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

SUPPLEMENT – BIOINFORMATICS AND HEALTH

Review Articles

Comparing genomes: databases and computational tools for comparative analysis of prokaryotic genomes

DOI: 10.3395/reciis.v1i2.Sup.105en

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

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

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

Abstract
Since the 1990's, the complete genetic code of more than 600 living organisms has been deciphered, such as bacteria, yeasts, protozoan parasites, invertebrates and vertebrates, including *Homo sapiens*, and plants. More than 2,000 other genome projects representing medical, commercial, environmental and industrial interests, or comprising model organisms, important for the development of the scientific research, are currently in progress. The achievement of complete genome sequences of numerous species combined with the tremendous progress in computation that occurred in the last few decades allowed the use of new holistic approaches in the study of genome structure, organization and evolution, as well as in the field of gene prediction and functional classification. Numerous public or proprietary databases and computational tools have been created attempting to optimize the access to this information through the web. In this review, we present the main resources available through the web for comparative analysis of prokaryotic genomes. We concentrated on the group of mycobacteria that contains important human and animal pathogens. The birth of Bioinformatics and Computational Biology and the contributions of these disciplines to the scientific development of this field are also discussed.

Keywords
Bioinformatics, computational biology, databases, genomes, prokaryotes

The beginning of a new era: the birth of Bioinformatics and Computational Biology

Bioinformatics and Computational Biology (BISTIC Definition Committee 2000) emerged in the 1960's when computers became essential tools for the development of Molecular Biology. According to HAGEN (2000), this emergence was motivated by three main factors: (i) the increasing availability of protein sequences, providing both a source of data and a set of relevant challenges impossible to cope without computer assistance; (ii) the idea that macromolecules carry information had became fundamental in the Molecular Biology conceptual framework; (iii) the availability of powerful computers in research centres.

Several algorithms and computational programs for the analysis of structure, function, and evolution at the molecular level, as well as rudimentary protein sequences databases, were already available towards the end of the 1960's (HAGEN, 2000; reviewed by OUZOUNIS & VALENCIA, 2003). New algorithms and computational approaches were introduced in the following decades, such as algorithms for sequence alignments, public databases, efficient data retrieval systems, sophisticated protein structure prediction methods, gene annotation and genome comparison tools, and systems for functional genome analysis (OUZOUNIS, 2002).

However, Bioinformatics and Computational Biology might only be recognized as independent disciplines, with their own problems and achievements, by the decade of 1980 when, for the first time, efficient algorithms were developed to cope with the increasing amount of information, and computer implementations of these algorithms (programs) were made available for the entire scientific community (OUZOUNIS & VALENCIA, 2003). The consolidation of both new disciplines occurred in the 1990's, with the emergence of powerful personal computers, supercomputers, the World Wide Web, huge biological databases and the so-called *ome* projects: genome, transcriptome, and proteome, supported by the continuous progress in DNA sequencing, the development of microarrays and biochip technologies, and mass spectrometry.

Actually, the achievement of (i) numerous complete genome sequences, (ii) gene and protein expression data of cells, tissues and organs, combined with the (iii) development of high-throughput computing technologies and (iv) more efficient algorithms, allowed holistic approaches (which consider the whole body of available information, such as all genes encoded by a group of genomes) to be used in the study of genome structure, organization and evolution, in differential expression analyses of genes and proteins, in protein three-dimensional structure predictions, in the process of metabolic reconstruction, and in the functional prediction of genes. As a result, at least two general rules about biological systems (summarizing a number of experimental evidences) can be derived from the exercise of Bioinformatics and Computational Biology over these almost five decades, given the existence of numerous corollaries resulting from them with direct application in biological researches: (i) the three-dimensional structures of proteins are much more conserved than their biochemical functions; (ii) in contrast to genomic sequences, the comparison of the total number of genes encoded by each individual in a group of organisms do not reflects the phylogeny of the species involved (OUZOUNIS, 2002).

New challenges, new approaches: the comparative analysis of prokaryotic genomes

The pioneering initiative of the U.S. Department of Energy (DOE) to obtain a reference human genome sequence culminated in the launching of the Human Genome Project, in 1990. The initial plan was achieve a deeper understanding of potential health and environmental risks caused by the production and use of new energy resources and technologies. Later, the technological resources generated by this project stimulated the development of many other public and private genome project initiatives (HGP 2001).

So far, 70 eukaryotic genomes have already been completely sequenced. They include the human genome (VENTER et al., 2001; LANDER et al., 2001), some other vertebrates and plants. In addition, the complete genome sequence of 47 archaeobacteria and 543 eubacteria are also available, and 2,258 other projects are currently in progress (GOLD, 2007). Concerning the mycobacteria group (GOODFELLOW & MINNIKIN, 1984) in particular, the genomes of 16 species have already been entirely sequenced and 23 others are on going (Table 1).

Table 1 - Mycobacterial genome projects

Species or strain	Importance	Research Centre	URL	Status
M. tuberculosis H37Ra	Medical; human and animal pathogen; causes tuberculosis	Beijing Genomics Institute	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=21081	Complete
M. tuberculosis F11 (ExPEC)	Medical; animal, cattle, and human pathogen; causes tuberculosis.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Complete

cont.

Table 1 - Mycobacterial genome projects (cont.)

M. bovis BCG Pasteur 1173P2	Medical; animal, cattle, and human pathogen; causes tuberculosis.	Institut Pasteur	http://www.pasteur.fr/recherche/unites/Lgmb/mycogenomics.html	Complete
M. ulcerans Agy99	Medical; human pathogen; causes Buruli ulcer.	Institut Pasteur	http://www.pasteur.fr/recherche/unites/Lgmb/mycogenomics.html	Complete
M. flavenscens PYR-GCK	Biotechnological; isolated from soil.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/mycfl/mycfl.home.html	Complete
M. vanbaalenii PYR-1	Biotechnological; isolated from soil.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/mycva/mycva.home.html	Complete
Mycobacterium sp JLS	Biotechnological; isolated from creosote-contaminated soil.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/myc_j/myc_j.home.html	Complete
Mycobacterium sp KMS	Biotechnological; isolated from creosote-contaminated soil.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/myc_k/myc_k.home.html	Complete
Mycobacterium sp MCS	Biotechnological; isolated from creosote-contaminated soil.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/myc_k/myc_k.home.html	Complete
M. tuberculosis H37Rv	Medical; human and animal pathogen; causes tuberculosis.	Sanger Institute	http://www.sanger.ac.uk/Projects/M_tuberculosis/	Complete
M. bovis AF2122/97	Medical; animal, cattle, and human pathogen; causes tuberculosis	Sanger Institute/ Institut Pasteur	http://www.sanger.ac.uk/Projects/M_bovis/	Complete
M. leprae TN	Medical; human pathogen; causes leprosy.	Sanger Institute/ Institut Pasteur	http://www.sanger.ac.uk/Projects/M_leprae/	Complete
M. avium 104	Medical; animal pathogen; causes respiratory infection.	The Institute for Genomic Research	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=20086	Complete
M. smegmatis MC2 155	Medical; human pathogen; opportunistic infection.	The Institute for Genomic Research	http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gms	Complete
M. tuberculosis CDC1551	Medical; animal and human pathogen; causes tuberculosis.	The Institute for Genomic Research	http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gmt	Complete
M. avium paratuberculosis k10	Medical; animal and cattle pathogen; causes Johne's disease, paratuberculosis and enteritis.	University of Minnesota	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=380	Complete
M. tuberculosis A1	Medical; human pathogen; causes tuberculosis.	Broad Institute	-	Incomplete
M. tuberculosis C	Medical; human pathogen; causes tuberculosis.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incomplete
M. tuberculosis Ekat-4	Medical; human pathogen; causes tuberculosis.	Broad Institute	-	Incomplete
M. tuberculosis Haarlem	Medical; human pathogen; causes tuberculosis.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incomplete
M. tuberculosis KZN 1435 (MDR)	Medical; human pathogen; causes tuberculosis.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incomplete
M. tuberculosis KZN 4207 (DS)	Medical; human pathogen; causes tuberculosis.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incomplete

cont.

Table 1 - Mycobacterial genome projects (cont.)

M. tuberculosis KZN 605 (XDR)	Medical; human pathogen; causes tuberculosis.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incomplete
M. tuberculosis Peruvian1	Medical; human pathogen; causes tuberculosis.	Broad Institute	-	Incomplete
M. tuberculosis Peruvian2	Medical; human pathogen; causes tuberculosis.	Broad Institute	-	Incomplete
M. tuberculosis W-148	Medical; human pathogen; causes tuberculosis.	Broad Institute	-	Incomplete
M. ulcerans	Medical; human pathogen; causes Buruli ulcer.	Clamson University	http://www.genome.clemson.edu/projects/stc/m.ulcerans/MU__Ba/index.html	Incomplete
M. bovis BCG Moreau[a]	Medical; animal, cattle, and human pathogen; causes tuberculosis.	Fundação Oswaldo Cruz / Fundação Ataulpho de Paiva	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=18279	Incomplete
M. abscessus CIP 104536	Medical; human pathogen; causes broncho-pulmonary and respiratory infection.	Genoscope	http://www.genoscope.cns.fr/externe/English/Projets/Projet_LU/organisme_LU.html	Incomplete
M. chelonae CIP 104535	Medical; human pathogen; causes broncho-pulmonary and respiratory infection.	Genoscope	http://www.genoscope.cns.fr/externe/English/Projets/Projet_LU/organisme_LU.html	Incomplete
Mycobacterium sp. Spyr1	Biotechnological; isolated from creosote-contaminated soil.	Joint Genome Institute / University of Ioannina	-	Incomplete
M. liflandii 128FXT	Medical; frog and animal pathogen; causes systemic disease.	Monash University	-	Incomplete
M. marinum DL240490	Medical; fish and human pathogen; causes tuberculosis-like infection and skin infection.	Monash University	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=20229	Incomplete
M. ulcerans 1615	Medical; human pathogen; causes Buruli ulcer.	Monash University	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=20231	Incomplete
M. africanum GM041182	Medical; human, cattle and animal pathogen; causes tuberculosis.	Sanger Institute	http://www.sanger.ac.uk/sequencing/Mycobacterium/africanum/	Incomplete
M. canetti CIPT140010059	Medical; human, cattle and animal pathogen; causes tuberculosis.	Sanger Institute	http://www.sanger.ac.uk/sequencing/Mycobacterium/canetti/	Incomplete
M. microti OV254	Medical; animal, cattle and human pathogen; causes tuberculosis.	Sanger Institute / Institut Pasteur	http://www.sanger.ac.uk/Projects/M_microti/	Incomplete
M. marinum M	Medical; animal and human pathogen; causes tuberculosis.	Sanger Institute / University of Washington / Institut Pasteur / Monash University / University of Tennessee	http://www.sanger.ac.uk/Projects/M_marinum/	Incomplete
M. tuberculosis 210	Medical; human pathogen; causes tuberculosis.	The Institute for Genomic Research	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=273	Incomplete

Sources: Genomes Online Database (GOLD 2007), NCBI Entrez Genome Project Database (Genome Project 2007), and Comprehensive Microbial Resource (CMR 2007).

Complete genome sequences constitute a singular source of data since they comprise in principle all that is required to create an organism along with epigenetic factors and their interaction with these factors (STROHMAN, 1997). However, what could be actually done with all this information is not immediately clear. For instance, it is believed that the comprehensive analysis of entire genomes has the potential to provide a complete understanding of genetics, biochemistry, physiology and pathogenesis of microorganisms (BROSCH et al., 2001). Nevertheless, it is argued that such potential can only be fulfilled by comparative studies of genomic sequences or syntenic regions in a pair or group of related species, subspecies or strains, as the genome of a single organism considered alone, without its phylogenetic context of the evolutionary process, merely provides an incomplete understanding of those issues (WEI et al., 2002).

Concerning this subject, Fraser et al. (2000) clearly showed how an evolutionary perspective could benefit genomic analyses. They gave examples of how it could assist (i) the identification of biological function of new genes, (ii) the inference of recombination patterns within species, (iii) the detection of lateral gene transfers between species and (iv) loss of genetic material, as well as (v) distinguishing similarities due to homology from those due to convergence (FITCH, 1970; 2000). On the other hand, KONDRASHOV (1999) and KOONIN et al. (2000) drew our attention to the significance of comparative genome analyses to Evolutionary Biology. According to KONDRASHOV (1999), comparative genomics has supplied the best available evidences for some evolutionary phenomena and, in some circumstances, has eventually led to the refinement of older concepts. More recently, new phylogenetic analysis strategies based on the entire gene content of completely sequenced genomes have been developed, and new methods for calculating inter-genomic distance have also been considered (OTU & SAYOOD, 2003; HENZ et al., 2005; KUNIN et al., 2005a and references therein; KUNIN et al., 2005b; TEKAIA et al., 2005). These methods overcome some recognizable problems of traditional phylogenetic approaches, such as saturation at certain codon positions, selection of suitable evolutionary markers, and biases yielded by these factors in phylogenetic analyzes. Hence, there is a feedback process between evolutionary and genomic analyses, as stated by FRASER et al. (2000).

It is important to stress that since the sequencing of the first bacterial genome in 1995, comparative analyses of prokaryotic genomes have gradually uncovered the complex nature of their genome structures and organization, and the enormous genetic diversity among these microorganisms (considerably higher than one could expect, even among isolates of a single species). This brings up important questions regarding the mechanisms by which prokaryotes are evolving and how taxonomists should actually classify them (COENYE et al., 2005; BINNEWIES et al., 2006).

Regarding pathogenic microorganisms in general and mycobacteria in particular, a number of potential applications of comparative genome analysis have been reported, aimed especially at the prevention (development of more effective vaccines), treatment (development of new drugs), and diagnosis (development of faster and more accurate methods) of tuberculosis and other mycobacterial diseases. Some of these applications include: (i) identification of unique genes and virulence factors, and metabolism reconstruction (GORDON et al., 2002); (ii) characterization of pathogens and identification of new diagnostic and therapeutic targets (FITZGERALD & MUSSER, 2001); (iii) investigation of the molecular basis of pathogenesis and host range, and differences in phenotypes between clinical isolates and natural populations of pathogens (BEHR et al., 1999; BROSCH et al., 2001; COLE, 2002; KATO-MAEDA et al., 2001); and (iv) investigation of the genetic basis of virulence and drug resistance in tuberculosis-causing bacteria (RANDHAWA & BISHAI, 2002).

Comparative genome analysis is a relatively recent approach that emerged with the sequencing of the first genomes in the 1990's. However, its constitutive tools derive from classical sequence analysis techniques: (i) global and local pairwise or multiple sequence alignment algorithms, (ii) phylogenetic analysis methods, and (iii) computer implementations of such algorithms and methods (NEEDLEMAN & WUNSCH, 1970; SMITH & WATERMAN, 1981; LIPMAN & PEARSON, 1985; PEARSON & LIPMAN, 1988; FENG & DOOLITTLE, 1987; ALTSCHUL et al., 1990; 1997; THOMPSON et al., 1994; FELSENSTEIN, 1981; 1989). Actually, comparative genome analysis not only benefits from preceding sequence analysis tools, but also from the creation of new tools and from improvements made in existing tools, which has been largely stimulated by the complex wealth of data yielded by large-scale sequencing projects.

Comparative genome analyses can be performed following different approaches, offering multiple perspectives on the organisms investigated (reviewed by WEI et al., 2002). Such approaches involve: (i) comparison of genome structure including description of DNA structural features, analysis of content and distribution of DNA repeats and other low complexity regions, identification of conserved synteny and genome rearrangement events, and analysis of breakpoints; (ii) comparison of coding regions comprising identification of gene-coding regions, comparison of gene and protein contents, identification/analysis of conserved orthologous and paralogous gene families (FITCH, 1970; 2000) across species, analysis of conservation of gene clusters (co-occurrence of genes in potential operons) and gene order across species, and identification/analysis of gene fusion/fission events and functionally linked genes (co-occurrence of genes) across species (MARCOTTE et al., 1999; ENRIGHT et al., 1999); (iii) comparison of non-coding regions consisting of identification of regulatory elements.

Since genomes are basically very long sequences, one might align them just as normal sequences, using one of the aforementioned algorithms. However, this task can only be accomplished with genomes of very closely related species, as changes in DNA structure (insertions, deletions, inversions, rearrangements, exchanges and

duplications) occur at a fast rate. In addition, because of their large sequence size, the alignment of more than one genome pair is computationally impracticable, even if efficient algorithms, especially developed to cope with very large sequences, are used (MORGENSTERN et al., 1998; 1999; 2002; JAREBORG et al., 1999; DELCHER et al., 1999; 2002; KENT & ZAHLER, 2000; BATZOGLOU et al., 2000; MA et al., 2002; BRAY et al., 2003; 2004; SCHWARTZ et al., 2003b; BRUDNO et al., 2003a; 2003b; KURTZ et al., 2004). Hence, in the majority of circumstances, it is more convenient to compare constitutive parts of several genomes than their whole sequences. The comparison of the complete set of genes encoded by various species is a classical example.

The crucial step of such analysis is to establish whether the sequences under comparison are homologous or not, that is, whether they descend from a common ancestral sequence or not. Since homologous sequences tend to have similar functions (BORK & KOONIN, 1998), one can use sequence homology to predict the function of an unknown gene. This nontrivial task is carried out by comparing one or more query sequences with an unrestricted number of other sequences stored in a database (subject sequences). The comparison is accomplished by aligning each query sequence with each subject sequence using a local alignment algorithm (SMITH & WATERMAN, 1981; PEARSON & LIPMAN, 1988; ALTSCHUL et al., 1997). For each alignment, the achieved score is calculated according to a substitution matrix (usually PAM [DAYHOFF et al., 1978] or BLOSUM [HENIKOFF & HENIKOFF, 1992]) and arbitrary values of gap opening/extension penalties; the number of different alignments with scores equivalent to or better than the one achieved by the alignment under consideration that are expected to occur in a database search by chance alone (E-value) is also calculated, based on the normalized score (bitscore), and the size and composition of the database. Finally, homology is inferred according to the calculated sequence alignment parameters: score, bitscore, E-value, fraction of identical positions and overlapped regions within the aligned pair, etc. The existence of conserved domains (modules that form evolutionary, functionally and structurally independent units) in proteins should not be overlooked, since it may cause serious difficulties in such an analysis.

Comparing genomes: available computational resources for comparative analysis of prokaryotic genomes

Numerous public (mostly) or proprietary databases and computational tools have been created aiming to integrate, organize and optimize the access to the wealth of information generated by the aforementioned high-throughput projects (exhaustively reviewed by HIGGINS & TAYLOR, 2000), as well as allowing the comparative analysis of this massive amount of data (Table 2). The creation and maintenance of biological databases is a challenge by itself, not only because it usually involves a large number of data, but mostly because it requires the designing of schemes and frameworks that accurately represent the complexity of biological systems, which is frequently a hard task to be accomplished (MACÊDO et al., 2003). Another difficulty is the development of efficient data retrieval systems, implemented in user-friendly interfaces and intended for complex and massive database searching. It is worth noting that in many circumstances the authors and curators of such databases receive little or no remuneration for their productive efforts. In addition, to obtain financial support for creation and maintenance of biological databases is still a difficult task (GALPERIN, 2005).

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes

Name	Description	Reference(s)	URL
DATABASES			
Generic and multifunctional			
BacMap	Interactive atlas (collection of high-resolution genomic maps) designed for visual exploration of bacterial genomes. Provides extensive gene annotation, and offers for each genome graphics representing global statistics, such as base and amino acid composition, protein length distribution, strand preference, among others.	Stothard et al., 2005	http://wishart.biology.ualberta.ca/BacMap/
CMR	Comprehensive Microbial Resource. Provides access to a range of information about and analyses of all completely sequenced prokaryotic genomes. Queries can be done by gene, genome, genomic regions and gene properties. Comparison of multiple genomes can be accomplished using distinct strategies, such as sequence similarity and gene attributes.	Peterson et al., 2001	http://cmr.tigr.org/

cont.

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes (cont.)

Genome Atlas Database	Developed for visualise and compare DNA structural features of completely sequenced microbial genomes, such as base composition, stacking energy, strand preference, DNase I sensitivity, intrinsic curvature, among others.	Hallin & Ussery 2004	http://www.cbs.dtu.dk/services/GenomeAtlas/
IMG	Integrated Microbial Genomes. Platform for comparative analysis of genomes sequenced by the Joint Genome Institute (DOE). Dedicated to facilitate visualisation and exploration of genomes according to a functional and evolutionary perspective.	Markowitz et al., 2006	http://img.jgi.doe.gov
MBGD	Microbial Genome Database. The system offers orthologous gene clustering using self-developed algorithm (DomClust), based on precomputed sequence similarity data and user-defined parameters. MBGD provides phylogenetic profile analysis, gene order and structure comparison, and functional classification.	Uchiyama 2003, 2006	http://mbgd.genome.ad.jp/
MicrobesOnline	Database for comparative analysis of prokaryotic genomes. Integrates several available sequence/genomic analysis tools, and provides precomputed data of operon prediction and orthologous groups in hundreds of prokaryotic genomes.	Alm et al., 2005	http://www.microbesonline.org/
PLATCOM	Platform for computational comparative genomics. Workspace where users are allowed to select groups of genomes, among hundreds of genomes, and compare them with a set of interconnected sequence analysis tools and local databases, establishing their own experimental protocol to investigate sequence similarities, synteny, conservation of metabolic pathways, and putative gene fusion/fission events.	Choi et al., 2005	http://platcom.informatics.indiana.edu/platcom/
PUMA2	Interactive and integrated bioinformatics system for massive sequence analysis and metabolic reconstruction. Provides a framework for comparative and evolutionary analyses of genomes and metabolic networks, within a taxonomic and phenotypic context. It presents more than 1,000 prokaryotic and eukaryotic genomes, as well as viral and mitochondrial genomes.	Maltsev et al., 2006	http://compbio.mcs.anl.gov/puma2/
	Organism or group–specific		
GenoList	Collection of databases dedicated to microbial genome analysis. Provides a complete dataset of protein and nucleotide sequences for selected species, as well as annotation and functional classification of such sequences. Searching/retrieval options include: gene name, gene localization, keywords, functional category, pattern searching, and sequence similarity searching.	Fang et al., 2005	http://genolist.pasteur.fr/
GenoMycDB	Relational database for large-scale comparative analysis of six completely sequenced mycobacterial genomes based on their predicted protein content. The database provides for each protein sequence the predicted subcellular localization, the assigned COG(s), features of the corresponding gene and links to several important databases. Tables containing pairs or groups of inferred homologs between selected species/strains can be created dynamically based on user-defined criteria.	Catanho et al., 2006	http://www.dbbm.fiocruz.br/GenoMycDB
LEGER	Database for comparative analysis of Listeria genomes. Provides precomputed genome comparison results and inferred orthologs, offering: functional analyses (including metabolic pathways), data searching/retrieval and data mining based on self-developed systems, among others. The database also provides integrated proteomic analysis results.	Dieterich et al., 2006	http://leger2.gbf.de/cgi-bin/expLeger.pl

cont.

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes (cont.)

MolliGen	Database for comparative analysis of Mollicutes genomes. Provides precomputed genome comparison results and inferred orthologs, offering: functional analyses (including metabolic pathways), data searching/retrieval and data mining based on self-developed systems, among others.	Barré et al., 2004	http://cbi.labri.fr/outils/molligen/
ShiBASE	Database for comparative analysis of Shigella genomes. Provides precomputed genome comparison results and inferred orthologs, offering: functional analyses (including metabolic pathways), data searching/retrieval and data mining based on self-developed systems, among others. The database also provides integrated large-scale comparative hybridization analysis (microarray) results.	Yang et al., 2006	http://www.mgc.ac.cn/ShiBASE/
xBASE	Collection of databases dedicated to bacterial comparative genome analysis. Provides precomputed data of comparative genome analyses among selected bacterial genera, as well as inferred orthologous groups and functional annotations. It also provides precomputed analyses of codon usage, base composition, CAI (codon adaptation index), hydropathy and aromaticity of their protein coding sequences. Searching/retrieval options include: gene name, gene localization, gene annotation, etc.	Chaudhuri & Pallen 2006	http://xbase.bham.ac.uk/
	Specialized		
COG	Clusters of Orthologous Groups. Represents an attempt to phylogenetically classify groups of predicted proteins encoded by completely sequenced prokaryotic (and also eukaryotic) genomes. Provides a range of precomputed data, such as phylogenetic patterns, functional classification, and clusters of orthologous groups (COGs) according to both functional categories and metabolic pathways, among others.	Tatusov et al., 1997, 2003	http://www.ncbi.nlm.nih.gov/COG
FusionDB	The FusionDB presents comprehensive analyses of gene fusion/fission events in prokaryotes, providing resources to investigate potential protein-protein interactions and regulatory metabolic networks.	Suhre & Claverie 2004	http://igs-server.cnrs-mrs.fr/FusionDB/
HAMAP	High-Quality Automated and Manual Annotation of Microbial Proteomes. Collection of microbial orthologous protein families manually created by experts (curators). Provides for each family extensive annotation, alignments, profiles and computed attributes (transmembrane regions, signal peptide, etc).	Gattiker et al., 2003	http://www.expasy.org/sprot/hamap/
Hogenom	Database of homologous sequences of completely sequenced genomes. Provides retrieval of homologous sequences among species and visualization of multiple sequence alignments and phylogenetic trees.	Dufayard et al., 2005	http://pbil.univ-lyon1.fr/databases/hogenom.html
IslandPath	The system integrates features frequently associated to genomic islands - as anomalous GC content, nucleotide composition biases, etc - in a graphical representation of prokaryotic genomes, assisting the recognition of genomic islands.	Hsiao et al., 2003	http://www.pathogenomics.sfu.ca/islandpath/

cont.

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes (cont.)

KEGG	Kyoto Encyclopedia of Genes and Genomes. Integrates several databases grouped in three main categories: molecular interaction networks in biological processes (biochemical pathways); information concerning the universe of genes and proteins; and information about the range of chemical components and reactions. It provides a collection of manually created maps of biochemical pathways, and precomputed results of comparative sequence analysis, pattern/motif searching, and orthologous gene clusters, among others.	Kanehisa 1997; Kanehisa & Goto 2000; Kanehisa et al., 2006	http://www.genome.jp/kegg
MetaCyc	Non-redundant database of experimentally verified metabolic pathways, involving 700 pathways in more than 600 organisms. Provides a range of information about metabolic pathways, enzymatic reactions, enzymes, chemical compounds, genes, etc, as well as a range of applications, such as computational prediction of metabolic pathways, and comparative analysis of biochemical networks, among others.	Caspi et al., 2006	http://metacyc.org/
OMA Browser	Web interface that provides access to and exploration of pairs and groups of orthologs in a database that integrates the results of the OMA project of identification of orthologs in completely sequenced genomes.	Schneider et al., 2007	http://omabrowser.org/
ORFanage	Database developed to analyse and classify orphan genes, i.e. genes that are exclusive of a particular species, family or lineage (taxonomically restricted genes). Searches can be accomplished using predefined orphan gene classes (unique, paralogs or orthologs).	Siew et al., 2004	http://www.cs.bgu.ac.il/~nomsiew/ORFans/
OrphanMine	Database dedicated to comparative analysis of orphan genes. Users are able to detect orphan genes according to several criteria (sequence similarity, sequence length, GC content, etc).	Wilson et al., 2005	http://www.genomics.ceh.ac.uk/orphan_mine/faq.php
OrthoMCL-DB	Database of orthologous groups involving 55 species (prokaryotes and eukaryotes). The groups were established on the basis of sequence similarity using a self-developed algorithm (OrthoMCL). The system provides visualization and analysis of phylogenetic profiles, domain architecture, and sequence similarity, among others.	Chen et al., 2006	http://orthomcl.cbil.upenn.edu
ProtRepeatsDB	Database of amino acid repetitions in protein sequences of completely sequenced genomes. Provides a set of tools for large-scale identification of amino acid repetitions, facilitating comparative and evolutionary analyses of such repetitions.	Kalita et al., 2006	http://bioinfo.icgeb.res.in/repeats/
RoundUp	Repository of orthologous gene groups and their evolutionary distances involving hundreds of species, achieved with a self-developed algorithm (Reciprocal Smallest Distance). The system provides data searching/retrieval based on genes or genomes, displaying the results as phylogenetic profiles, combined with gene annotation and molecular function.	Deluca et al., 2006	https://rodeo.med.harvard.edu/tools/roundup/
SEED	Extensively curated, non-redundant database developed by the Fellowship for Interpretation of Genomes (FIG), that integrates data from diverse sources (GenBank, RefSeq, UniProt, KEGG, and other genome sequencing centres). Provides a platform to support comparative analyses of genomes, opened to contributions from the whole scientific community, in which the genome annotation is performed according to subsystems (biochemical pathways and functionally linked genes).	Overbeek et al., 2005	http://theseed.uchicago.edu/FIG/index.cgi

cont.

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes (cont.)

STRING	Search Tool for the Retrieval of Interacting Genes/Proteins. Database of predicted protein interactions, including direct (physical) and indirect (functional) linkages, based on: genomic context, high-throughput experiments, co-expression data, and previous knowledge.	von Mering et al., 2005, 2007	http://string.embl.de/
TransportDB	Relational database describing the predicted cytoplasmic membrane transport protein complement for organisms whose the genome has been completely sequenced. For each organism, the complete set of membrane transport systems was identified and classified into different types and families according to putative membrane topology, protein family, bioenergetics, and substrate specificities. The database provides similarity searching, comparison of transport systems from different organisms and phylogenetic trees of individual transporter families.	Ren et al., 2004,2007	http://www.membranetransport.org/
	Phylogenomic		
BPhyOG	Bacterial Phylogenies Based on Overlapping Genes. Interactive web server dedicated to phylogeny reconstruction of completely sequenced bacterial genomes, based on their shared overlapping gene content.	Luo et al., 2007	http://cmb.bnu.edu.cn/BPhyOG/
PHOG	Phylogenetic Orthologous Groups. Database of homologous genes, involving dozens of species (prokaryotes and eukaryotes), built according to the taxonomy tree representing these organisms. The system implements a completely automated procedure that creates clusters of orthologous groups at each node of the taxonomy tree.	Merkeev et al., 2006	http://bioinf.fbb.msu.ru/phogs/index.html
Phydbac	Phylogenomic Display of Bacterial Genes. Provides interactive visualization and comparison of phylogenetic profiles derived from protein sequences of hundreds of bacteria, allowing detection of functionally related proteins and conservation patterns across these organisms.	Enault et al., 2004	http://igs-server.cnrs-mrs.fr/phydbac/
SHOT	System developed for genome phylogeny reconstruction, providing construction of phylogenetic trees for hundreds of organisms whose the genome has been completely sequenced, based on the shared gene content or the conservation of gene order across the species.	Korbel et al., 2003	http://www.Bork.EMBL-Heidelberg.de/SHOT
	Genomic metadata		
Genome Properties	System developed to present key aspects of prokaryotic biology using standardized computational methods and controlled vocabularies. Properties reflect gene content, phenotype, phylogeny and computational analyses. Comparisons can be accomplished based on several attributes.	Haft et al., 2005; Selengut et al., 2007	http://www.tigr.org/Genome_Properties/
GenomeMine	This database integrates a range of information about all completely sequenced genomes, derived from heterogeneous sources, such as Genome (NCBI) and GOLD (Genomes Online Database) databases, and data achieved from genomic sequences. Comparisons can be accomplished based on several attributes.	-	http://www.genomics.ceh.ac.uk/GMINE/
SACSO	Systematic Analysis of Completely Sequenced Organisms. Provides comparative analysis of completely sequenced organisms including base composition, amino acid composition, ancestral duplication, ancestral conservation, and classification of organisms as obtained from their intra and inter predicted proteome comparisons.	Tekaia et al., 2002	http://www.pasteur.fr/~tekaia/sacso.html

cont.

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes (cont.)

COMPUTATIONAL TOOLS			
Interactive genome browsing			
ABC	Application for Browsing Constraints. Program for interactive exploration of genomic multiple sequence alignments. Provides simultaneous display of quantitative data (sequence similarities or evolutionary rates, for instance) and annotation (such as gene localization and repeats).	Cooper et al., 2004	http://mendel.stanford.edu/sidowlab/downloads.html
ACT	Artemis Comparison Tool. The program allows interactive visualisation of comparisons between complete genome sequences and associated annotations. The comparison data can be generated with several different alignment programs, making it possible to identify syntenic regions, insertions and rearrangements.	Carver et al., 2005	http://www.sanger.ac.uk/Software/ACT/
AutoGRAPH	Integrated web server for multi-species comparative genomic analysis, based on precomputed or user supplied data. Provides construction and visualization of synteny maps between two or three species, determination and displaying of macrosynteny and microsynteny relationships among species, and highlighting of evolutionary breakpoints.	Derrien et al., 2007	http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/
CGAT	Comparative Genome Analysis Tool. Program for interactive visualization and comparison of aligned genome pairs, along with their associated annotation. It offers a generic framework for processing/visualizing genomic alignments using several exiting programs.	Uchiyama et al., 2006	http://mbgd.genome.ad.jp/CGAT/
Cinteny	Web server dedicated to find syntenic regions across multiple genomes and measure the extent of genome rearrangement using reversal distance as a measure.	Sinha & Meller 2007	http://cinteny.cchmc.org/
ComBo	Comparative Genome Browser. Provides dynamic view of whole genome alignments along with their associated annotations.	Engels et al., 2006	http://www.broad.mit.edu/annotation/argo/
DNAVis	Program package that provides interactive and real-time visualization of DNA sequences and their comparative genome annotations.	Fiers et al., 2006	http://www.win.tue.nl/dnavis/
GECO	Provides linear visualization of multiple prokaryotic genomes, allowing detection of lateral gene transfer, pseudogenes, and insertion/deletion events among related species. The program is able to display ortholog relations (calculated using the algorithm implemented in the software BLASTCLUST, which is part of the BLAST program package), and identify irregularities on the genomic level based on anomalous GC composition.	Kuenne et al., 2007	http://bioinfo.mikrobio.med.uni-giessen.de/geco2/GecoMainServlet
GenColors	Program developed to improve and accelerate annotation of prokaryotic genomes, considering information on related genomes and making extensive use of genome comparison. The available comparative tools provide detection of bidirectional best hits, conserved genes, and synteny, among others.	Romualdi et al., 2005	http://gencolors.imb-jena.de
GeneOrder3.0	Program for comparison of gene order and synteny between pairs of small bacterial genomes.	Celamkoti et al., 2004	http://binf.gmu.edu/genometools.html

cont.

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes (cont.)

GenomeViz	The program allows visualization of both qualitative and quantitative information - deriving from studies on genomic islands, gene/protein classifications, GC content, GC skew, whole genome alignments, microarrays and proteomics - from completely and partially sequenced microbial genomes.	Ghai et al., 2004	http://www.uniklinikum-giessen.de/genome/genomeviz/intro.html
G-InforBIO	Integrated system for microbial genomics. The system can import genome data (annotation or sequences) from diverse sources and formats, creating a local database to store the data. It provides a range of searching/retrieval mechanisms, data exporting tools, and visualization and comparative analysis tools.	Tanaka et al., 2006	http://rhodem17.ddbj.nig.ac.jp/inforbio/
inGeno	Interactive visualization platform for sequence comparisons between complete genome sequences and all associated annotations and features. Comparisons can be made with several different sequence analysis programs, providing identification of syntenic regions, inversions and rearrangements.	Liang & Dandekar 2006	http://ingeno.bioapps.biozentrum.uni-wuerzburg.de/
MuGeN	Program package built for navigating through multiple annotated genomes, able to retrieve annotated sequences in several formats, including user supplied data.	Hoebeke et al., 2003	http://genome.jouy.inra.fr/MuGeN/
SynBrowse	Synteny Browser for comparative sequence analysis. Software for visualization and comparative analysis of aligned genomes, providing identification of conserved sequences, syntenic regions, inversions and rearrangements.	Pane et al., 2005	http://www.synbrowse.org/
SynView	Interactive and customizable software for visualization and comparative analysis of multiple genomes, providing identification of conserved sequences, syntenic regions, inversions and rearrangements.	Wang et al., 2006	http://www.ApiDB.org/apps/SynView/
Large-scale genomic sequences comparison			
BioParser	Provides a set of user-friendly interfaces for parsing and analysing sequence alignment data in large-scale. The program is able to parse sequence alignment reports obtained with several local alignment programs (BLAST, FASTA, SSEARCH, and HMMER). Users are allowed to dynamically select/retrieve pairs or groups of sequences based on user-defined criteria (computed similarity indices, sequence description and length, among others). The program simplifies the analysis of data produced by the most common sequence similarity searching softwares, making it easier, for instance, to identify evolutionary, structural or functional relationships among the compared sequences, based on their degree of similarity.	Catanho et al., 2006	http://www.dbbm.fiocruz.br/BioParser.html
BSR	The BLAST Score Ratio Analysis Tool. The program allows visual evaluation of the level of conservation of any three predicted proteomes and the degree to which the genome structure among the three genomes is similar, based on a self-developed algorithm (BLAST Score Ratio).	Rasko et al., 2005	http://www.microbialgenomics.org/BSR/
COMPAM	Tool for visualizing relationships among multiple whole genomes by combining all pairwise genome alignments.	Lee et al., 2006	http://bio.informatics.indiana.edu/projects/compam/

cont.

Table 2 - Main databases and computational tools available for comparative analysis of prokaryotic genomes (cont.)

GenomeBlast	Web tool developed for comparative analysis of multiple small genomes, based on user supplied data. The program allows identification of unique genes and homologous genes, visualization of their distribution across the compared genomes, and genome phylogeny reconstruction.	Lu et al., 2006	http://bioinfo-srv1.awh.unomaha.edu/genomeblast/
GenomeComp	Tool for parsing and visual comparison of large-scale data derived from BLAST local alignments of genomic sequences from multiple organisms. It provides detection of repeat regions, insertions, deletions and rearrangements of genomic segments.	Yang et al., 2003	http://www.mgc.ac.cn/GenomeComp/
GenomePixelizer	Visualization tool that generates custom images of physical or genetic positions of specified sets of genes in whole genomes or genomic segments. The program allows the analysis of duplication events within and between species based on sequence similarities.	Kozik et al., 2002	http://www.atgc.org/GenomePixelizer/
M-GCAT	Multiple Genome Comparison and Alignment Tool. Program for multiple alignment and visualization of whole genomes or large DNA segments, based on self-developed algorithm.	Treangen & Messeguer 2006	http://alggen.lsi.upc.es/recerca/align/mgcat/intro-mgcat.html
MUMmer	System for multiple alignment and visualization of whole genomes or large DNA segments, based on self-developed algorithm (Space efficient suffix trees).	Kurtz et al., 2004	http://www.tigr.org/software/mummer/
PipMaker, PipTools, MultiPipMaker, zPicture	Set of tools for aligning and visualizing, in various formats, whole genomes or genomic segments. It allows the dynamically creation of conservation profiles and identification of evolutively conserved regions.	Schwartz et al., 2000 2003a; Elnitski et al., 2002; Ovcharenko et al., 2004	http://bio.cse.psu.edu/
PyPhy	Set of tools for automatic, large-scale reconstructions of phylogenetic relationships of completely sequenced microbial genomes.	Sicheritz-Ponten & Andersson 2001	http://www.cbs.dtu.dk/staff/thomas/pyphy/
VISTA	Set of computational tools for comparative genomics. Provides algorithms for aligning and visualizing large genomic fragments, along with their associated functional annotations.	Frazer et al., 2004; Brudno et al., 2007	http://www-gsd.lbl.gov/vista/

Overall, databases for comparative analyses of prokaryotic genomes can be divided in five main categories, according to their purposes and functionalities: (i) generic and multifunctional, (ii) organism or group-specific, (iii) specialized, (iv) phylogenomic, and (v) genomic metadata (Table 2). On the other hand, the computational tools can be grouped in (i) interactive genome browsing programs and (ii) large-scale genomic sequences comparison programs (Table 2). Certainly, these classifications are not definitive or perhaps the most suitable, since the purposes and functionalities of these databases and tools are naturally overlapped. Alternative classification schemes are therefore feasible and equally valid (FIELD et al., 2005; GALPERIN, 2005).

Most generic and multifunctional databases presented in this review are dedicated to cover the universe of prokaryotic species (and sometimes also eukaryotic species) whose genomes have been completely sequenced, and to offer the required resources to search/retrieve precomputed (mostly) and/or experimentally achieved data for each species (BacMap, CMR, Genome Atlases, IMG, MBGD, Microbes Online, PLATCOM, PUMA2). The information offered and the available searching/retrieval and analysis tools vary significantly from one database to another. They may comprise, for instance, (i) physico-chemical, structural, statistical, functional, evolutionary, taxonomic, and/or phenotypical features associated to entire genomes or to their coding and/or non-coding regions (Figure 1), and (ii) searching/retrieval mechanisms based on keywords, gene/coding sequence and/or species name/identification number, and based on pairwise comparison of entire genomes, genomic sequences or coding regions using local or global alignment algorithms. All these particularities also apply to the organism or group-specific databases (GenoList, GenoMycDB, LEGER, MolliGen, ShiBASE, xBASE), which in contrast are dedicated to particular microbes.

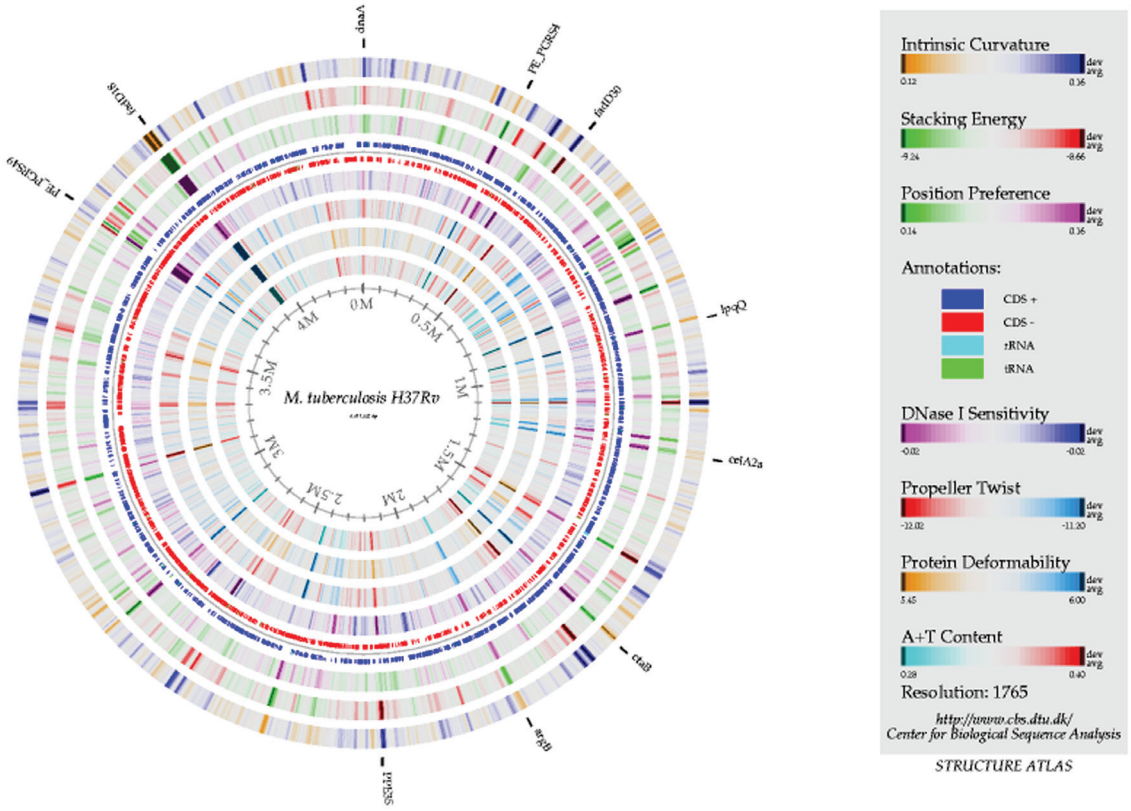


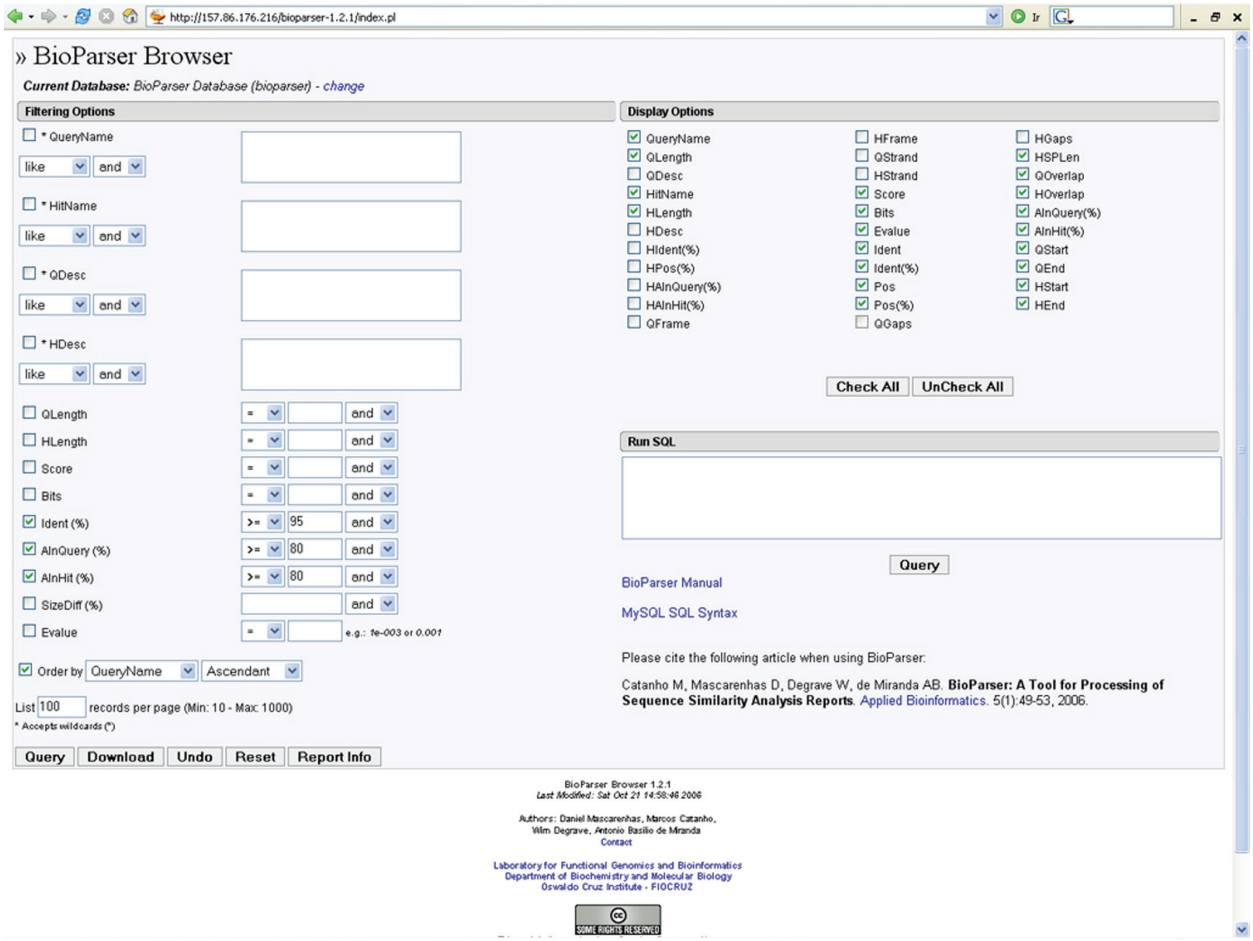
Figure 1 - DNA structural atlas of *Mycobacterium tuberculosis* H37Rv genome. The concentric circles represent seven distinct DNA structural features (see legend in the figure). The fourth and fifth circles, from the outer to the inner circle, represent the distribution of coding regions annotated in the DNA strand (positive strand in blue and negative strand in red, respectively) and the distribution of ribosomal (light-blue) and transfer (green) RNA coding regions in the genome. The measured values of each structural parameter are represented by colour scales, allowing visual inspection of their variation within the genome. Similar maps, representing these and other biological features, can be easily retrieved (or achieved with user supplied data) from the Genome Atlas Database (GenomeAtlas 2007) and, then visually compared. Detailed explanation about DNA structural features and their relevance can be found in the Genome Atlas Database website.

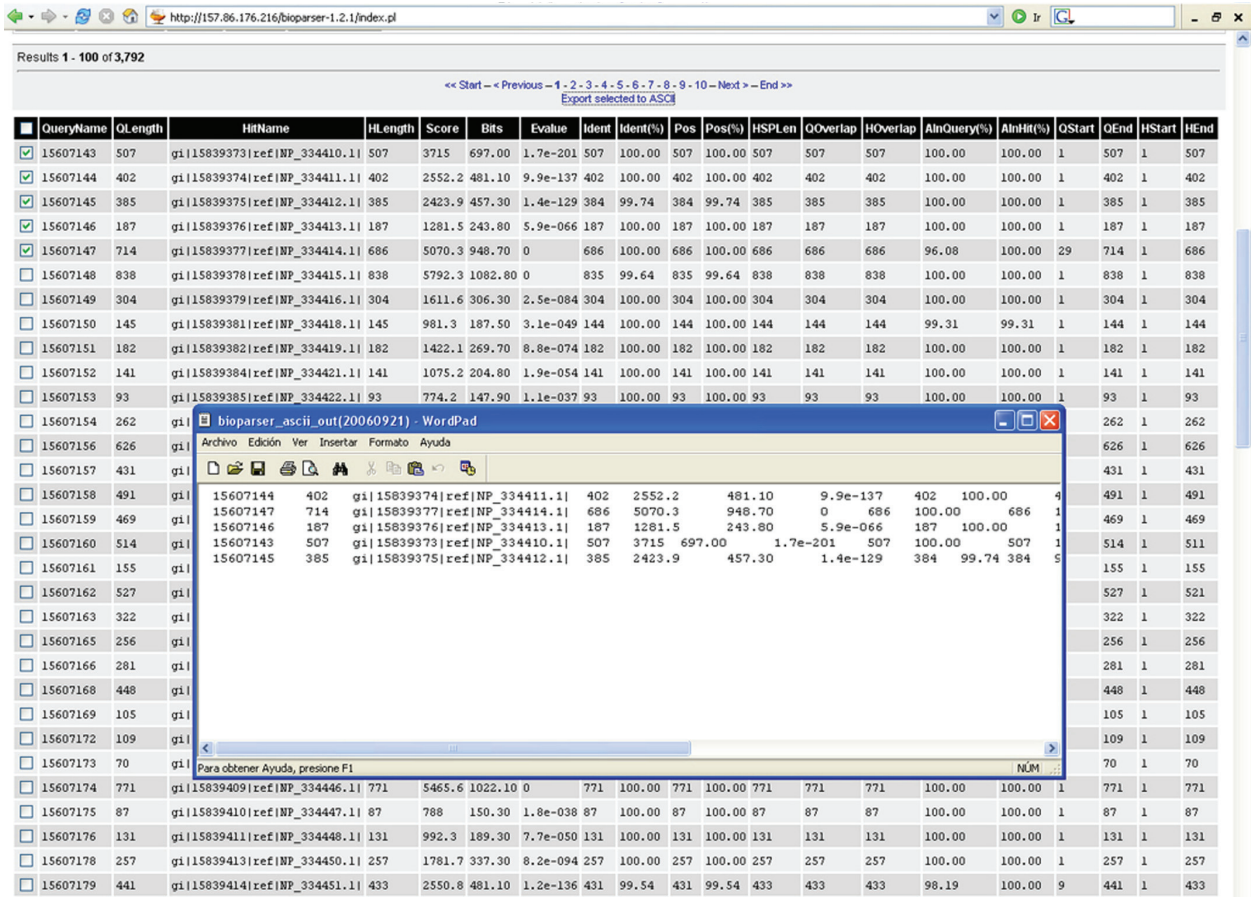
On the other hand, there are an increasing number of databases dedicated to the comparative analysis of particular features of genomes and their components. Among the features explored by these specialized databases, one may distinguish: (i) conservation of orthologous genes (or proteins) across species (COG, HAMAP, Hogenom, OMA Browser, OrthoMCL-DB, RoundUp); (ii) gene fusion/fission events (FusionDB); (iii) occurrence of genomic islands (IslandPath); (iv) incidence of amino acid repetitions in proteins (ProtRepeatsDB); (v) incidence and characterization of orphan genes (ORFanage, OrphanMine) or functional groups, such as genes involved in cellular subsystems (SEED) or even membrane transport proteins (TransportDB); (vi) configuration of protein interaction networks (STRING); and (vii) incidence and conservation of metabolic pathways (KEGG, MetaCyc).

In the last years, the development of phylogenetic methods that explore the entire gene content of completely sequenced genomes (phylogenomics, as opposed to classical approaches employing only a few selected genes) has originated several phylogenomic databases, providing: (i) visualization and comparison of phylogenetic profiles (co-occurrence of genes across species (MARCOTTE et al., 1999) (Phydbac); (ii) phylogeny reconstruction on the basis of conserved gene content (BPhyOG, SHOT) or conservation of gene order (SHOT) across species; or (iii) analysis of *phylogenetic orthologous groups*, orthologous clusters built according to the taxonomy tree of numerous

organisms (PHOG). In addition, databases dedicated to comparative studies of genomic metadata has also been developed in recent years, based on analyses of (i) information achieved from genomes and particular groups of genes in hundreds of microbial species, and also partially based on (ii) information compiled from published scientific studies. These databases make it possible to investigate interesting relationships among lifestyle, evolutionary history and genomic features (Genome Properties, GenomeMine, SACSO).

Most computational tools developed for comparative genome analyses are dedicated to interactive visualization and browsing. They offer different graphical environments for (i) visual comparison and browsing of pairs (ATC, Cinteny, DNAVis, GeneOrder3.0, G-InforBIO, inGeno, SynBrowse) or groups (AutoGRAPH, GECO, GenColors, GenomeViz, MuGeN, SynView) of genomes or genomic sequences, and for (ii) visual investigation of multiple alignments of genomic sequences (ABC, CGAT, ComBo). Another group of tools are based on (i) large-scale sequence comparison involving multiple genomes using local (BioParser, BSR, COMPAM, GenomeBlast, GenomeComp) (Figura 2) or global (M-GCAT, MUMmer, PipMaker/PipTools/MultiPipMaker/zPicture, VISTA, PyPhy) alignment algorithms, or (ii) physical or genetic positions of specified groups of genes in whole genomes or genomic sequences and their similarity matrix (GenomePixelizer). Similarly to the aforementioned databases, the provided searching/retrieval and analysis mechanisms vary significantly from one tool to another, overlapping in many circumstances. For instance, they provide: (i) searching/retrieval mechanisms based on keywords, gene/coding sequence and/or species name/identification number; (ii) acquisition of functional gene annotations; (iii) phylogenetic reconstruction; (iv) detection of colinearity, synteny, gene duplication, orthologous and paralogous clusters, rearrangements, repetitions, inversions, insertions, deletions, restriction sites, motifs and profiles (Gribskov et al. 1987), etc. These tools are available as on-line services and/or stand-alone applications.





QueryName	QLength	HitName	HLength	Score	Bits	Evalue	Ident	Ident(%)	Pos	Pos(%)	HSPLen	QOverlap	HOverlap	AlnQuery(%)	AlnHit(%)	QStart	QEnd	HStart	HEnd
15607143	507	gi|15839373|ref|NP_334410.1|	507	3715	697.00	1.7e-201	507	100.00	507	100.00	507	507	507	100.00	100.00	1	507	1	507
15607144	402	gi|15839374|ref|NP_334411.1|	402	2552.2	481.10	9.9e-137	402	100.00	402	100.00	402	402	402	100.00	100.00	1	402	1	402
15607145	385	gi|15839375|ref|NP_334412.1|	385	2423.9	457.30	1.4e-129	384	99.74	384	99.74	385	385	385	100.00	100.00	1	385	1	385
15607146	187	gi|15839376|ref|NP_334413.1|	187	1281.5	243.80	5.9e-066	187	100.00	187	100.00	187	187	187	100.00	100.00	1	187	1	187
15607147	714	gi|15839377|ref|NP_334414.1|	686	5070.3	948.70	0	686	100.00	686	100.00	686	686	686	96.08	100.00	29	714	1	686
15607148	838	gi|15839378|ref|NP_334415.1|	838	5792.3	1082.80	0	835	99.64	835	99.64	838	838	838	100.00	100.00	1	838	1	838
15607149	304	gi|15839379|ref|NP_334416.1|	304	1611.6	306.30	2.5e-084	304	100.00	304	100.00	304	304	304	100.00	100.00	1	304	1	304
15607150	145	gi|15839381|ref|NP_334418.1|	145	981.3	187.50	3.1e-049	144	100.00	144	100.00	144	144	144	99.31	99.31	1	144	1	144
15607151	182	gi|15839382|ref|NP_334419.1|	182	1422.1	269.70	8.8e-074	182	100.00	182	100.00	182	182	182	100.00	100.00	1	182	1	182
15607152	141	gi|15839384|ref|NP_334421.1|	141	1075.2	204.80	1.9e-054	141	100.00	141	100.00	141	141	141	100.00	100.00	1	141	1	141
15607153	93	gi|15839385|ref|NP_334422.1|	93	774.2	147.90	1.1e-037	93	100.00	93	100.00	93	93	93	100.00	100.00	1	93	1	93
15607154	262	gi|															262	1	262
15607156	626	gi|															626	1	626
15607157	431	gi|															431	1	431
15607158	491	gi|															491	1	491
15607159	469	gi|															469	1	469
15607160	514	gi|															514	1	511
15607161	155	gi|															155	1	155
15607162	527	gi|															527	1	521
15607163	322	gi|															322	1	322
15607165	256	gi|															256	1	256
15607166	281	gi|															281	1	281
15607168	448	gi|															448	1	448
15607169	105	gi|															105	1	105
15607172	109	gi|															109	1	109
15607173	70	gi|															70	1	70
15607174	771	gi|15839409|ref|NP_334446.1|	771	5465.6	1022.10	0	771	100.00	771	100.00	771	771	771	100.00	100.00	1	771	1	771
15607175	87	gi|15839410|ref|NP_334447.1|	87	788	150.30	1.8e-038	87	100.00	87	100.00	87	87	87	100.00	100.00	1	87	1	87
15607176	131	gi|15839411|ref|NP_334448.1|	131	992.3	189.30	7.7e-050	131	100.00	131	100.00	131	131	131	100.00	100.00	1	131	1	131
15607178	257	gi|15839413|ref|NP_334450.1|	257	1781.7	337.30	8.2e-094	257	100.00	257	100.00	257	257	257	100.00	100.00	1	257	1	257
15607179	441	gi|15839414|ref|NP_334451.1|	433	2550.8	481.10	1.2e-136	431	99.54	431	99.54	433	433	433	98.19	100.00	9	441	1	433

Figure 2 - Comparison of the entire predicted protein content between two strains of *M. tuberculosis*, H37Rv and CDC1551 using a stand-alone version of the program BioParser. The protein sequences were obtained from RefSeq (2007) (accessions NC_000962 and NC_002755, respectively) and compared against each other using the FASTA sequence alignment program (UVa FASTA Server 2007). The FASTA sequence alignment report was parsed with BioParser and the parsed information were automatically stored in a local database, created and configured according to the instruction manual. The figure shows the result of a database searching using the provided database web interface, BioParser Browser, where only records representing alignment pairs with at least 95% identity and 80% overlap were returned. The first five records were selected and exported as flat files using the provided *Export selected to ASCII* tool. Sequence alignment reports up to 5 megabytes can be remotely parsed and analysed using the BioParser web service (BioParserWeb 2007). Details about the creation, application, use and local installation of this tool can be found in the program website (BioParser 2007) and in the paper describing it (Catanho et al. 2006).

Thinking of the future: concluding remarks and perspectives

As outlined in this review, comparative genome analysis has a range of applications in different fields, from analyses of genome structure, organization and evolution to development of more accurate methods of prevention, treatment and diagnosis of parasitic diseases, for instance. It was also shown that this holistic approach of comparative genomics benefits from the outcomes of high-throughput technologies, such as genomics, proteomics and transcriptomics. Their methods, algorithms and tools find their roots in the emergence and consolidation of new disciplines, as Bioinformatics and Computational Biology. However, despite its scientific relevance, the massive comparison of genomic data carries a range of important scientific and technical challenges, such as data storage capacity, data structure and representation, data access and manipulation, data processing speed, format compatibility and multiple tool integration.

Numerous databases and computational tools have been created in order to provide the scientific community access to a range of genomic data, as well as to the results of comparative analyses of such data. Diverse options to visualize, search, retrieve and analyse these data are offered, providing the opportunity to acquire more detailed knowledge about genomes and their respective organisms. However, this wealth of information is presently fragmented, dispersed across all these computational resources, and is redundant in many circumstances,

clearly requiring unification in order to provide a global and integral picture of the biology of such genomes and species. Ideally, the upcoming databases and computational tools should (i) offer data integration, providing multi-perspective genome analyses; (ii) combine data achieved *in silico* and curated data, improving the quality of our research; (iii) present efficient data structure, storage and processing, providing dynamic, flexible and fast data visualization, data searching, data retrieval and data analysis, via user-friendly graphical interfaces; (iv) implement a consistent and controlled vocabulary to describe the data and standardized data format, providing full data interchanging and integration with other data sources. In this way, a fruitful field for interactions and cooperation among researches from distinct areas might emerge, providing the required support to interpret and analyse this wealth of data according to a truly multidisciplinary approach.

Bibliographic references

ALM, E.J. et al. The MicrobesOnline Web site for comparative genomics. **Genome Research**, v.15, n.7, p.1015-22, 2005.

ALTSCHUL S.F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, v.215, n.3, p.403-10, 1990.

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

BARRE A., de DA; BLANCHARD, A. MolliGen, a database dedicated to the comparative genomics of Mollicutes. **Nucleic Acids Research**, v.32(Database issue), p.D307-D310, 2004.

BATZOGLOU, S. Human and mouse gene structure: comparative analysis and application to exon prediction. **Genome Research**, v.10, n.7, p.950-8, 2000.

BEHR, M.A. et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. **Science**, v.284, n.5419, p.1520-3, 1999.

BENSON, D.A. GenBank. **Nucleic Acids Research**, v.33(Database issue), p.D34-D38, 2005.

BINNEWIES, T.T. et al. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. **Functional & Integrative Genomics**, v.6, n.3, p.165-85, 2006.

BIOPARSER. Available at: <http://www.dbbm.fiocruz.br/BioParser> Accessed: 8 Oct. 2007.

BIOPARSERWEB. Available at: <http://www.dbbm.fiocruz.br/BioParserWeb> Accessed: 8 Oct. 2007.

BISTIC Definition Committee. NIH working definition of bioinformatics and computational biology. 2000. Available at: <http://www.bisti.nih.gov/CompuBioDef.pdf> Accessed: 8 Oct. 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.

BORK, P.; KOONIN, E.V. Predicting functions from protein sequences--where are the bottlenecks? **Nature Genetics**, v.18, n.4, p.313-8, 1998.

BRAY, N.; DUBCHAK, I.; PACHTER, L. AVID: A global alignment program. **Genome Research**, v.13, n.1, p.97-102, 2003.

BRAY, N.; PACHTER, L. MAVID: constrained ancestral alignment of multiple sequences. **Genome Research**, v.14, n.4, p.693-9, 2004.

BROSCH, R. et al. The evolution of mycobacterial pathogenicity: clues from comparative genomics. **Trends Microbiol**, v.9, n.9, p.452-8, 2001.

BRUDNO, M. et al.. Fast and sensitive multiple alignment of large genomic sequences. **BMC Bioinformatics**, v.4, n.1, p.66, 2003a.

BRUDNO, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. **Genome Research**, v.13, n.4, p.721-31, 2003b.

BRUDNO, M. et al. Multiple whole genome alignments and novel biomedical applications at the VISTA portal. **Nucleic Acids Research**, v.35, p.W669-W674, 2007.

CARVER, T.J. et al. ACT: the Artemis Comparison Tool. **Bioinformatics**, v.21, n.16, p.3422-3, 2005.

CASPI, R. et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. **Nucleic Acids Research**, v.34, p.D511-D516, 2006.

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, 2006.

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

CELAMKOTI S. et al. GeneOrder3.0: software for comparing the order of genes in pairs of small bacterial genomes. **BMC Bioinformatics**, v.5, p.1, p.52, 2004.

CHAUDHURI, R.R.; PALLEN, M.J. xBASE, a collection of online databases for bacterial comparative genomics. **Nucleic Acids Research**, v.34, p.D335-D337, 2006.

CHEN, F. et al. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. **Nucleic Acids Research**, v.34, p.D363-D368, 2006.

CHOI, K. et al. PLATCOM: a Platform for Computational Comparative Genomics. **Bioinformatics**, Mar 15, 2005.

CMR. Comprehensive Microbial Resource. Available at: <http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl> Accessed: 8 Oct. 2007.

COENYE, T. et al. Towards a prokaryotic genomic taxonomy. **FEMS Microbiology Reviews**, v.29, n.2, p.147-67, 2005.

COLE, S.T. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. **European Respiratory Journal**, v.36, Suppl., p.78s-86s, 2002.

COOPER, G.M.; SINGARAVELU, S.A.; SIDOW, A. ABC: software for interactive browsing of genomic multiple sequence alignment data. **BMC Bioinformatics**, v.5, n.1, p.192, 2004.

DAYHOFF, M.O.; SCHWARTZ, R.M.; ORCUTT, B.C. A model of evolutionary change in proteins. In: DAYHOFF, M.O. (ed.) **Atlas of Protein Sequence and Structure**. Washington DC: National Biomedical Research Foundation, 1978. v.5. Suppl.3. p.345-352.

DELCHER, A.L. et al. Alignment of whole genomes. **Nucleic Acids Research**, v.27, n.11, p.2369-76, 1999.

DELCHER, A.L. et al. Fast algorithms for large-scale genome alignment and comparison. **Nucleic Acids Research**, v.30, n.11, p.2478-83, 2002.

DELUCA, T.F. et al. Roundup: a multi-genome repository of orthologs and evolutionary distances. **Bioinformatics**, v.22, n.16, p.2044-6, 2006.

DERRIEN, T. et al. AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. **Bioinformatics**, v.23, n.4, p.498-499, 2007.

DIETERICH G, et al.. LEGER: knowledge database and visualization tool for comparative genomics of pathogenic and non-pathogenic Listeria species. **Nucleic Acids Research**, v.34, p.D402-D406, 2006.

DUFAYARD, J.F. et al. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. **Bioinformatics**, v.21, n.11, p.2596-603, 2005.

ELNITSKI, L. et al. PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. **Genomics**, v.80, n.6, p.681-90, 2002.

ENAULT, F. et al. Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. **Nucleic Acids Research**, v.32, p.W336-W339, 2004.

ENGELS, R. et al. Combo: a whole genome comparative browser. **Bioinformatics**, v.22, n.14, p.1782-3, 2006.

ENRIGHT, A.J. et al. Protein interaction maps for complete genomes based on gene fusion events. **Nature**, v.402, n.6757, p.86-90, 1999.

FANG, G, et al. Specialized microbial databases for inductive exploration of microbial genome sequences. **BMC Genomics**, v.6, n.1, p.14, 2005.

FELSENSTEIN, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. **Journal of Molecular Evolution**, v.17, n.6, p.368-76, 1981.

FELSENSTEIN, J. PHYLIP -- Phylogeny Inference Package (Version 3.2). **Cladistics**, v.5, p.164-6, 1989.

FENG; D.F.; DOOLITTLE, R.F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. **Journal Molecular Evolution**, v.25, n.4, p.351-60, 1987.

FIELD, D.; FEIL, E.J.; WILSON, G.A. Databases and software for the comparison of prokaryotic genomes. **Microbiology**, v.151, n.Pt 7, p.2125-32, 2005.

FIERS, M.W. et al. DNAVis: interactive visualization of comparative genome annotations. **Bioinformatics**, v.22, n.3, p.354-5, 2006.

FITCH, W.M. Distinguishing homologous from analogous proteins. **Systematic Zoology**, v.19, n.2, p.99-113, 1970.

FITCH, W.M. Homology a personal view on some of the problems. **Trends in Genetics**, v.16, n.5, p.227-31, 2000.

FITZGERALD, J.R.; MUSSER, J.M. Evolutionary genomics of pathogenic bacteria. **Trends Microbiol**, v.9, n.11, p.547-53, 2001.

FRASER, C.M. et al. Comparative genomics and understanding of microbial biology. **Emerging Infectious Diseases**, v.6, n.5, p.505-12, 2000.

FRAZER, K.A. et al. VISTA: computational tools for comparative genomics. **Nucleic Acids Research**, v.32, p.W273-W279, 2004.

GALPERIN, M.Y. The Molecular Biology Database Collection: 2005 update. **Nucleic Acids Research**, v.33, p.D5-24, 2005.

GATTIKER, A. et al. Automated annotation of microbial proteomes in SWISS-PROT. **Comput Biol Chem**, v.27, n.1, p.49-58, 2003.

GENOMEATLAS. CBS Genome Atlas Database. Available at: <http://www.cbs.dtu.dk/services/GenomeAtlas/>. Accessed: 8 Oct. 2007.

GENOME PROJECT. NCBI Entrez Genome Project Database. Available at: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj>. Accessed: 8 Oct. 2007.

GHAI, R.; HAIN, T.; CHAKRABORTY; T. GenomeViz: visualizing microbial genomes. BMC Bioinformatics, v.5, n.1, p.198, 2004.

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

GOODFELLOW, M.; MINNIKIN, D.E. Circunscription of the genus. In: KUBICA, G.P.; WAYNE, L.G. (eds.) **The Mycobacteria: A Source Book**. New York: Marcel Dekker; 1984. p.1-24.

GORDON, S.V. et al. Royal Society of Tropical Medicine and Hygiene Meeting at Manson House, London, 18th January 2001. Pathogen genomes and human health. Mycobacterial genomics. **Transactions of the Royal**

Society of Tropical Medicine and Hygiene, v.96, n.1, p.1-6, 2002.

GRIBSKOV, M.; MCLACHLAN, A.D.; EISENBERG, D. Profile analysis: detection of distantly related proteins. **Proceedings of National Academy of Science**, v.84, n.13, p.4355-8, 1987.

HAFT, D.H. et al. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. **Bioinformatics**, v.21, n.3, p.293-306, 2005.

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

HALLIN, P.F.; USSERY, D.W. CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. **Bioinformatics**, v.20, n.18, p.3682-6, 2004.

HENIKOFF, S.; HENIKOFF, J.G. Amino acid substitution matrices from protein blocks. **Proceedings of National Academy of Science**, v.89, n.22, p.10915-9, 1992.

HENZ, S.R. et al. Whole-genome prokaryotic phylogeny. **Bioinformatics**, v.21, n.10, p.2329-35, 2005.

HGP. HUMAN GENOME PROGRAM (USA). U.S. Department of Energy. Genomics and Its Impact on Medicine and Society: A 2001 Primer; 2001.

HIGGINS, D.; TAYLOR, W.R. Bioinformatics sequence, structure, and databanks: a practical approach. Oxford: Oxford University Press, 2000.

HSIAO, W. et al.. IslandPath: aiding detection of genomic islands in prokaryotes. **Bioinformatics**, v.19, n.3, p.418-20, 2003.

HOEBEKE, M.; NICOLAS, P.; BESSIERES, P. MuGeN: simultaneous exploration of multiple genomes and computer analysis results. **Bioinformatics**, v.19, n.7, p.859-64, 2003.

JAREBORG, N.; BIRNEY, E.; DURBIN, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res, v.9, n.9, p.815-24, 1999.

KALITA, M.K. et al. ProtRepeatsDB: a database of amino acid repeats in genomes. BMC **Bioinformatics**, v.7, p.336, 2006.

KANEHISA, M. A database for post-genome analysis. **Trends Genet**, v.13, n.9, p.375-6, 1997.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, v.28, n.1, p.27-30, 2000.

KANEHISA, M. et al. From genomics to chemical genomics: new developments in KEGG. **Nucleic Acids Research**, v.34, p.D354-7, 2006.

KATO-MAEDA, M. et al. Comparing genomes within the species Mycobacterium tuberculosis. Genome Res, v.11, n.4, p.547-54, 2001.

KENT, W.J.; ZAHLER, A.M. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res, v.10, n.8, p.1115-25, 2000.

KONDRASHOV, A.S. Comparative genomics and evolutionary biology. **Current Opinion in Genetics and Development**, v.9, n.6, p.624-9, 1999.

KOONIN, E.V.; ARAVIND, L.; KONDRASHOV; A.S. The impact of comparative genomics on our understanding of evolution. Cell, v.101, n.6, p.573-6, 2000.

KORBEL, J.O. et al. SHOT: a web server for the construction of genome phylogenies. **Trends in Genetics**, v.18, n.3, p.158-62, 2002.

KOZIK, A.; KOCHETKOVA, E.; MICHELMORE, R. GenomePixelizer--a visualization program for comparative genomics within and between species. **Bioinformatics**, v.18, n.2, p.335-6, 2002.

KUENNE, C.T. et al. GECO--linear visualization for comparative genomics. Bioinformatics, v.23, n.1, p.125-126, 2007.

KUNIN, V. et al. Measuring genome conservation across taxa: divided strains and united kingdoms. **Nucleic Acids Research**, v.33, n.2, p.616-21, 2005a.

KUNIN, V. et al. The net of life: reconstructing the microbial phylogenetic network. **Genome Research**, v.15, n.7, p.954-9, 2005b.

KURTZ, S. et al. Versatile and open software for comparing large genomes. **Genome Biology**, v.5, n.2, R12, 2004.

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

LEE D. et al. COMPAM: visualization of combining pairwise alignments for multiple genomes. **Bioinformatics**, v.22, n.2, p.242-4, 2006.

LIANG, C.; DANDEKAR, T.; inGeno--an integrated genome and ortholog viewer for improved genome to genome comparisons. **BMC Bioinformatics**, v.7, p.461, 2006.

LIPMAN, D.J.; PEARSON, W.R. Rapid and sensitive protein similarity searches. **Science**, v.227, p.4693, p.1435-41, 1985.

LU, G. et al. GenomeBlast: a web tool for small genome comparison. **BMC Bioinformatics**, v.7, Suppl 4, p.S18, 2006.

LUO, Y. et al. BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. **BMC Bioinformatics**, v.8, p.266, 2007.

MA, B.; TROMP, J.; LI, M. PatternHunter: faster and more sensitive homology search. **Bioinformatics**, v.18, n.3, p.440-5, 2002.

MACÊDO, J.A. et al. A Molecular Biology Conceptual Model for Information Integration. **Revista Tecnologia da Informação**, v.3, n.2, p.41-8, 2003.

MALTSEV, N. et al. PUMA2--grid-based high-throughput analysis of genomes and metabolic pathways. **Nucleic Acids Research**, v.34, p.D369-D372, 2006.

MARCOTTE, E.M. et al. Detecting protein function and protein-protein interactions from genome sequences. **Science**, v.285, n.5428, p.751-3, 1999.

MARKOWITZ, V.M. et al. The integrated microbial genomes (IMG) system. **Nucleic Acids Research**, v.34, p.D344-D348, 2006.

MERKEEV, I.V.; NOVICHKOV, P.S.; MIRONOV, A.A. PHOG: a database of supergenomes built from proteome complements. **BMC Evoltionary Biology**, v.6, p.52, 2006.

MORGENSTERN, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. BIOINFORMATICS, v.15, n.3, p.211-8, 1999.

MORGENSTERN, B. DIALIGN: finding local similarities by multiple sequence alignment. **Bioinformatics**, v.14, n.3, p.290-4, 1998.

MORGENSTERN, B. et al. Exon discovery by genomic sequence alignment. **Bioinformatics**, v.18, n.6, p.777-87, 2002.

NEEDLEMAN, S.B.; WUNSCH, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, v.48, n.3, p.443-53, 1970.

OTU, H.H.; SAYOOD, K. A new sequence distance measure for phylogenetic tree construction. **Bioinformatics**, v.19, n.16, p.2122-30, 2003.

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.

OVCHARENKO, I. et al. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. Genome Res 2004 Mar;14(3):472-7.

OVERBEEK, R. et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. **Nucleic Acids Research**, v.33, n.17, p.5691-702, 2005.

PAN, X.; STEIN, L.; BRENDEL, V. SynBrowse: a synteny browser for comparative sequence analysis. **Bioinformatics**, v.21, n.17, p.3461-8, 2005.

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.

PETERSON, J.D. et al. The Comprehensive Microbial Resource. **Nucleic Acids Research**, v.29, n.1, p.123-5, 2001.

RANDHAWA, G.S.; BISHAI, W.R. Beneficial impact of genome projects on tuberculosis control. **Infectious Disease Clinics of North America**, v.16, n.1, p.145-61, 2002.

RASKO, D.A.; MYERS, G.S.; RAVEL, J. Visualization of comparative genomic analyses by BLAST score ratio. **BMC Bioinformatics**, v.6, n.1, p.2, 2005.

REFSEQ. NCBI Reference Sequence. Available at: <http://www.ncbi.nlm.nih.gov/RefSeq/> Accessed: 8 Oct. 2007.

REN, Q.; KANG, K.H.; PAULSEN, I.T. TransportDB: a relational database of cellular membrane transport systems. **Nucleic Acids Research**, v.32, p.D284-D288, 2004.

Ren Q, Chen K, Paulsen IT. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 2007 Jan;35(Database issue):D274-279.

ROMUALDI, A. et al. GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. **Bioinformatics**, v.21, n.18, p.3669-71, 2005.

SCHWARTZ, S. et al. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. **Nucleic Acids Research**, v.31, n.13, p.3518-24, 2003a.

SCHWARTZ, S. et al. Human-mouse alignments with BLASTZ. **Genome Research**, v.13, n.1, p.103-7, 2003b.

SCHWARTZ, S. et al. PipMaker--a web server for aligning two genomic DNA sequences. Genome Res, v.10, n.4, p.577-86, 2000.

SCHNEIDER A.; DESSIMOZ, C.; GONNET, GH. OMA Browser--exploring orthologous relations across 352 complete genomes. **Bioinformatics**, v.23, n.16, p.2180-2182, 2007.

SELENGUT, J.D. et al. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* 2007 January;35(Database issue): D260-D264.

SICHERITZ-PONTEN, T.; ANDERSSON, S.G. A phylogenomic approach to microbial evolution. **Nucleic Acids Research**, v.29, n.2, p.545-52, 2001.

SIEW, N.; AZARIA, Y.; FISCHER, D. The ORFanage: an ORFan database. **Nucleic Acids Research**, v.32, p.D281-D283, 2004.

SINHA, A.U.; MELLER, J. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. **BMC Bioinformatics**, v.8, p.82, 2007.

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

STOTHARD, P.; et al. BacMap: an interactive picture atlas of annotated bacterial genomes. **Nucleic Acids Research**, v.33, p.D317-D320, 2005.

STROHMAN, R.C. The coming Kuhnian revolution in biology. **Nature Biotechnology**, v.15, n.3, p.194-200, 1997.

SUHRE, K.; CLAVERIE, J.M. FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. **Nucleic Acids Research**, v.32, p.D273-D276, 2004.

TANAKA, N. et al. G-InforBIO: integrated system for microbial genomics. BMC Bioinformatics 2006;7:368.

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

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

TEKAIA, F.; YERAMIAN, E.; DUJON, B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. **Gene**, v.297, n.1-2, p.51-60, 2002.

TEKAIA, F.; YERAMIAN, E. Genome trees from conservation profiles. **PLoS Computational Biology**, v.1, n.7, p.e75, 2005.

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.

TREANGEN, T.J.; MESSEGUER, X. M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. **BMC Bioinformatics**, v.7, p.433, 2006.

UCHIYAMA, I. MBGD: microbial genome database for comparative analysis. **Nucleic Acids Research**, v.31, n.1, p.58-62, 2003.

UCHIYAMA, I.; HIGUCHI. T.; KOBAYASHI, I. CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. BMC Bioinformatics, v.7, p.472, 2006.

UVA FASTA SERVER. Disponível em: <http://fasta.bioch.virginia.edu/> Acesso em: 08 out. 2007.

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

VON MERING, C. et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. **Nucleic Acids Research**, v.33, p.D433-D437, 2005.

VON MERING, C. et al. STRING 7: recent developments in the integration and prediction of protein interactions. **Nucleic Acids Research**, v.35, p.D358-D362, 2007.

WANG, H. et al. SynView: a GBrowse-compatible approach to visualizing comparative genome data. **Bioinformatics**, v.22, n.18, p.2308-9, 2006.

WEI, L. et al. Comparative genomics approaches to study organism similarities and differences. **Journal of Biomedical Informormatics**, v.35, n.2, p.142-50, 2002.

WILSON, G.A. et al. Orphans as taxonomically restricted and ecologically important genes. **Microbiology**, v.151, n.Pt 8, p.2499-501, 2005.

YANG, J. et al. ShiBASE: an integrated database for comparative genomics of Shigella. **Nucleic Acids Research**, v.34, p.D398-D401, 2006.

YANG, J. et al. GenomeComp: a visualization tool for microbial genome comparison. **Journal of Microbiological Methods**, v.54, n.3, p.423-6, 2003.

About the authors

Antonio Basílio de Miranda

Holds a DSc and a MSc in Biological Sciences(Genetics) from the Federal University of Rio de Janeiro. He has also spent one year at the Sanger Institute as a postdoc student. Currently, he is an Associate Researcher at the Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Fiocruz. He is also the manager of the PDTIS Bioinformatics Platform, and coordinator of several post-graduation courses at Fiocruz. His research lines include comparative genomics and molecular evolution.

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.