: <http://www.ddbj.nig.ac.jp/>. Accessed: 8 Oct. 2007.

EMBL. Available at: <http://www.ebi.ac.uk/embl/>. Accessed: 8 Oct. 2007.

GenBank. Available at: <http://www.ncbi.nlm.nih.gov/Genbank/>. Accessed: 8 Oct. 2007.

GENOME COMPARISON PROJECT. Available at: <http://www.worldcommunitygrid.org/projects_showcase/archives/fgc/viewFgcOverview.do>. Accessed: 8 Oct. 2007.

GOLD. Genomes Online Database. Available at: <http://www.genomesonline.org/>. Accessed: 8 Oct. 2007.

HAGEN, J.B. The origins of bioinformatics. **Nature Reviews Genetics**, v.1, n.3, p.231-6, 2000.

LANDER E.S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n.6822, p.860-921, 2001.

LOCALCOG. Relational database system for local use of COG data, 2005. Available at: <http://www.dbbm.fiocruz.br/labwim/bioinfoteam/index.pl?action=softwares>. Accessed: 2005.

OUZOUNIS, C. Bioinformatics and the theoretical foundations of molecular biology. **Bioinformatics**, v.18, n.3, p.377-8, 2002.

OUZOUNIS, C.A.; VALENCIA, A. Early bioinformatics: the birth of a discipline--a personal view. **Bioinformatics**, v.19, n.17, p.2176-90, 2003.

PDTIS. Plataforma de Desenvolvimento Tecnológico em Insumos para a Saúde. Available at: <http://www.presidencia.fiocruz.br/vppdt1/pdtis.php>. Accessed: 8 Oct. 2007.

PEARSON, W.R.; LIPMAN, D.J. Improved tools for biological sequence comparison. **Proceedings of National Academy of Science**, v.85, n.8, p.2444-8, 1988.

PEARSON, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. **Methods Enzymol**, 183, p.63-98, 1990.

Plataforma PDTIS de Bioinformática. Available at: <http://www.dbbm.fiocruz.br/bioinfo.pdtis/>. Accessed: 8 Oct. 2007.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends in Genetics**, v.16, n.6, p.276-277, 2000.

SMITH, T.F.; WATERMAN, M.S. Comparison of Biosequences. **Advances in Applied Mathematics**, 2: p.482-9, 1981.

SWISS-PROT. Available at: <http://www.expasy.ch/sprot/>. Accessed: 8 Oct. 2007.

TATUSOV, R.L.; KOONIN, E.V., LIPMAN, D.J. A genomic perspective on protein families. **Science**, v.278, n.5338, p.631-7, 1997.

TATUSOV, R.L. et al. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics** v.4, n.1, p.41, 2003.

THE BIOINFORMATICS TEAM. Available at: <http://www.dbbm.fiocruz.br/labwim/bioinfoteam/>. Accessed: 8 Oct. 2007.

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 Research**, v.22, n.22, p.4673-80, 1994.

UNIPROT. Available at: <http://www.expasy.uniprot.org/>. Accessed: 8 Oct. 2007.

VENTER, J.C. et al. The sequence of the human genome. **Science**, v.291, n.5507, p.1304-51, 2001.

WORLD COMMUNITY GRID. Available at: <http://www.worldcommunitygrid.org/>. Accessed: 8 Oct. 2007.

About the authors

Thomas D. Otto

Master in Computer Sciences at the University at Lubeck, Germany. He is currently concluding his Ph.D degree in Cellular and Molecular Biology at the Oswaldo Cruz Institute (IOC/Fiocruz), in the Field of Bioinformatics. He has experience in the fields of Informatics, Bioinformatics and Molecular Biology, and research in genome assembly, comparative genome analysis and evolution, developing algorithms and applications for comparative and functional genomics of prokaryotes.

Marcos Catanho

Holds a Master degree in Cellular and Molecular Biology (focusing on Bioinformatics and Comparative Genome Analysis) from the Instituto Oswaldo Cruz (IOC/Fiocruz) and a Baccalaureate degree in Pharmacy from Universidade Federal do Rio de Janeiro (UFRJ). Currently he is a PhD student in Cellular and Molecular Biology (focusing on Bioinformatics and Comparative Genome Analysis) at Instituto Oswaldo Cruz (IOC/Fiocruz). He is active in the area of Biological Sciences with emphasis on: comparative genome analysis and evolution; development of computational approaches and tools for comparative analysis of microbial genomes.